Knowledge-Augmented Question Error Correction for Chinese Question Answer System with QuestionRAG

Longpeng Qiu[1] **Ting Li**[2] **Shuai Mao**[2] **Nan Yang**[2] **Xiaohui Yan**[2*]
[1]University of Chinese Academy of Sciences
[2]Huawei Technologies Co., Ltd.
qiulongpeng23@mails.ucas.ac.cn
{liting142, maoshuai5, yangnan16, yanxiaohui2}@huawei.com

Abstract

Input errors in question-answering (QA) systems often lead to incorrect responses. Large language models (LLMs) struggle with this task, frequently failing to interpret user intent (misinterpretation) or unnecessarily altering the original question's structure (over-correction). We propose QuestionRAG, a framework that tackles these problems. To address misinterpretation, it enriches the input with external knowledge (e.g., search results, related entities). To prevent over-correction, it uses reinforcement learning (RL) to align the model's objective with precise correction, not just paraphrasing. Our results demonstrate that knowledge augmentation is critical for understanding faulty questions. Furthermore, RL-based alignment proves significantly more effective than traditional supervised fine-tuning (SFT), boosting the model's ability to follow instructions and generalize. By integrating these two strategies, QuestionRAG unlocks the full potential of LLMs for the question correction task.

1 Introduction

The performance and reliability of any Question Answering (QA) system are fundamentally constrained by a critical, universal challenge: errors inherent within the input questions. This is not a minor issue but a central bottleneck that affects systems of all complexities. These pervasive errors originate from two inevitable sources. The first is system-induced, where automated processes like Automatic Speech Recognition (ASR) and Optical Character Recognition (OCR) misinterpret spoken or written words. The second is user-induced, which covers the vast and unpredictable spectrum of human expression, including simple typographical mistakes, the use of regional dialects or colloquialisms, and ambiguous phrasing. An ill-posed question triggers a cascade of failures throughout the entire QA pipeline. It immediately compromises the initial stage of question comprehension, which in turn derails the subsequent analysis and information retrieval, ultimately leading to an incorrect or irrelevant answer. Therefore, the ability to proactively identify and correct these input errors is not just a desirable feature—it is a foundational requirement for building robust, effective, and trustworthy QA systems.

Error type	question example
phonetic error	高四 和李白有什么关系 (高四-> 高适)
graphemic error	OpenAI 01 模型 (01->O1)
missing word	界m7 和 m9 哪个好 (界-> 问界)
wrong order	有哈弗车吗纯电版（哈弗车有纯电版吗）
repeating word	哈哈马斯 的状况 (哈哈马斯-> 哈马斯)
ill-expression	手机SM 码 (SM->SN)

Table 1: Typical error types of the questions in Chinese QA systems. The underlined text is miss-spelling. The parenthetical text describes how to correct the error.

The challenge of question correction is rooted in the diverse nature of errors, which can range from simple typographical mistakes to complex grammatical and semantic inaccuracies, as shown in Table 1. Traditional methods, often built upon foundational models like BERT (Devlin et al., 2019) and sequence-to-sequence architectures (Sutskever et al., 2014), have attempted to address this by creating specialized solutions. These approaches, however, are heavily reliant on explicit rule sets, hand-crafted features, and bespoke model architectures tailored to specific error types, such as phonetic-based correctors or grammatical taggers (Hong et al., 2019; Zhang et al., 2021; Omelianchuk et al., 2020). While effective for their intended narrow tasks, this specialization inherently limits their generalizability and scalability (Li et al., 2024). Consequently, there is a compelling need to develop a universal correction

*Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1533–1542
November 4-9, 2025

framework capable of addressing a broad spectrum of errors without requiring task-specific customization.

