Retrieval-Augmented Generation for Large Language Model based Few-shot Chinese Spell Checking

Ming Dong[1,2,3], **Zhiwei Cheng**[1,2,3], **Changyin Luo**[1,2,3], **Tingting He**[1,2,3,*]
[1]Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning,
[2]National Language Resources Monitoring and Research Center for Network Media,
[3]School of Computer, Central China Normal University, Wuhan, China
{dongming,changyinluo,tthe}@ccnu.edu.cn
{thesolution}@mails.ccnu.edu.cn

Abstract

Large language models (LLMs) are naturally suitable for Chinese spelling check (CSC) task in few-shot scenarios due to their powerful semantic understanding and few-shot learning capabilities. Recent CSC research has begun to use LLMs as foundational models. However, most current datasets are primarily focused on errors generated during the text generation process, with little attention given to errors occurring in the modal conversion process. Furthermore, existing LLM-based CSC methods often rely on fixed prompt samples, which limits the performance of LLMs. Therefore, we propose a framework named RagID (Retrieval-Augment Generation and Iterative Discriminator Strategy). By utilizing semantic-based similarity search and an iterative discriminator mechanism, RagID can provide well-chosen prompt samples and reduce over-correction issues in LLM-based CSC. RagID demonstrates excellent effectiveness in few-shot scenarios. We conducted comprehensive experiments, and the results show that RagID achieves the best performance on dataset that include data from multiple domains and dataset containing modal conversion spelling errors. The dataset and code are available online [1].

1 Introduction

Automatic Speech Recognition (ASR) and Optical Character Recognition (OCR) play key roles in various real-world intelligent applications. However, the effectiveness of ASR and OCR applications is restricted by recognition errors caused by technical bottlenecks, complex recognition environments, and individual differences. Therefore, using Chinese Spelling Check (CSC) to correct these errors in Chinese has become an essential and universal approach.

Many of the existing CSC models are implemented based on BERT architecture, treating the correction task as a sequence labeling problem and achieving impressive results by introducing abundant features and fine-grained tuning on training datasets. BERT-based CSC methods are mainly divided into two categories based on the presence of an independent detection phase: one-stage methods (Wu et al., 2023; Liu et al., 2021) and two-stage methods (Huang et al., 2023; Zhang et al., 2020). However, the generalization ability of these methods to real-world application data is challenging, as current evaluation benchmarks do not prioritize the assessment of generalization ability in few-shot scenarios. Moreover, the meticulous design and fine-tuning on existing datasets inadvertently narrow the scope of generalization ability for these methods, rendering them less effective in the scenarios of few-shot learning. Real-world spelling errors in Chinese are complex and changeable, resulting in that most of them are few-shot scenarios. Therefore, improving the generalization ability of the CSC methods is crucial.

LLMs have demonstrated excellent performance in various NLP domains, showing a huge knowledge base and strong semantic comprehension ability. Therefore, using LLMs as foundation models to solve CSC problems has become a valuable research direction (Li et al., 2023b; Dong et al., 2024). LLM-based CSC faces several significant challenges. Firstly, it is essential to establish objective and unbiased criteria for evaluating the efficacy of various CSC methods within the context of few-shot scenarios. Secondly, it is crucial to design effective prompts that enable LLMs to comprehend the specific demands of the CSC task. Lastly, there is a need to prevent over-correction and to determine the optimal point at which to cease alterations.

To address the challenges, we propose a framework named RagID (Retrieval-Augment Generation and Iterative Discriminator Strategy).

* Corresponding author.
[1]https://github.com/ViTsing/RagID

Proceedings of the 31st International Conference on Computational Linguistics, pages 10767–10780
January 19–24, 2025.

RagID utilizes RAG based ICL (In Context Learning) to guide LLMs to detect and correct errors in Chinese sentences, achieving excellent performance on real datasets. Our contributions are summarized as follows:

- We build a modal conversion dataset OAD, which is generated by recognizing Mandarin speech data and handwritten Chinese data. Besides, we conduct a thorough analysis of the differences between OAD and existing CSC datasets.
- We propose an framework to adopt Retrieval-Augmented Generation technique into LLMs based CSC. This framework provides well-chosen fine-grained context for each error-corrected sentence and performs well in few-shot scenarios.
- We propose an iterative discriminator based self-reflection strategy for CSC to avoid over-correction.

2 Related Work

2.1 Chinese Spelling Check Methods

Early work on CSC is conducted from linguistic and statistical perspectives (Jiang et al., 2012; Yu and Li, 2014; Chang et al., 2015). Models detect and correct Chinese spelling errors by formulating grammatical rules or analyzing word frequency. Performance of these rule-based models depends on ability of rule designers and quality of the corpus used for statistics.

With the development of neural networks, data-driven neural network models demonstrate advantages in CSC, but still lack CSC training data. Wang et al. (2018) address this issue by generating a large amount of training data using automatic generation methods. The subsequent neural network models are divided into two classes. One is two-stage method (Zhang et al., 2020), where error detection and correction processes are performed on separate neural networks, and the effectiveness of the correction is limited by the detection ability of the detection neural network. The other is one-stage method (Hong et al., 2019), where the model independently handles detection and correction tasks, predicts results directly. Then, many representative two-stage works have been proposed. Liu et al. (2024) point that CSC is treated as a sequence tagging task, establishing mappings from erroneous characters to correct characters based on semantic context. Xu et al. (2021) add phonetic, graphic, and semantic information to BERT. Wang et al. (2024b) increase the number of error correction attempts and limits the gap between distributions of two correct results to alleviate the problem of lack-correction and over-correction.

After LLMs demonstrating excellent performance in various NLP domains, Wang et al. (2024a) and Li et al. (2023a) investigate the performance of LLMs in CSC. Dong et al. (2024) introduce rich semantic information of Chinese characters into LLMs.

2.2 Chinese Spelling Check Datasets

SIGHAN13 (Wu et al., 2013), SIGHAN14 (Yu et al., 2014), and SIGHAN15 (Chang et al., 2015) are collected from Chinese writing, these datasets are widely used to assess CSC tasks. Lv et al. (2023) extracts data from legal, medical, and official governmental domains. Then volunteers artificially create spelling errors in the sentences. Wu et al. (2023) proposes the dataset LEMON, which is a large-scale multi-domain dataset collected from the everyday life writing corpus. When further investigating spelling errors in modal conversion, image recognition data in Wang et al. (2018) undergoes an unusual text-image-text process, which differs from common recognition scenarios. Moreover, this dataset is widely used for CSC model training, making it unsuitable for evaluation. Therefore, it is necessary to bulid a new modal conversion dataset.

3 Dataset Generation and Profiling

In this section, we introduce a new CSC benchmark dataset collected from OCR and ASR recognition results named OAD (OCR and ASR Dataset), and then show the generation process in detail. We also conduct a fine-grained statistics and analysis of the differences between OAD and other datasets.