Large Language Models (LLMs) represent a promising paradigm for question error correction, capable of addressing diverse error types within a single, unified framework. Building on this promise, a line of recent work has emerged to apply LLMs to this challenge. (Li et al., 2023) first investigates the effectiveness of GPT-3.5 on Chinese Text Correction, and reveals that LLMs handle fluency well and are more fault-tolerant to input data quality, but still have a noticeable gap when compared to previous state-of-the-art, fine-tuned smaller models. Their analysis result identified the main reasons:

- While LLMs possess vast general knowledge, they often fail when it comes to highly specific or nuanced errors. This is particularly evident when correcting errors involving proper nouns or domain-specific terminology. As shown in Table 2, the LLM could not distinguish between " 市大" (Shìdà - City University) and " 师大" (Shdà - Normal/Teacher's University). This correction requires specific background knowledge about place names that a general-purpose LLM may not have prioritized. Moreover, as a text-only model, LLMs cannot access information about a character's pronunciation (e.g., Pinyin) or its visual structure, which is a major handicap for identifying and correcting these types of errors. We collectively refer to the problem of a large model being unable to or incorrectly understanding a user's query due to insufficient background knowledge as *misinterpretation*.

- LLMs are trained to generate the most probable or fluent text based on their massive training data. This causes them to "correct" expressions that are grammatically valid but less common, replacing them with more conventional phrases (Fang et al., 2023; Wu et al., 2023). As shown in Table 2, the LLM changed "remarried wife" (再婚妻子) to "ex-wife" (前妻). In the context of question correction, this behavior is a form of error. We refer this problems as *over-correction*.

To address the misinterpretation and over-correction issues of LLMs when applying to Chinese question correction, we propose a knowledge-augmented LLM approach, called QuestionRAG, which can handle various types of errors in a simple and unified way. Specifically, to overcome the misinterpretation problem, QuestionRAG leverages Retrieval-Augmented Generation (RAG) (Lewis et al., 2021) by introducing rich external knowledge for each question, such as the search results, entity description, similar questions. As the retrieval process takes into account a combination of textual, visual, and phonetic similarities, QuestionRAG enables the LLM to generate a correction based solely on the retrieved context, eliminating the need for the LLM itself to possess visual or phonetic discrimination capabilities.

To further address the over-correction problems, we designed a reinforcement learning training method for the question error correction task. This method can automatically stimulate the LLM's intermediate thought processes for question correction, reducing the reliance on annotated samples for the reasoning process.

Experimental results show that the model trained with RL significantly outperforms SFT-trained models, and the over-correction problem is substantially mitigated.

Error Type	example	question
misinterpretation	input	湖南市大 怎么走
	ground truth	湖南师大 怎么走
	correction	湖南四大怎么走
over-correction	input	摩克多 再婚妻子
	ground truth	默克多 再婚妻子
	correction	默克多前妻

Table 2: Examples of the misinterpretation and over-correction problems in LLM-based question correction. In the first example, the LLM misinterpreted the term " 湖南市大" as " 湖南四大", but the " 湖南四大" does not exit in the real word. Instead, the correct should be " 湖南师大" , which is a normal university and pronounced exactly the same as the user's original input. In the second example, in addition to correcting the erroneous text (" 摩克多" to " 默克多"), the LLM also change the correct text from " 再婚妻子" to " 前妻", a needless modification that changed the meaning of the original question.

2 Related Work

Chinese text Correction (CTC) is a challenging task since thousands of characters that can have

similar pronunciations or visual appearances. In question answer system, errors in user questions are highly varied—stemming from colloquial language, input devices, and environmental conditions—making the task significantly more challenging.

Most conventional methods are tailored to address specific error types with small models, such as BERT(Devlin et al., 2019) and Seq-to-Seq models(Sutskever et al., 2014). For instance, SpellBERT (Ji et al., 2021) is a lightweight pre-trained model utilizing BERT with additional graph-based visual and phonetic features for Chinese Spelling Check. Similarly, other BERT-based approaches (Hong et al., 2019; Zhang et al., 2020; Cheng et al., 2020) have integrated phonological and morphological knowledge into contextual embeddings in BERT-like models for identifying and correcting textual errors. These task-specific models are limited by their parameter scale and capabilities, thereby underperform in low-frequency and complex semantic scenarios.