3.1 Dataset Generation Process

Data source. In order to deeply analyze the possible spelling errors in the context of handwritten Chinese character recognition and Mandarin speech recognition. We collect image and audio files from publicly available data sources as the original files for creating the dataset OAD. Chinese handwriting data are derived from HWDB2.0, HWDB2.1, HWDB2.2 (Liu et al., 2013) . These datasets contain handwritten articles from hundreds of experimenters, paragraphs of the article

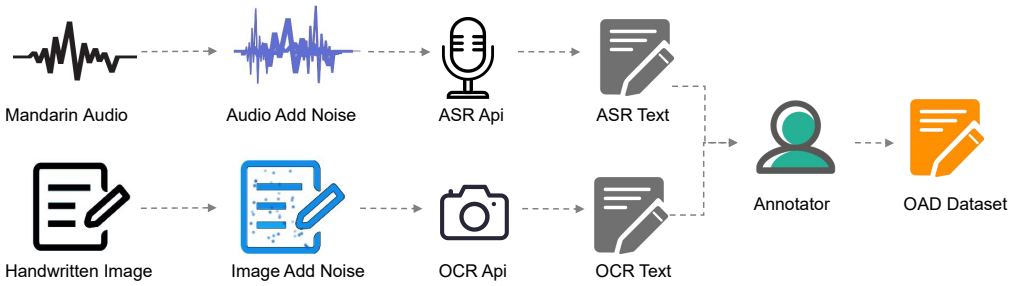

Figure 1: The procedure for constructing dataset OAD. Icons with blobs indicate image data with noise added, and icons with multiple waving lines indicate speech data with interference.

are processed separately and store in the form of image. The sources of the audio files are THCHS-30 (Wang and Zhang, 2015) and AISHELL-1 (Bu et al., 2017). THCHS-30 is a Mandarin speech data resource produced by Tsinghua University, and the main content of the audio is news information. AISHELL-1 is the largest open-source Mandarin audio corpus to date, released by Beijing Beike, containing 400 speakers and more than 170 hours of Mandarin speech data, covering smart home, autonomous driving, financial real estate market, and other fields.

Identification Tool. The Mandarin audio data from THCHS-30 and AISHELL-1 is processed using the short speech recognition interface from Baidu AI Cloud. This interface employs an end-to-end speech and language integrated modeling approach with very high recognition accuracy. This interface has already been successfully applied commercially. Chinese handwritten images from the HWDB dataset are recognized through Baidu AI Cloud's high-precision Universal Character Recognition (UCR) interface, which can recognize almost all commonly used Chinese characters and most rare characters.

Identification Strategy. To collect a wider variety of errors, we simulate a real recognition environment and increase the error rate of recognition tools. During identification, we add some disturbances such as Gaussian noise and salt-and-pepper noise into the image data. Before speech recognition begins, we select some noise from the NOISE-92 audio dataset and mix it into the Mandarin audio data. Fig. 1 illustrates the specific execution process of the recognition strategy.

Post-processing. We employ several experienced annotators to process 4,356 pieces of image recognition data and 2,700 pieces of speech recognition data. The detailed requirements are as follows:

- Restoring the text recognition results to form complete sentences and excluding sentences with fewer than 5 characters.
- Removing duplicate recognized sentences and keeping the first recognition result.
- Correcting punctuation errors in the recognition results and rectifying errors in proper nouns such as place name.

After subsequent annotation and processing, we combine 300 obtained audio recognition results and 200 handwritten image recognition results into a dataset called OAD. This dataset serves as an evaluation dataset for tasks involving handwritten Chinese recognition and Mandarin audio recognition.

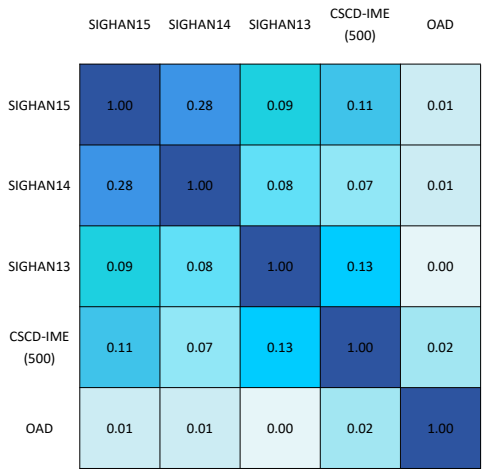

Figure 2: Statistics on the overlap rate between different datasets.

3.2 Dataset Comparison

Shallow Features Analysis. We compare shallow features between widely used existing datasets

Set	SenNum	ISRatio	MaxLen	MinLen	AveLen	MN	PSMN
SIGHAN15	1100	0. 50	108	5	30	703	0. 63
SIGHAN14	1062	0. 49	150	6	50	771	0. 72
SIGHAN13	1000	0. 97	158	17	74	1224	1. 22
CSCD-IME(500)	500	0. 47	122	11	57	255	0. 51
OAD	500	0. 50	113	11	42	480	0. 96

Table 1: Shallow features in different evaluation datasets. SenNum is the number of sentences, ISRatio is the percentage of sentences with errors, MN is the number of wrong characters, and PSMN is the average number of wrong characters per sentence.

and OAD, including SIGHAN13, SIGHAN14, SIGHAN15, CSCD-IME (500). (The details of all datasets are introduced in Section 5.1) As shown in Table 1, OAD dataset has a higher error character density, indicating that it is more challenging to correct spelling errors in OAD than other datasets. Because recognition tools are more prone to consecutive errors, which are different from errors caused by humans.

Error Character Overlap Rate. We collect the 100 error pairs with the highest frequency for each datasets and calculate the overlap ratios between different datasets based on the error pairs from each dataset. Fig. 2 shows that the overlapping proportions among the datasets of the SIGHAN series is high, but there are basically no overlap errors between OAD and other datasets. It indicates that the spelling errors in the OAD dataset are different from other evaluation datasets.

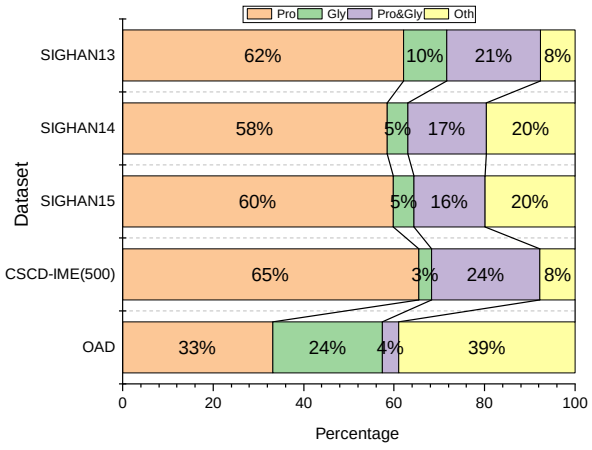

Figure 3: Distribution of error types.

Error Type Distribution. We divide CSC errors into four different types: pronunciation similar misspelling, glyph similar misspelling, both pronunciation and glyph similar misspelling, and others according to (Wang et al., 2024b). The distribution is shown in Fig. 3. We found that the distributions of error types are very similar among the SIGHAN13-15 datasets because they have similar sources. Most of the data in CSCD-IME is generated by the pinyin input method, so there is a large proportion pronunciation similar misspelling errors. In OAD, the proportion of pronunciation similar errors and glyph similar errors are very close, while other types of errors have a higher proportion, resulting in a more balanced error type distribution.