While recent Large Language Models (LLMs) show promise for this task, standard in-context learning approaches that directly prompt the model for corrections (Li et al., 2023; Sun et al., 2024) have performed poorly, even underperforming traditional small models (Li et al., 2023; Zhou et al., 2024b). More effective strategies include task-specific pre-training (Zhou et al., 2024b) and fine-tuning (Li et al., 2023). (Zhou et al., 2024a) combines an LLM with a Minimal Distortion Model for scoring corrections, but this is limited to candidates of identical length to the error. Addressing this, (Liang and Zhou, 2025) proposed a multi-turn CTC framework to maintain length consistency. Nevertheless, all of these methods still suffer from problems of misinterpretation and over-correction.

Recently, few work has begun to explore the integration of RAG (Lewis et al., 2021) into LLM-based text correction. These approaches typically enhance model performance by retrieving relevant data, such as domain-specific entities (Pusateri et al., 2025) or labeled corrections from existing training sets (Liang and Zhou, 2025). However, the applicability of them is often limited by their reliance on curated, domain-specific knowledge. In contrast, our method leverages broad, general knowledge drawn from public sources like the web and Wikipedia, supplemented by unlabeled questions. Furthermore, to the best of our knowledge, our work is the first to integrate RAG with reinforcement learning (RL) to align an LLM specifically for the text correction task. This novel combination successfully alleviates the aforementioned problems of both misinterpretation and over-correction.

3 Methodology

We first introduce the QuestionRAG framework, and then discuss how reinforcement learning is utilized within it to address the challenge of over-correction.

3.1 QuestionRAG

Most prior research on LLM-based Chinese text correction has focused on enhancing the model's ability to understand and generate text based on character graphemes (Zhang et al., 2025) and phonetics (Yamashita et al., 2025). This task is inherently challenging, as large models are primarily trained for text generation. In contrast, the core idea of QuestionRAG is to fully leverage the LLM's strengths in text comprehension and generation, while offloading the processing of graphemes and phonetics—areas where LLMs are less proficient—to the retrieval stage. By integrating similarity factors such as string overlap, semantics, graphemes, and phonetics during search and ranking, QuestionRAG gathers relevant texts to serve as context, providing valuable guidance for the LLM to infer the correct correction.

We introduce general knowledge for a question by retrieval relevant texts from the following sources:

- **Web**: We take the question as a query, and obtain the search results form a commercial search engine. For simplification, we only kept the titles of relevant web pages as context.
- **Log of the QA system**: We search similar questions from historical questions extracted from the logs of a commercial question answer system. To filter low-quality questions, we filter ones with frequency less than 5.
- **Wikipedia**: we gather entities and their description from Wikipedia. To retrieval relevant entities, we first extract candidate entities from the input questions, and then search the entity base with these candidate entities.

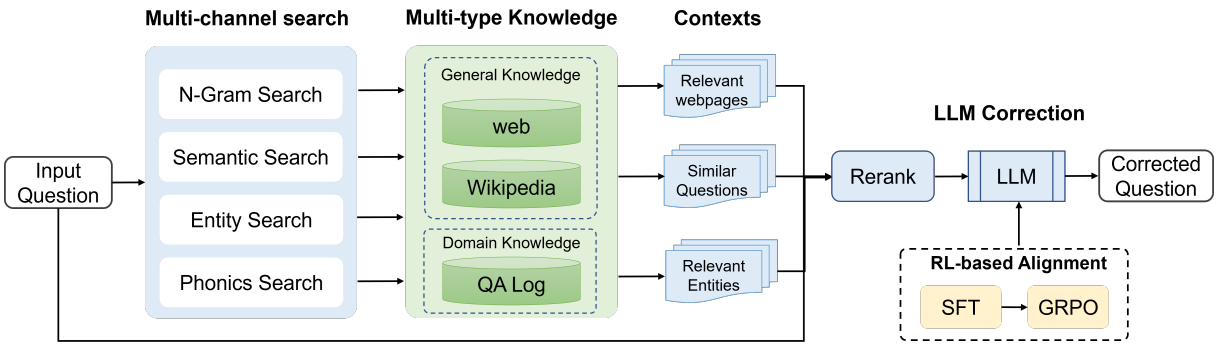

Figure 1: The workflow of QuestionRAG includes a search stage collecting relevant webpages, questions, and entities from external knowledge sources. It utilizes multi-facet search (n-gram for lexical similarity, embedding from semantic similarity, entity for conceptual similarity, and Pinyin for phonics similarity, etc.) from multiple knowledge sources (either general or domain-specific). With the search results as augmented knowledge, a LLM trained with reinforcement learning is utilized to generate the correction.

QuestionRAG utilizes a multi-channel search system to handle multiple knowledge sources, as shown in Figure 1. Its indexing stage employs a multi-faceted strategy combining three methods: (1) a ngram-based inverted index for lexical matching, (2) semantic vector similarity for conceptual retrieval, and (3) a Pinyin-based inverted index for phonetic matching. After an initial retrieval pass, the GTE-multilingual-reranker(Zhang et al., 2024) is used to semantically filter and re-rank all candidate texts based on their relevance to the question. The resulting top-ranked texts then serve as the contextual input for the subsequent generation stage.

3.2 Model Training with RL

Following the paradigm of deepseek R1(Guo et al., 2025), we adopt a two-stage training strategy consisting of Supervised Fine-Tuning (SFT) for cold-start and GRPO-based reinforcement learning for post-training.

3.2.1 Cold-Start SFT

To prepare our training data, we curated a small-scale, high-quality set by sampling both erroneous and error-free questions from QA logs. We manually corrected the erroneous samples to create their ground truth labels. Each question was then augmented with reasoning traces generated by deepseek R1(Guo et al., 2025). During this process, any trace that led to an incorrect correction was discarded. Finally, to ensure the quality of the dataset, every training sample underwent a thorough manual review.

Instead of directly generating the corrected text, in cold-start SFT, the LLM first reason if the question contains an error, and then identifies the error's location if erroneous. Only after outputting this complete reasoning process does the model provide the final, corrected result.

3.2.2 GRPO Training

To enhance the model's generalization and reasoning, we employ Group Relative Policy Optimization (GRPO) (Guo et al., 2025) in a post-training phase. This resource-efficient method eliminates the need for a separate evaluator model by directly comparing candidate responses, with each response scored by a predefined reward function. The design of this reward function is therefore critical for guiding the model toward high-quality, accurate corrections. Our reward function is built on two key components:

- **Format Rewards.** This component systematically verifies whether model-generated outputs strictly adhere to predefined structural templates: <think>{reasoning process}</think>\n\n{final answer}. The reward function of output $\boldsymbol{c}$ is defined as:

$$R_f(\boldsymbol{c}|\boldsymbol{q}) = \begin{cases} 1 & \text{if } \mathbf{c} \text{ format adhered} \\ 0 & \text{otherwise} \end{cases}$$

- **Accuracy Rewards.** The accuracy reward is designed for a nuanced evaluation of the generated correction against the ground truth. Drawing inspiration from the use of Character Error Rate (CER) in error correction(Leng et al., 2022), our function leverages a normalized edit distance(Wang et al., 2024) to quantify the similarity of output $\boldsymbol{c}$ to the target $\boldsymbol{g}$,

defined as:

$$R_a(\boldsymbol{c},\boldsymbol{g}|\boldsymbol{q})=\begin{cases}2.0 & \text{if } \boldsymbol{c}=\boldsymbol{g}\\ (1-d_c)+\beta(1-d_c)^2 & \text{if} d_c<d_q\\ -\lambda\cdot d_c & \text{if} d_c>d_q\\ 0.0 & \text{otherwise}\end{cases}$$

where d_q and d_c represent the normalized edit distances between the ground truth $\boldsymbol{g}$ and the input question $\boldsymbol{q}$, and the model's output $\boldsymbol{c}$, respectively. A smaller edit distance indicates a closer proximity to ground truth. β is a hyperparameter that scales a non-linear bonus for improvements. λ is a hyperparameter that scales the penalty for regressions.