We conclude from the comparison experiment that OAD dataset has a value for further study.

4 Method

4.1 Task Definition

Given a Chinese sentence $X = \{x_1, x_2, ..., x_n\}$ including errors, the correction model aims to generate the corresponding correct result denoted as $Y = \{y_1, y_2, ..., y_n\}$. It must be noted that, the current CSC studies all stipulate that X and Y must have the same length. This task typically involves two sub-steps: first, detecting the location of the error x_i, and then providing the correct character y_i for that position. In the final result statement, if the erroneous character has changed, it is considered to be successfully detected. If this erroneous character is changed to the correct character, it is considered to be successfully corrected.

4.2 Overview of RagID

The structure of RagID is shown in Fig 4. RagID includes three key modules: Retriever, Corrector, and Discriminator. We collect a large number of Chinese spelling examples and convert these examples into low-dimensional vector representations through vectorization technique. These representations are stored in a pre-established vector database as our external knowledge base in RAG. Retriever uses the same vectorization technique to convert the given sentence X into a vector and searches for related correction examples. Corrector is an LLM

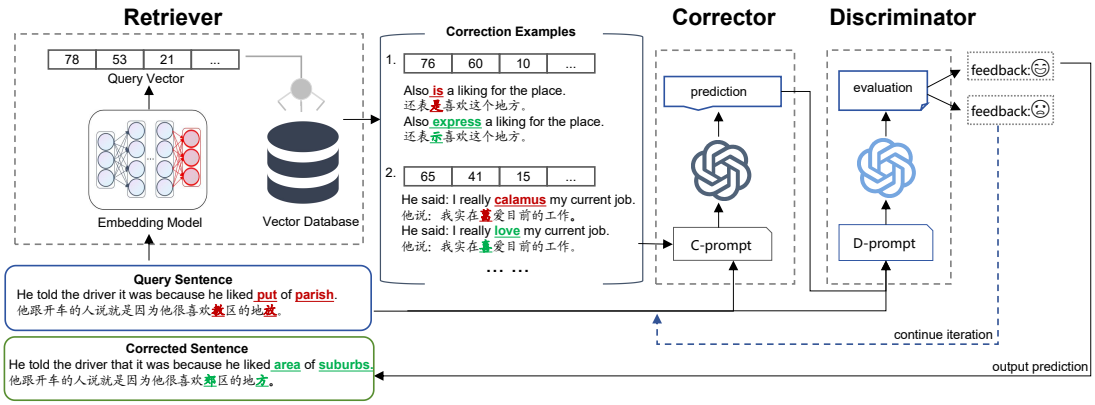

Figure 4: Structure diagram of RagID, including three main modules: Retriever, Corrector, and Discriminator.

that learns from these examples to understand the specific rules of the CSC task. In the final step, Discriminator evaluates the reasonableness of the correction result and provides feedback to the Corrector if the result is unreasonable, prompting a new round of correction.

4.3 RAG for CSC

The core idea of ICL based CSC is to extract rule vectors θ from the examples and rules in prompt (Li, 2023; Hendel et al., 2023). Therefore, LLM constructs a new correction function $f(X;\theta;T)$ after ICL process, where T denotes the LLM and X denotes the query sentence. LLMs accomplish CSC tasks through this principle.

We design prompt templates for ICL to enhance the utilization of correction examples in LLMs. These templates outline the roles of the corrector and the discriminator, background of the CSC task, task objectives, sequential correction process steps, and the desired output format. The prompt template used in this work is shown in the Appendix.

In RagID, we obtain context examples using the Retrieval-Augmented Generation (RAG) strategy from vector database. The CSC datasets used for RAG are introduced in Table C of the Appendix. We use the vectorization model '*bge-large-zh*' to convert these CSC examples into low-dimensional vector representations, which are stored in the vector database '*milvus*'. These data stored in *milvus* serve as the external knowledge base for subsequent correction. When the corrector requires prompt examples, we employ the nearest neighbor search method to find n similar sentences in the external knowledge base and get $E = \{G_1, G_2, ..., G_n\}$. Each $G_i = (x_i, y_i)$ includes uncorrected CSC sample x_i and its gold answer y_i. Guided by CSC prompt, the corrector generates correction predictions $P = \{y_1, y_2, ..., y_n\}$ based on external examples E and query sentence X, completing one cycle of detection and correction. This strategy offers the advantage of enabling LLMs to rapidly adapt to CSC tasks without fine-tuning, facilitating a swift integration with more advanced foundational models to continuously enhance performance.

4.4 Iterative Discriminator Strategy

Due to the complexity of CSC task (Kiyono et al., 2019), models often suffer from both over-correction and under-correction to some extent. To address these issues, we introduce a discriminator following the corrector, and it collaborate using an iterative reflection strategy. Discriminator will evaluate the prediction, emulating human evaluation to provide either affirmative or critical feedback. Specifically, after receiving query sentence $X = \{x_1, x_2, ..., x_n\}$ and the corresponding auxiliary samples $E = \{G_1, G_2, ..., G_n\}$, the corrector generates the first prediction $Y^1 = \{y_1, y_2, ..., y_n\}$. Discriminator learns discrimination methods from fixed discrimination examples $F = \{D_1, D_2, D_3, D_4\}$ and then evaluates predictions from corrector. If Y^1 does not meet the criteria, the result will be returned to the corrector along with feedback, corrector continues to iterate and get Y^i until the predicted results satisfy the requirements of the discriminator or maximum number of times is exceeded, where i denotes the time of iteration.

IDS can fully encourage the corrector LLM to rethinking and reflecting, thereby enhancing effectiveness. LLMs for discrimination can be the same as corrector or an external LLM, our work use the former. A specific example of the operation of the discriminator can be seen in the Fig. 7 of Appendix.

By using Iterative Discriminator Strategy, it can ensure that the length of the output sentence remains unchanged. In addition, this mechanism also ensures that ICL information can be better utilized.

5 Experiments

5.1 Datasets

SIGHAN13 (Wu et al., 2013) was a dataset proposed in 2013. SIGHAN13 consists of sentences written by Chinese students aged 13 to 14 in a language examination. Average sentence length in the SIGHAN13 dataset is 70.
SIGHAN14 (Yu et al., 2014) was a dataset proposed in 2014. SIGHAN14 consists of articles written by foreigners in the process of learning Chinese. Sentences of SIGHAN14 contains both Chinese spelling errors and grammatical mistakes.
SIGHAN15 (Chang et al., 2015) was a dataset proposed in 2015. Although sourced from non-native Chinese speakers, SIGHAN15 is one of the most commonly used benchmarks in CSC.
CSCD-IME(500) (Hu et al., 2022) was a dataset proposed in 2022, consists of official posts from Sina Weibo. The data spans multiple domains, including law, politics, entertainment, and daily life. Phonetic similar spelling errors account for a large proportion. We randomly extract 500 samples from CSCD-IME as evaluation dataset CSCD-IME(500).
OAD is the dataset generated by us, consists of OCR and ASR recognition results from Mandarin speech data and handwritten Chinese data across multiple domains, including environment, politics, medicine, and history.