The overall reward function summarizes the format reward and the accuracy reward, i.e.,

$$R(\boldsymbol{c},\boldsymbol{g}|\boldsymbol{q}) = R_f(\boldsymbol{c}|\boldsymbol{q}) + R_a(\boldsymbol{c},\boldsymbol{g}|\boldsymbol{q}).$$

The reward function described above encourages the gradual optimization of correction results during the RL training process, while penalizing incorrect corrections.

4 Experiments

In this section, we present the specific experiments conducted for error correction tasks.

4.1 Datasets

QCSet: QCSet is a dataset derived from the logs of a commercial question-answering system. It comprises approximately 10,000 error-correction pairs, meticulously annotated by human experts to ensure high accuracy.

MCSCSet: MCSCSet(Jiang et al., 2022) is an open-source dataset focused on medical Chinese spelling correction. It is a large-scale dataset containing about 200k samples, manually annotated by medical specialists.

Qspell: Qspell(Ye et al., 2025) is a public-domain dataset of approximately 250k error-correction pairs. It covers lots of topics, like formal terms, casual speech, and idioms, in both Chinese and English.

4.2 Evaluation Metrics

Character Error Rate. To quantitatively evaluate the performance of our error correction model, we employed the Character Error Rate (CER). CER is a widely used metric for assessing the accuracy of text transcription and correction tasks. It measures the minimum number of edits—such as insertions, deletions, or substitutions—required to align the generated output into the ground truth.

4.3 Implementation Details

We utilized Qwen3-8B(Yang et al., 2025) as our base model and trained it on the proprietary dataset QCSet. Specifically, for the cold-start phase, we used around 1,000 QCSet samples enhanced with Chain-of-Thought reasoning generated via DeepSeek R1 distillation(Guo et al., 2025). For the GRPO post-training phase, we employed an additional 7,000 QCSet samples, ensuring no overlap with the cold-start phase. This two-stage approach allowed us to progressively enhance the model's capabilities. To rigorously evaluate the model's performance, we tested it on three distinct test sets, each comprising around 1,000 randomly selected samples, with approximately 90% of the samples requiring correction. These test sets were carefully chosen to ensure no overlap with the training data, providing a comprehensive assessment of the model's error correction capabilities.

In terms of training details, during the cold-start stage, we performed full parameter fine-tuning with a learning rate of 2e-6 and a total batch size of 8. For the GRPO training stage, the learning rate is 5e-6 and batch size is 32. The hyperparameters in our reward function, β and λ, were both set to 1.0 throughout our experiments.

We also detail the computational requirements and training load for the post-training process, addressing the complexity of implementation. Our training, consisting of Cold-Start SFT and GRPO, was executed on Ascend 910B NPUs. The Cold-Start SFT stage involved training the model with 1k samples for 4 epochs on 8 NPUs, costing approximately 50 minutes. The subsequent GRPO phase, which represents the primary contributor to the total computational load, required training with 7k samples for 12 epochs on 32 NPUs, taking approximately 18 hours.

4.4 Results

4.4.1 In-Context Learning (ICL) Result

To verify the effectiveness of knowledge argumentation of LLM on the question correction task, we evaluate QuestionRAG without any fine-tuning (referred to as QuestionRAG$_{ICL}$) by adding different types of knowledge into prompt, as de-

Configuration	QCSet	MCSCSet	QSpell
Original Question	16.13	16.35	14.78
No-RAG	15.37	10.77	10.71
QuestionRAG$_{ICL}$(Entity)	14.31	9.91	10.5
QuestionRAG$_{ICL}$(Entity + Web Info)	13.6	7.8	9.04
QuestionRAG$_{ICL}$(Entity + Web Info + Similar Questions)	**12.53**	**7.14**	**8.63**

Table 3: Comparison of QuestionRAG$_{ICL}$ with different types of knowledges(CER %). No-RAG refers denotes Qwen3-8B without any knowledge argumentation.

Question	Similar Web page titles	Similar Questions	Related Entities	Correction
湖南市大怎么走	湖南大学正门怎么走 湖南市政府怎么走	**到湖南师大怎么走** **到湖南师范大学怎么走**	**湖南师大: 湖南师范大学**	湖南师大怎么走
摩克多再婚妻子	**93 岁默克多再婚** 摩克多塔拉	摩克多检测	**默克多: 默克多也就是鲁伯特·默多克**	默克多再婚妻

Table 4: Two examples of question correction by QuestionRAG. Bold text represents the useful information that has been introduced.

Approach	QCSet	MCSCSet	QSpell
Original Question	16.13	16.35	14.78
Fine-tuned T5	13.08	7.8	13.65
QuestionRAG$_{ICL}$	12.53	7.14	8.63
QuestionRAG$_{SFT}$	11.19	5.43	8.28
QuestionRAG$_{GRPO}$	**9.04**	**5.08**	**6.37**

Table 5: Comparing of SFT and GRPO against the other baselines on Chinese question correction(CER %).

tailed in Appendix A. From Table 3, it is clear that progressively enhancing the RAG framework with additional knowledge sources consistently improved error correction performance on all the three dataset. Firstly, related entity information provided relevant or homophonic named entities and descriptions, which allowed the model to understand related concepts, especially by supplementing knowledge of rare entities. Secondly, the inclusion of web information provided real-time, diverse data, enhancing the model' s ability to handle novel errors. Finally, the introduction of similar questions from Our QA System also significantly improved error correction on the two open-source datasets. These findings suggest that external knowledge from either general or domain sources is crucial for LLM on the question correction task.

4.4.2 Post Training Results

We further aligned the LLM in QuestionRAG for question correction task with reinforcement learning. Its results are shown in Table 5. QuestionRAG$_{GRPO}$ achieved the lowest CER on all the datasets, while QuestionRAG$_{SFT}$ slightly outperforms QuestionRAG$_{ICL}$. It suggests that both SFT and reinforcement learning alignment can further improve the effectiveness of QuestionRAG. Compared with SFT alone, reinforcement learning with GRPO exhibits stronger generalization capability. Table 4 presents the relevant knowledge introduced for the incorrect questions, along with the useful information and the final corrected results. As can be seen, introducing additional context significantly reduces the difficulty for the LLM to perform the correction.

Beyond the quantitative improvements, we observed a surprising and critical qualitative advantage: QuestionRAG$_{GRPO}$ exhibits superior robustness even when faced with highly misleading and erroneous retrieval context. Table 6 clearly demonstrates this capability. The model successfully navigates a scenario where the majority of the entire retrieved context is erroneous and collectively favors common, yet wrong interpretations. The ability of QuestionRAG$_{GRPO}$to successfully filter this dominant, misleading information and adhere to the user's likely intent confirms that our approach guides the model toward a more reasoned and reliable correction trajectory.

4.4.3 The impact of Model Scale

We investigated the impact of model size within the QuestionRAG framework using various Qwen3 models. The results, presented in Figure 2, show that larger LLMs have a clear advantage over smaller ones, though this performance gap narrows as model size increases. This is because while all models share the same external knowledge provided by the RAG component, larger LLMs also possess more internal knowledge and stronger inherent capabilities for linguistic under-

Question	Similar Web page titles	Similar Questions	Related Entities
终南山是谁 *(Ground Truth:* 钟南山是谁*)*	我如何来到终南山? 谁都想到终南山隐居 终南山的开派祖师是谁	终南山王维 西安终南山 终南山	**钟南山**: 抗击非典、新冠疫情的领军人物 终南山: 终南山是秦楚古道的所在地 终南山:《终南山》是王维创作的一首五律

QuestionRAG$_{ICL}$: 终南山是谁 (Incorrect); QuestionRAG$_{GRPO}$: **钟南山是谁** (*Correct*)

Table 6: Robustness case study: QuestionRAG$_{GRPO}$ successfully corrects ambiguous queries despite highly noisy contextual information.