5.2 Baselines

BERT (Devlin et al., 2019) has a softmax layer added to the top to predict the characters distribution.
ReaLiSe (Xu et al., 2021) integrates phonetic encoder, semantic encoder and glyph encoder to capture and fuse information from three modalities. The fusion weights of different information are continuously updated during training process.
CRASpell (Liu et al., 2022) proposes a noise modeling module to generate noisy context in training process to deal with Contextual Typo Disturbance. Furthermore, CRASpell incorporates a copy block in the correction model, which encourages model to prefer to keep the input character when the corrected characters and input characters are both valid in context.
SCOPE (Li et al., 2022) builds two parallel decoders on top of the shared encoder, for the main CSC task and fine-grained auxiliary Chinese pronunciation prediction (CPP) task. Furthermore, SCOPE adopts a novel adaptive weighting scheme to balance the two tasks.

5.3 Evaluation Metrics

Sentence-level Evaluation Metric is utilized for measuring CSC performance of models in this work. Sentence-level evaluation metric includes both detection and correction aspects. A sentence is considered correct at the detection aspect only if all errors are detected and considered correct at the correction aspect only if all errors are corrected to the target characters. Each evaluation aspect consists of four parameters: accuracy, precision, recall, and F1 score.
F1 Score is calculated from the precision and recall. Therefore, detection F1 and correction F1 scores are typically used to represent level of detection and correction ability, respectively. In our experiments, the specific formulation of the F1 score is: $f_1 = 2 * p * r/(p + r)$.

5.4 Detail Settings

We selected n sentences from the test dataset as few-shot samples for fine-tuning baseline models and for In-Context Learning with LLMs. The baseline models are fine-tuned on the samples with a learning rate of {3e-5, 5e-5}, for 10 to 30 epochs. When using LLMs like GPT3.5 or GLM4 for CSC task, we only modify the prompt or replace the fixed few-shot samples with RAG to optimize effectiveness of In-Context Learning. Both methods have access to the fixed few-shot samples on the target dataset, ensuring fairness.

Among the four baseline models, BERT and ReaLiSe are pre-trained only on unsupervised data but SCOPE and CRASpell undergo a second step of training on annotated CSC data (Wang et al., 2018). The foundation model for RagID is api of GPT3.5 Turbo and api of GLM4 (Du et al., 2022; Zeng et al., 2023). The local experimental environment is a workstation equipped with an Intel-12600k CPU, 64GB memory, and RTX-4090 GPU.

Dataset	Model	d-a	d-p	d-r	d-f	c-a	c-p	c-r	c-f
SIGHAN13	BERT+FixedFew-Shot	1.30	1.11	1.12	0.20	0.00	0.00	0.00	0.00
	ReaLiSe+FixedFew-Shot	0.20	0.20	0.20	0.20	0.00	0.00	0.00	0.00
	CRASpell+FixedFew-Shot	32.30	43.40	31.51	36.52	30.90	41.42	30.07	34.84
	SCOPE+FixedFew-Shot	**64.40**	**72.37**	**63.95**	**67.90**	**56.69**	**63.40**	**56.02**	**59.48**
	GLM4+RAGFew-Shot+IDS	47.40	62.14	46.65	53.29	42.90	55.97	42.02	48.00
	GPT3.5+RAGFew-Shot+IDS	48.10	58.52	47.37	52.36	45.40	55.09	44.59	49.29
SIGHAN14	BERT+FixedFew-Shot	12.42	3.14	5.57	4.01	9.69	0.00	0.00	0.00
	ReaLiSe+FixedFew-Shot	1.41	0.57	1.15	0.76	0.84	0.00	0.00	0.00
	CRASpell+FixedFew-Shot	52.73	31.97	25.58	28.42	51.79	29.57	23.65	26.28
	SCOPE+FixedFew-Shot	**64.97**	**50.78**	**43.65**	**46.94**	**62.24**	**44.29**	**38.07**	**40.95**
	GLM4+RAGFew-Shot+IDS	49.34	30.16	26.15	28.01	47.74	26.39	22.88	24.51
	GPT3.5+RAGFew-Shot+IDS	56.87	40.99	41.54	41.26	54.99	37.19	37.69	37.44
SIGHAN15	BERT+FixedFew-Shot	22.90	6.69	10.53	8.18	17.81	0.11	0.18	0.14
	ReaLiSe+FixedFew-Shot	5.54	1.91	3.69	2.52	3.81	0.09	0.18	0.12
	CRASpell+FixedFew-Shot	63.36	53.69	39.00	45.18	62.09	50.13	36.41	42.18
	SCOPE+FixedFew-Shot	**73.09**	**67.19**	**55.26**	**60.64**	**69.27**	**57.75**	**47.5**	**52.12**
	GLM4+RAGFew-Shot+IDS	60.27	46.8	41.96	44.25	57.64	40.82	36.6	38.6
	GPT3.5+RAGFew-Shot+IDS	63.00	51.50	50.65	51.07	59.55	44.36	43.62	43.99
CSCD-IME(500)	BERT+FixedFew-Shot	10.40	4.58	8.97	6.06	6.06	0.00	0.00	0.00
	ReaLiSe+FixedFew-Shot	0.40	0.40	0.85	0.54	0.20	0.20	0.42	0.27
	CRASpell+FixedFew-Shot	51.20	35.35	**32.48**	33.85	50.60	33.95	**31.20**	32.52
	SCOPE+FixedFew-Shot	28.19	16.81	24.78	20.03	27.20	15.36	22.64	18.30
	GLM4+RAGFew-Shot+IDS	58.00	50.47	23.08	31.67	57.20	46.73	21.37	29.33
	GPT3.5+RAGFew-Shot+IDS	**61.60**	**60.71**	29.06	**39.31**	**60.80**	**57.14**	27.35	**36.99**
OAD	BERT+FixedFew-Shot	15.60	3.75	6.40	4.73	12.40	0.00	0.00	0.00
	ReaLiSe+FixedFew-Shot	1.60	0.80	1.60	1.07	0.80	0.00	0.00	0.00
	CRASpell+FixedFew-Shot	49.40	19.33	11.60	14.50	47.60	13.33	8.00	10.00
	SCOPE+FixedFew-Shot	59.59	38.13	**39.20**	38.65	52.20	23.73	**24.40**	24.06
	GLM4+RAGFew-Shot+IDS	**65.20**	**49.47**	37.20	**42.47**	**58.40**	**31.38**	23.60	**26.94**
	GPT3.5+RAGFew-Shot+IDS	57.60	37.31	28.80	32.51	52.20	23.32	18.00	20.32

Table 2: Main results of all baselines. **Bold** indicates the best results. The test results of LLMs are the average performance of 3 times.

5.5 Main Results

Table 2 shows the CSC evaluation results of our RagID and other baselines on various datasets. We observe that:

In SIGHAN13, SIGHAN14, and SIGHAN15 datasets, deep learning model SCOPE achieves the best results, outperforming the other models in both detection and error correction F1 values. These datasets are all sourced from exam essays, essays of the latter two datasets are written by non-native speakers. Spelling errors in the SIGHAN datasets are limited in difficulty and also domain-specific. Therefore, both the deep learning model SCOPE and the large-scale language model GLM4 excel at detecting and correcting these errors. We surmise that SCOPE performs better because it has encountered more similar errors during training process.

GPT-3.5 and GLM4, based on the RagID framework, achieve the best results on CSCD-IME and OAD, respectively. In contrast, SCOPE experiences performance degradation, with an error-correcting F1 score 18.69 points behind GPT-3.5 on CSCD-IME and 2.88 points behind GLM4 on OAD. We believe SCOPE's performance decline is due to the multi-domain nature of CSCD-IME and OAD, which results in different spelling error distributions compared to single-domain datasets. The deep learning model struggles to generalize across these domains because of parameter constraints and other limitations. Additionally, OAD's advantage over CSCD-IME is smaller because it contains denser and more glyph-related errors, presenting greater challenges for spelling detection and correction.

We observe that deep learning models perform better when dataset is homogeneous and has ample training data related to dataset. However, experiments with BERT and ReaLiSe show that even pretrained models with rich knowledge struggle with the CSC task without training on large-scale annotated data. In contrast, LLMs based on the RagID framework still perform well in multi-domain and

Base Model	Strategy	D-Acc	D-Pre	D-Rec	D-F1	C-Acc	C-Pre	C-Rec	C-F1
GPT3.5	FixedFew-Shot	57.09	43.23	45.47	44.32	54.36	37.96	39.93	38.92
	RAGFew-Shot	61.55	50.0	47.13	48.53	57.82	41.96	39.56	40.72
	FixedFew-Shot+IDS	58.91	44.71	48.43	46.5	55.91	39.08	42.33	40.64
	RAGFew-Shot+IDS	**63.00**	**51.50**	**50.65**	**51.07**	**59.55**	**44.36**	**43.62**	**43.99**

Table 3: Ablation experiments of RagID framework. We test GPT3.5 based on RagID framework on SIGHAN15 evaluation dataset. We conduct the following modifications on RAGFew-Shot+IDS: Without RAG (FixedFew-Shot+IDS). Without IDS (RAGFew-Shot). Without both RAG and IDS, exclusively utilize fixed few-shot prompt samples as fine-grained context (FixedFew-Shot).

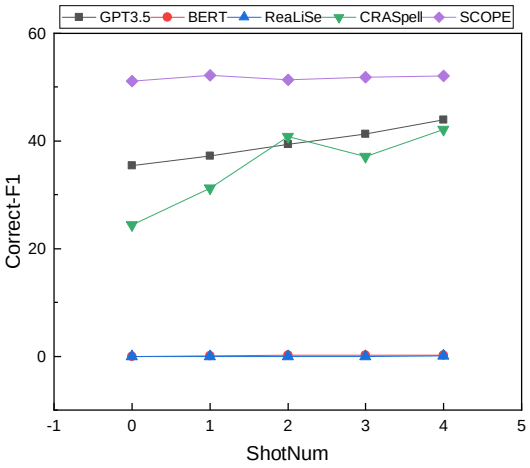

Figure 5: The performance of the model on the dataset SIGHAN15 when the fine-tuned few-shot changes. The ordinate is the F1 value of error correction.

low-data scenarios, demonstrating the superiority of LLMs and our framework.

5.6 Ablation Study

Samples Size. We research the impact of different sample sizes on the error correction performance of baseline models and GPT-3.5-based RagID. Fig. 5 illustrates that as samples increase, SCOPE, CRASPell and RagID based GPT3.5 show an increasing trend, but performance of SCOPE and CRASpell may fluctuate occasionally due to overfitting. However, there is no significant change in the correction ability of pre-training models BERT and ReaLiSe.

Strategy. The ablation experiment which investigates the contributions of different strategies based on LLMs are organized as follows: 1) Removing RAG. 2) Removing IDS. 3) Simultaneously removing RAG and IDS. Specific experimental results are shown in Table 3. The results in Table 3 from the SIGHAN15 test set show that using either IDS or RAG alone improves the correction performance, both strategies are effective. Furthermore, With the addition of both strategies, the effectiveness is further improved, which shows RAG and IDS offer assistance from distinct perspectives and there is a synergistic relationship between them.

5.7 Case Study

We present a case study in Table 4 of Appendix. From the first example, we can see that corrector based on LLMs cannot detect and correct unfamiliar spelling error "suffered" with fixed prompt samples. After identifying prompt samples from an external knowledge corpus with RAG technique, corrector successfully corrects "suffered" to "received". In the second example, the omission of a word causes a grammatical problem in the sentence. Discriminator based on LLMs bypasses the negative effects of grammatical errors and successfully understands meaning of the sentence, warns that there are still some spelling errors in this sentence.And then corrector successfully changes the error "bay" to the correct form "playing" at the third iteration with IDS. These examples demonstrate the effectiveness of our method.

6 Conclusion

This work concentrates on the LLMs-based CSC task. Initially, we introduce an evaluation dataset, OAD, which is derived from the recognition results of Mandarin speech data and handwritten Chinese data. It has been shown that OAD exhibits new features that are distinct from those of existing datasets. Besides, we propose a CSC method RagID that leverages the RAG technique and LLMs. This method harnesses the strong generalization ability of LLMs to correct errors and incorporates an iterative discriminator strategy (IDS) to enhance correction performance. RagID shows excellent effectiveness in few-shot CSC task with the best results on multi-domain and high-difficulty datasets.

7 Limitations

There is still a gap between LLMs based on RagID architecture and the best deep learning model SOCPE on SIGHAN15 evaluation dataset, we think current methods does not fully realize the potential of LLMs in few-shot scenarios. Current semantic-based RAG method cannot find best prompt samples as fine-grained context for each error, which limits the correction ability. The discriminant criteria of the discriminator can also be further optimized. We will work on improving these issues in future work.

8 Ethics Statement

We conduct research in strict adherence to the principles of the ACL Code of Ethics. The datasets used in our research is a certified public dataset to ensure that the data does not leak personal privacy or violate social ethics. During the research process, we strictly follow the harmlessness principle to avoid misleading large language models to produce harmful output, making our research does not negatively affect any participant, group or society. We anticipate that the research presented in this paper will not directly causes any social issues or ethical challenges.

Acknowledgments

This work was partly supported by China Postdoctoral Science Foundation (No. 2023M731253), Hubei Provincial Natural Science Foundation (No. 2023AFB487), and self-determined research funds of CCNU from the colleges' basic research and operation of MOE (Grant No. CCNU24ai011, CCNU24ai012).

References

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline. In *O-COCOSDA*, pages 1–5. IEEE.

Tao-Hsing Chang, Hsueh-Chih Chen, and Cheng-Han Yang. 2015. Introduction to a proofreading tool for chinese spelling check task of SIGHAN-8. In *SIGHAN@IJCNLP*, pages 50–55. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Ming Dong, Yujing Chen, Miao Zhang, Hao Sun, and Tingting He. 2024. Rich semantic knowledge enhanced large language models for few-shot chinese spell checking. *arXiv preprint arXiv:2403.08492*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *EMNLP (Findings)*, pages 9318–9333. Association for Computational Linguistics.