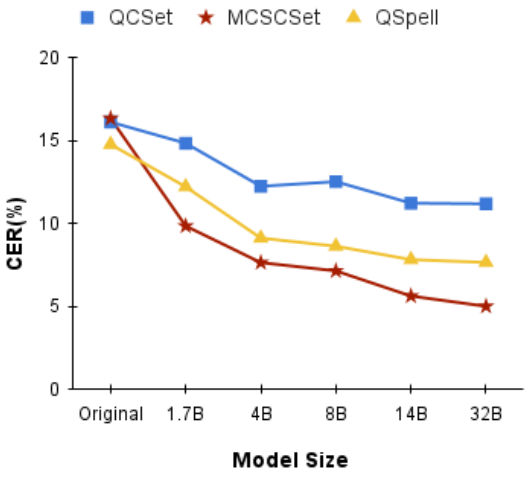

Figure 2: Impact of model size on CER(%) for Qwen3 models (1.7B to 32B) within the QuestionRAG framework, illustrating how increasing model size affects performance.

standing and reasoning.

Moreover, a key finding demonstrates the power of knowledge augmentation: a smaller model equipped with QuestionRAG can outperform a much larger model that lacks it. For instance, a comparison between the "No RAG" baseline (Qwen3-8B) in Table 3 and the 1.7B parameter QuestionRAG model in Figure 2 reveals that the smaller model with QuestionRAG often achieves superior performance, highlighting that external knowledge can be more impactful than simply increasing model scale.

5 Conclusion

This paper introduces QuestionRAG, a novel framework that significantly enhances language model performance by addressing two of their core failures: misinterpretation and over-correction. Its primary innovation lies in a dual-pronged strategy: it leverages Retrieval-Augmented Generation to provide essential external context while employing Reinforcement Learning to meticulously align model behavior. This approach allows QuestionRAG to overcome the inherent deficiencies of LLMs in understanding character graphemes and phonetics, enabling it to fully exploit their strengths in text comprehension and generation within a single, unified framework that handles diverse error types. Beyond simple question correction, its methodology is directly applicable to broader tasks such as question rewriting, planning, and enhancing query understanding in search engines, opening new avenues for future research.

Limitations

The performance of QuestionRAG is critically dependent on the quality of its search results. If the retrieved information is irrelevant or noisy, the accuracy of the final output will be compromised. Furthermore, because the RAG process significantly increases prompt length, it also increases latency, particularly the time to first token.

References

Xingyi Cheng, Weidi Xu, Kunlong Chen, Shaohua Jiang, Feng Wang, Taifeng Wang, Wei Chu, and Yuan Qi. 2020. SpellGCN: Incorporating phonological and visual similarities into language models for Chinese spelling check. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 871–881, Online. 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 *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *Preprint*, arXiv:2304.01746.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong

Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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 *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 160–169, Hong Kong, China. Association for Computational Linguistics.

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.

Wangjie Jiang, Zhihao Ye, Zijing Ou, Ruihui Zhao, Jianguang Zheng, Yi Liu, Bang Liu, Siheng Li, Yujiu Yang, and Yefeng Zheng. 2022. Mcscset: A specialist-annotated dataset for medical-domain chinese spelling correction. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*.

Yichong Leng, Xu Tan, Wenjie Liu, Kaitao Song, Rui Wang, Xiang-Yang Li, Tao Qin, Ed Lin, and Tie-Yan Liu. 2022. Softcorrect: Error correction with soft detection for automatic speech recognition. In *AAAI Conference on Artificial Intelligence*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Kunting Li, Yong Hu, Liang He, Fandong Meng, and Jie Zhou. 2024. C-llm: Learn to check chinese spelling errors character by character. *arXiv preprint arXiv:2406.16536*.

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

Junhong Liang and Yu Zhou. 2025. Mtcsc: Retrieval-augmented iterative refinement for chinese spelling correction. *Preprint*, arXiv:2504.18938.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

Ernest Pusateri, Anmol Walia, Anirudh Kashi, Bortik Bandyopadhyay, Nadia Hyder, Sayantan Mahinder, Raviteja Anantha, Daben Liu, and Sashank Gondala. 2025. Retrieval augmented correction of named entity speech recognition errors. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *Preprint*, arXiv:1409.3215.