Yuzhong Hong, Xianguo Yu, Neng He, Nan Liu, and Junhui Liu. 2019. Faspell: A fast, adaptable, simple, powerful chinese spell checker based on dae-decoder paradigm. In *W-NUT@EMNLP*, pages 160–169. Association for Computational Linguistics.

Yong Hu, Fandong Meng, and Jie Zhou. 2022. CSCD-IME: correcting spelling errors generated by pinyin IME. *CoRR*, abs/2211.08788.

Haojing Huang, Jingheng Ye, Qingyu Zhou, Yinghui Li, Yangning Li, Feng Zhou, and Hai-Tao Zheng. 2023. A frustratingly easy plug-and-play detection-and-reasoning module for chinese spelling check. In *EMNLP (Findings)*, pages 11514–11525. Association for Computational Linguistics.

Ying Jiang, Tong Wang, Tao Lin, Fangjie Wang, Wenting Cheng, Xiaofei Liu, Chenghui Wang, and Weijian Zhang. 2012. A rule based chinese spelling and grammar detection system utility. In *2012 International Conference on System Science and Engineering (ICSSE)*, pages 437–440. IEEE.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. In *EMNLP/IJCNLP (1)*, pages 1236–1242. Association for Computational Linguistics.

Jiahao Li, Quan Wang, Zhendong Mao, Junbo Guo, Yanyan Yang, and Yongdong Zhang. 2022. Improving chinese spelling check by character pronunciation prediction: The effects of adaptivity and granularity. In *EMNLP*, pages 4275–4286. Association for Computational Linguistics.

Kunting Li, Yong Hu, Shaolei Wang, Hanhan Ma, Liang He, Fandong Meng, and Jie Zhou. 2023a. Eval-gcsc: A new metric for evaluating chatgpt's performance in chinese spelling correction. *CoRR*, abs/2311.08219.

Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Yangning Li, Feng Zhou, Hai-Tao Zheng, and Qingyu Zhou. 2023b. On the (in)effectiveness of large language models for chinese text correction. *CoRR*, abs/2307.09007.

Yinheng Li. 2023. A practical survey on zero-shot prompt design for in-context learning. In *RANLP*, pages 641–647. INCOMA Ltd., Shoumen, Bulgaria.

Cheng-Lin Liu, Fei Yin, Da-Han Wang, and Qiu-Feng Wang. 2013. Online and offline handwritten chinese character recognition: Benchmarking on new databases. *Pattern Recognit.*, 46(1):155–162.

Linfeng Liu, Hongqiu Wu, and Hai Zhao. 2024. Chinese spelling correction as rephrasing language model. In *AAAI*, pages 18662–18670. AAAI Press.

Shulin Liu, Shengkang Song, Tianchi Yue, Tao Yang, Huihui Cai, Tinghao Yu, and Shengli Sun. 2022. Craspell: A contextual typo robust approach to improve chinese spelling correction. In *ACL (Findings)*, pages 3008–3018. Association for Computational Linguistics.

Shulin Liu, Tao Yang, Tianchi Yue, Feng Zhang, and Di Wang. 2021. PLOME: pre-training with misspelled knowledge for chinese spelling correction. In *ACL/IJCNLP (1)*, pages 2991–3000. Association for Computational Linguistics.

Qi Lv, Ziqiang Cao, Lei Geng, Chunhui Ai, Xu Yan, and Guohong Fu. 2023. General and domain-adaptive chinese spelling check with error-consistent pretraining. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 22(5):124:1–124:18.

Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang. 2018. A hybrid approach to automatic corpus generation for chinese spelling check. In *EMNLP*, pages 2517–2527. Association for Computational Linguistics.

Dong Wang and Xuewei Zhang. 2015. THCHS-30 : A free chinese speech corpus. *CoRR*, abs/1512.01882.

Xi Wang, Ruoqing Zhao, Hongliang Dai, and Piji Li. 2024a. An empirical investigation of domain adaptation ability for chinese spelling check models. *CoRR*, abs/2401.14630.

Yue Wang, Zilong Zheng, Zecheng Tang, Juntao Li, Zhihui Liu, Kunlong Chen, Jinxiong Chang, Qishen Zhang, Zhongyi Liu, and Min Zhang. 2024b. Towards better chinese spelling check for search engines: A new dataset and strong baseline. In *WSDM*, pages 769–778. ACM.

Hongqiu Wu, Shaohua Zhang, Yuchen Zhang, and Hai Zhao. 2023. Rethinking masked language modeling for chinese spelling correction. In *ACL (1)*, pages 10743–10756. Association for Computational Linguistics.

Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee. 2013. Chinese spelling check evaluation at SIGHAN bake-off 2013. In *SIGHAN@IJCNLP*, pages 35–42. Asian Federation of Natural Language Processing.

Heng-Da Xu, Zhongli Li, Qingyu Zhou, Chao Li, Zizhen Wang, Yunbo Cao, Heyan Huang, and Xian-Ling Mao. 2021. Read, listen, and see: Leveraging multimodal information helps chinese spell checking. In *ACL/IJCNLP (Findings)*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 716–728. Association for Computational Linguistics.

Junjie Yu and Zhenghua Li. 2014. Chinese spelling error detection and correction based on language model, pronunciation, and shape. In *CIPS-SIGHAN*, pages 220–223. Association for Computational Linguistics.

Liang-Chih Yu, Lung-Hao Lee, Yuen-Hsien Tseng, and Hsin-Hsi Chen. 2014. Overview of SIGHAN 2014 bake-off for chinese spelling check. In *CIPS-SIGHAN*, pages 126–132. Association for Computational Linguistics.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2023. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li. 2020. Spelling error correction with soft-masked BERT. In *ACL*, pages 882–890. Association for Computational Linguistics.

A Detailed Prompt

Corrector Few_Shot Prompt

User: 我想让你担任一个强大的中文拼写纠错模型，你的职责是对中文句子中的汉字进行拼写检查和纠错。

User: I want you to serve as a powerful Chinese spelling correction model, and your role is to spell check and correct Chinese characters in Chinese sentences.

User: 待处理中文句子:***我很高兴受到你们结婚的邀请单，你们到底决定结婚了！*** 待处理中文句子的字符串长度:***25*** 处理指令:结合中文语言学规则、上下文信息对上文中的待处理中文句子进行检查，判断句子中是否错误使用了拼音相似或者形状相似的汉字，导致中文句子原本的语义发生改变。如果句子中有这种错误使用的汉字，就找到正确的汉字，对错误汉字进行一对一替换。有几个错误汉字就替换几次，替换操作发生前后中文句子的字符串长度不发生改变；如果没有错误使用的汉字，则不进行替换。 结果模板：是否存在错误汉字:<> 处理后的中文句子:<> 处理后的中文句子的字符串长度:<>

User: Pending Chinese sentence: ***I am very happy to suffer your invitation to get married, you have finally decided to get married!*** String length of the pending Chinese sentence: ***25*** Processing instructions: Check the above Chinese sentences in combination with Chinese linguistic rules and context information to determine whether Chinese characters with similar pinyin or similar shapes are used incorrectly in the sentence, resulting in changes in the original semantics of Chinese sentences. If there is such a misused kanji in the sentence, find the correct kanji and replace the wrong kanji one-to-one. There are several wrong Chinese characters are replaced several times, and the string length of the Chinese sentence does not change before and after the replacement operation; If there are no incorrectly used kanji, they will not be replaced. Result Template: Whether there are wrong Chinese characters: <> Processed Chinese sentences: <> length of processed Chinese sentences: <>

Assistant: 是否存在错误汉字:***是*** 处理后的中文句子:***我很高兴收到你们结婚的邀请单，你们到底决定结婚了！*** 处理后的中文句子的字符串长度:***25***

Assistant: Is there a wrong Chinese character: ***Yes***. Processed Chinese sentence: ***I am very happy to receive your invitation to get married, you have finally decided to get married!*** String length of processed Chinese sentence: ***25***

User: 待处理中文句子:***我很高兴你们邀请我参加你们的婚礼。*** 待处理中文句子的字符串长度:***17*** 处理指令:结合中文语言学规则、上下文信息对上文中的待处理中文句子进行检查，判断句子中是否错误使用了拼音相似或者形状相似的汉字，导致中文句子原本的语义发生改变。如果句子中有这种错误使用的汉字，就找到正确的汉字，对错误汉字进行一对一替换。有几个错误汉字就替换几次，替换操作发生前后中文句子的字符串长度不发生改变；如果没有错误使用的汉字，则不进行替换。 结果模板：是否存在错误汉字:<> 处理后的中文句子:<> 处理后的中文句子的字符串长度:<>

User: Pending Chinese sentence: ***I'm glad you invited me to your wedding.*** String length of the pending Chinese sentence: ***17*** Processing instructions: Check the above Chinese sentences in combination with Chinese linguistic rules and context information to determine whether Chinese characters with similar pinyin or similar shapes are used incorrectly in the sentence, resulting in changes in the original semantics of Chinese sentences. If there is such a misused kanji in the sentence, find the correct kanji and replace the wrong kanji one-to-one. There are several wrong Chinese characters are replaced several times, and the string length of the Chinese sentence does not change before and after the replacement operation; If there are no incorrectly used kanji, they will not be replaced. Result Template: Whether there are wrong Chinese characters: <> Processed Chinese sentences: <> length of processed Chinese sentences: <>

Assistant: 是否存在错误汉字:***否*** 处理后的中文句子:***我很高兴你们邀请我参加你们的婚礼。*** 处理后的中文句子的字符串长度:***17***

Assistant: Is there a wrong Chinese character: ***No***. Processed Chinese sentence: ***I'm glad you invited me to your wedding.*** String length of processed Chinese sentence: ***17***

...

CSC Query Prompt

User: 待处理中文句子:***我受到了你们的结婚卡。*** 待处理中文句子的字符串长度:***11*** 处理指令:结合中文语言学规则、上下文信息对上文中的待处理中文句子进行检查，判断句子中是否错误使用了拼音相似或者形状相似的汉字，导致中文句子原本的语义发生改变。如果句子中有这种错误使用的汉字，就找到正确的汉字，对错误汉字进行一对一替换。有几个错误汉字就替换几次，替换操作发生前后中文句子的字符串长度不发生改变；如果没有错误使用的汉字，则不进行替换。 结果模板：是否存在错误汉字:<> 处理后的中文句子:<> 处理后的中文句子的字符串长度:<>

User: Pending Chinese sentence: ***I suffered your wedding card.*** String length of the pending Chinese sentence: ***11*** Processing instructions: Check the above Chinese sentences in combination with Chinese linguistic rules and context information to determine whether Chinese characters with similar pinyin or similar shapes are used incorrectly in the sentence, resulting in changes in the original semantics of Chinese sentences. If there is such a misused kanji in the sentence, find the correct kanji and replace the wrong kanji one-to-one. There are several wrong Chinese characters are replaced several times, and the string length of the Chinese sentence does not change before and after the replacement operation; If there are no incorrectly used kanji, they will not be replaced. Result Template: Whether there are wrong Chinese characters: <> Processed Chinese sentences: <> length of processed Chinese sentences: <>

Figure 6: Taskspecific Few_shot Correction prompt for CSC task. We've marked key information in italics and bold, originally incorrect words in red, correctly changed words in green, incorrectly changed words in blue.

Discriminator Few_Shot Prompt

User: 我想让你担任一个中文拼写纠错的评价模型，你的职责是结合相关领域知识，判断中文拼写纠错模型的纠错结果是否符合要求。

User: I want you to serve as an evaluation model for Chinese spelling correction, and your responsibility is to judge whether the error correction results of the Chinese spelling error correction model meet the requirements based on relevant domain knowledge.

User: 纠错前的中文句子:***在公车上有很多人，所以我们没有位子可以座。*** 纠错前的中文句子长度:***21*** 纠错后的中文句子:***在公车上有很多人，所以我们没有位子可以坐。*** 纠错后的中文句子长度:***21*** 评价标准:结合中文拼写纠错的知识，判断当前发生的纠错是否合理，纠错后的中文句子中是否仍然存在错误？纠错后的句子是否更加符合语义？纠错前后句子长度是否一致？只输出结果模板的内容。评价结果模板: 纠错后的句子已经不存在字符错误:<> 纠错后的句子是否符合语法逻辑:<> 纠错后的句子和纠错前句子长度相等:<>

User: Chinese sentence before error correction: ***There are many people on the bus, so we don't have seats to seat.*** Chinese sentence length before error correction: ***21*** Chinese sentence after error correction: ***There are many people on the bus, so we don't have seats to sit***. Length of corrected Chinese sentences: ***21*** Evaluation criteria: Combined with the knowledge of Chinese spelling correction, judge whether the current correction is reasonable, and whether there are still errors in the corrected Chinese sentences? Are the corrected sentences more semantic? Is the sentence length consistent before and after error correction? Only the contents of the result template are output. Evaluation Result Template: The corrected sentence no longer has a character error: <> The corrected sentence is more reasonable than the sentence before the correction: <> The length of the corrected sentence is equal to the sentence before the correction: <>

Assistant: 纠错后的句子已经不存在字符错误:***是*** 纠错后的句子是否符合语法逻辑:***是*** 纠错后的句子和纠错前句子长度相等:***是***

Assistant: The corrected sentence no longer has a character error: ***Yes***. The corrected sentence is reasonable: ***Yes***. The corrected sentence is equal in length to the sentence before the correction: ***Yes***

User: 纠错前的中文句子:***在补习班他昨天晚上到夜里两点还在读书，所以他一回家就累得睡着了。*** 纠错前的中文句子长度:32 纠错后的中文句子:***在补习班他昨天晚上到夜里两点还在读书，所以他一回家就累得不动了。*** 纠错后的中文句子长度:32 评价标准:结合中文拼写纠错的知识，判断当前发生的纠错是否合理，纠错后的中文句子中是否仍然存在错误？纠错后的句子是否更加符合语义？纠错前后句子长度是否一致？只输出结果模板的内容。评价结果模板: 纠错后的句子已经不存在字符错误:<> 纠错后的句子是否符合语法逻辑:<> 纠错后的句子和纠错前句子长度相等:<>