Yi-Cheng Wang, Hsin-Wei Wang, Bi-Cheng Yan, Chi-Han Lin, and Berlin Chen. 2024. Dancer: Entity description augmented named entity corrector for automatic speech recognition. *ArXiv*, abs/2403.17645.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark. *Preprint*, arXiv:2303.13648.

Natsuo Yamashita, Masaaki Yamamoto, Hiroaki Kokubo, and Yohei Kawaguchi. 2025. Llm-based generative error correction for rare words with synthetic data and phonetic context. *Preprint*, arXiv:2505.17410.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Dezhi Ye, Haomei Jia, Junwei Hu, Tian Bowen, Jie Liu, Haijin Liang, Jin Ma, and Wenmin Wang. 2025. QSpell 250K: A large-scale, practical dataset for Chinese search query spell correction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 148–155, Albuquerque, New Mexico. Association for Computational Linguistics.

Ruiqing Zhang, Chao Pang, Chuanqiang Zhang, Shuohuan Wang, Zhongjun He, Yu Sun, Hua Wu, and Haifeng Wang. 2021. Correcting Chinese spelling errors with phonetic pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2250–2261, Online. Association for Computational Linguistics.

Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li. 2020. Spelling error correction with soft-masked bert. In *Annual Meeting of the Association for Computational Linguistics*.

Xiaowu Zhang, Hongfei Zhao, Jingyi Hou, and Zhijie Liu. 2025. Unveiling the impact of multimodal features on chinese spelling correction: From analysis to design. *Preprint*, arXiv:2504.07661.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, and 1 others. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.

Houquan Zhou, Zhenghua Li, Bo Zhang, Chen Li, Shaopeng Lai, Ji Zhang, Fei Huang, and Min Zhang. 2024a. A simple yet effective training-free prompt-free approach to chinese spelling correction based on large language models. *Preprint*, arXiv:2410.04027.

Jie Zhou, Fandong Meng, Liang He, Yong Hu, and Kunting Li. 2024b. C-llm: Learn to check chinese spelling errors character by character. *Preprint*, arXiv:2406.16536.

A Prompt for Error Correction

The complete version of the prompt used in our experiments is provided in the following. It outlines the instructions and principles for question error correction, including the use of retrieved information and strict adherence to correction rules.

Prompt

You are a meticulous proofreading assistant. You need to combine your own knowledge with retrieved information (including similar questions, web titles, and entity information) to determine whether the user's original query requires correction. If no correction is needed, output the user's original query; otherwise, output the corrected query, strictly adhering to the following principles. Any changes must be based on phonetic similarity and preserve the framework of the user's original query.
Correction Principles (Strict Priority, Check from 1 to 4):
1. Minimal Modification Principle (Core): Only modify clearly erroneous parts of the user's original query (e.g., spelling errors, homophonic typos). Some special parts (including punctuation, spaces, capitalization, and word order) must remain unchanged. The corrected output should be similar in length and structure to the original query, with minimal changes. If the error cannot be identified, do not correct.
2. Corrections must strictly follow homophonic or near-homophonic rules (i.e., the corrected result must have the same or similar pronunciation as the original part). If an error is evident but does not satisfy the homophonic/near-homophonic rule, do not correct and output the user's original query.
3. Cautious Reference Principle (Supplementary Reference): Retrieved information (similar questions, web titles, entity information) is for reference only and may contain errors (especially similar questions, which may share the same mistakes). Use your own language knowledge to evaluate, and ignore retrieved information if it violates Principle 1 or 2 (e.g., suggests significant changes or alters pronunciation).
4. Semantic Preservation Principle (Non-Optimization Principle): Ensure the corrected query retains the same semantic meaning as the original, without altering the user's intent or structure for the sake of fluency. Only fix errors. If the intent is ambiguous, has multiple possibilities, or correction would distort the original meaning, do not correct (output the user's original query). Do not add or remove words to make the sentence more fluent.
Output Requirement: Directly output the corrected query or the user's original query without explanation.

Retrieved Information:
- Similar questions:
{similar questions}
- Related Web Titles:
{web titles}
- Related Entity Information:
{entity details}
User's Original Query:
{query}
Output the corrected result