User: Chinese sentence before error correction: ***In cram school, he was still studying from last night to two o'clock at night, so he fell asleep tired as soon as he got home.*** Chinese sentence length before error correction: 32 Chinese sentence after error correction: ***In cram school, he was still studying from last night to two o'clock at night, so he was too tired to move as soon as he got home.*** Chinese sentence length after error correction: 32 Evaluation criteria: Combined with the knowledge of Chinese spelling correction, judge whether the current correction is reasonable, and whether there are still errors in the corrected Chinese sentences? Are the corrected sentences more semantic? Is the sentence length consistent before and after error correction? Only the contents of the result template are output. Evaluation Result Template: The corrected sentence no longer has a character error: <> The corrected sentence is more reasonable than the sentence before the correction: <> The length of the corrected sentence is equal to the sentence before the correction: <>

Assistant: 纠错后的句子已经不存在字符错误:***否*** 纠错后的句子是否符合语法逻辑:***否*** 纠错后的句子和纠错前句子长度相等:***是***

Assistant: The corrected sentence no longer has a character error: ***No***. The corrected sentence is reasonable: ***No***. The corrected sentence is equal in length to the sentence before the correction: ***Yes***

...

Evaluate Query Prompt

User: 纠错前的中文句子:***我受到了你们的结婚卡。*** 纠错前的中文句子长度:***11*** 纠错后的中文句子:***我收到了你们的结婚卡。*** 纠错后的中文句子长度:***11*** 评价标准:结合中文拼写纠错的知识，判断当前发生的纠错是否合理，纠错后的中文句子中是否仍然存在错误？纠错后的句子是否更加符合语义？纠错前后句子长度是否一致？只输出结果模板的内容。评价结果模板: 纠错后的句子已经不存在字符错误:<> 纠错后的句子是否符合语法逻辑:<> 纠错后的句子和纠错前句子长度相等:<>

User: Chinese sentence before error correction: ***I suffered your wedding card.*** Chinese sentence length before error correction: ***11*** Chinese sentence after error correction: ***I received your marriage card.*** Chinese sentence length after error correction: ***11*** Evaluation criteria: Combined with the knowledge of Chinese spelling correction, judge whether the current correction is reasonable, and whether there are still errors in the corrected Chinese sentences? Are the corrected sentences more semantic? Is the sentence length consistent before and after error correction? Only the contents of the result template are output. Evaluation Result Template: The corrected sentence no longer has a character error: <> The corrected sentence is more reasonable than the sentence before the correction: <> The length of the corrected sentence is equal to the sentence before the correction: <>

Figure 7: Task-specific Few_shot Evaluation prompt for CSC task. We have marked key information in italics and bold, originally incorrect words in red, correctly changed words in green, and incorrectly changed words in blue.

B CSC Case

RAG	
Instance	strategy
原始句子:我受到了你们的结婚卡。 预测句子:我受到了你们的结婚卡。	FixedFew_Shot
Original sentence: I suffered your wedding card. Prediction sentence: I suffered your marriage card.	
原始句子:我受到了你们的结婚卡。 RAG辅助语句: 我很高兴受到你们结婚的邀请单，你们到底决定结婚了！->>我很高兴收到你们结婚的邀请单，你们到底决定结婚了! 预测句子:我收到了你们的结婚卡。	RAGFew_Shot
Original sentence: I suffered your wedding card. RAG Auxiliary Example: I'm glad to have suffered your invitation to get married, you've finally decided to get married! ->>I'm glad to have received your invitation to get married, you've finally decided to get married! Prediction sentence: I received your marriage card.	
IDS	
原始句子:我觉得我们湾很好，我们会唱歌，我们也做饭。 预测句子:我觉得我们湾很好，我们会唱歌，我们也做饭。	FixedFew_Shot
Original sentence: I think our bay is good. We can sing, we also cook. Prediction sentence: I think our bay is good. We can sing, we also cook	
原始句子:我觉得我们湾很好，我们会唱歌，我们也做饭。 中间句子:我觉得我们湾很好，我们会唱歌，我们也做饭。 纠错模块评价：修改后的句子更加合理：否 中间句子:我觉得我们湾很好，我们会唱歌，我们也做饭。 纠错模块评价：修改后的句子更加合理：否 中间句子:我觉得我们玩很好，我们会唱歌，我们也做饭。 纠错模块评价：修改后的句子更加合理：是 预测句子:我觉得我们玩很好，我们会唱歌，我们也做饭。	IDS+FixedFew_Shot
Original sentence: I think our bay is good. We sang, we also cook. Intermediate sentence:I think our bay is good. We sang, we also cook. Evaluator comment:The revised sentence is more reasonable: No Intermediate sentence:I think our bay is good. We sang, we also cook. Evaluator comment:The revised sentence is more reasonable: No Intermediate sentence:I think our playing is good. We sang, we also cook. Evaluator comment:The revised sentence is more reasonable: Yes Prediction sentence:I think our playing is good. We sang, we also cook.	

Table 4: Result examples of CSC experiment on SIGHAN15 with RAG and IDS method. We highlight incorrect characters in red, those remaining wrong after correction in blue, and those right after correction in green. The term Simple Few_Shot denotes no additional policies, RAG Few_Shot signifies the experiment with only RAG method, and IDS+Fixed Few_Shot indicates the experiment with only IDS method.

C Datasets for RAG

Name	Size	Discription	url
CSCD-IME(Train)	30000	consists of blogs posted on Weibo by a certified news media organization,The dataset only focuses on errors caused by "Pinyin Input Method".	https://github.com/nghuyong/cscd-ns?tab=readme-ov-file
SIGHAN13(Train)	700	consists of sentences written by Chinese students aged 13 to 14 in a language examination	http://ir.itc.ntnu.edu.tw/lre/sighan7csc.html
SIGHAN14(Train)	3437	consists of articles written by foreigners in the process of learning Chinese.	http://ir.itc.ntnu.edu.tw/lre/clp14csc.html
SIGHAN15(Train)	2339	consists of articles written by foreigners in the process of learning Chinese.	http://ir.itc.ntnu.edu.tw/lre/sighan8csc.html
ECSpell	572	Open multi-domain spelling correction dataset from Soochow University, including finance and medicine.	https://github.com/Aopolin-Lv/ECSpell
LEMON	21736	A dataset manually created to include seven domains.	https://github.com/gingasan/lemon/tree/main/lemon_v2
MCSC	196497	Medical field dataset, data source is TencentMedical Dictionary	https://github.com/yzhihao/MCSCSet/tree/main/data/mcsc_benchmark_dataset
Wang271k	271282	Based on the concepts of "similar in shape" and "similar in sound," this dataset is primarily used for training models and is typically not used as a test set.	https://github.com/wdimmy/Automatic-Corpus-Generation

Table 5: The CSC dataset used for RAG. Before conducting CSC evaluations, examples in the RAG dataset that have the same answers as the test set were removed.