QSpell 250K: A Large-Scale, Practical Dataset for Chinese Search Query Spell Correction

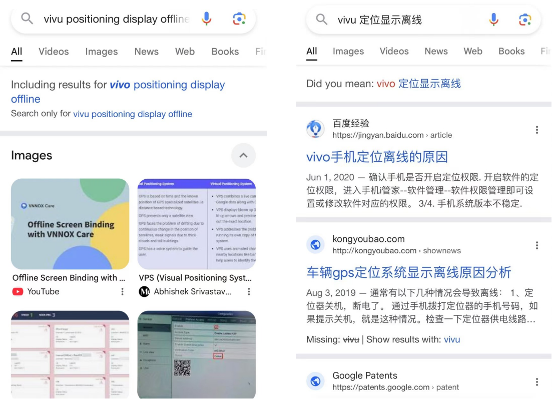
Dezhi Ye[1] **Haomei Jia**[2] **Junwei Hu**[1] **Bowen Tian**[1]
Jie Liu[1] **Haijin Liang**[1] **Jin Ma**[1] **Wenmin Wang**[2]
[1]Tencent [2]Macau University of Science and Technology
`{dezhiye,keewayhu,lukatian,jesangliu,hodgeliang,daniellwang}@tencent.com`
`3220003442@student.must.edu.mo` `wmwang@must.edu.mo`

Abstract

Chinese Search Query Spell Correction is a task designed to autonomously identify and correct typographical errors within queries in the search engine. Despite the availability of comprehensive datasets like Microsoft Speller and Webis, their monolingual nature and limited scope pose significant challenges in evaluating modern pre-trained language models such as BERT and GPT. To address this, we introduce **QSpell 250K**, a large-scale benchmark specifically developed for simplified Chinese Query Spelling Correction. QSpell 250K offers several advantages: 1) It contains over 250K samples, which is ten times more than previous datasets. 2) It covers a broad range of topics, from formal entities to everyday colloquialisms and idiomatic expressions. 3) It includes both Chinese and English, addressing the complexities of code-switching. Each query undergoes three rounds of high-fidelity annotation to ensure accuracy. Our extensive testing across three popular models demonstrates that QSpell 250K effectively evaluates the efficacy of representative spelling correctors. We believe that QSpell 250K will significantly advance spelling correction methodologies. The accompanying data and code will be made publicly available[1].

1 Introduction

Query Spelling Correction is essential for enhancing the efficacy of search engines by identifying and rectifying errors in user queries (Sharma et al., 2023; Yang et al., 2022). A misspelled search query can yield irrelevant results, thereby diminishing the user's ability to obtain satisfactory outcomes (Gong et al., 2019; Fourney et al., 2017; Gupta et al., 2019). For instance, given the query "vivu positioning display offline," "vivu" in Figure 1 should be corrected to "vivo". Should the model fail to rectify this, the search results would include "vivu," failing to meet the user's needs. By identifying common spelling errors, search engines can be better equipped to handle these inaccuracies by suggesting corrections or automatically adjusting queries.


vivu positioning display offline
All
Videos
Images
News
Web
Books
Fi
Including results for ***vivo*** positioning display offline
Search only for vivu positioning display offline
Images
VNNOX Care
Offline Screen Binding with VNNOX Care
Offline Screen Binding with ...
YouTube
VPS (Visual Positioning Syst...
Abhishek Srivastav...

(a) vivu positioning display offline

vivu 定位显示离线
All
Images
Videos
News
Web
Books
Fi
Did you mean: vivo 定位显示离线

百度经验
https://jingyan.baidu.com › article
vivo手机定位离线的原因
Jun 1, 2020 — 确认手机是否开启定位权限. 开启软件的定位权限，进入手机i管家--软件管理--软件权限管理即可设置或修改软件对应的权限。 3/4. 手机系统版本不稳定.

kongyoubao.com
http://kongyoubao.com › shownews
车辆gps定位系统显示离线原因分析
Aug 3, 2019 — 通常有以下几种情况会导致离线： 1、定位器关机，断电了。 通过手机拨打定位器的手机号码，如果提示关机，就是这种情况。检查一下定位器供电线路...
Missing: ~~vivu~~ | Show results with: vivu

Google Patents
https://patents.google.com › patent

(b) vivu 定位显示离线

Figure 1: The display format of query corrections on Google involves the search engine automatically correcting a misspelled query to the appropriate term, while simultaneously notifying the user with the prompt "Showing results for".

The field of query spelling correction has garnered considerable interest (Li, 2020), as evidenced by initiatives such as the Microsoft Speller Challenge (MSC) (Wang and Pedersen, 2011) and the wealth of research on the participating methodologies and their subsequent refinements. The MSC provides a corpus of approximately 6,000 annotated queries, of which 16% contain errors. To address the issue of limited data scale, Webis (Hagen et al., 2017) compiled a more extensive collection of 54,772 queries, with 16.74% marked for spelling inaccuracies. While these datasets have propelled advancements in the domain of query spelling correction, they predominantly cater to English, with minimal efforts extended to other

[1] https://github.com/dz1109/CQSpell

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 148–155
April 30, 2025

languages, such as Chinese. Existing datasets like SIGHAN13 (Wu et al., 2013), SIGHAN14 (Yu et al., 2014), and SIGHAN15 (Tseng et al., 2015) provide a Chinese Spelling Correction corpus collected from a computer-based test of Chinese as a foreign language. However, these datasets are primarily aimed at long texts in spelling exams, rather than user searches in the web domain. Additionally, the scale of the datasets is small, with each dataset containing approximately 1,000 entries. The MCSCSet (Jiang et al., 2022) introduces a Chinese corpus focus on medical domain, which unfortunately limits the assessment of error correction models in non-medical contexts. Consequently, there is a pressing need for a comprehensive Chinese Search Query Spelling Benchmark.

Furthermore, in the realm of real-world search engines, users frequently engage in code-switching, interspersing multiple languages within a single search, such as "ipap插入u盘无反应"("ipap insertion of USB drive unresponsive). This widespread practice introduces complex challenges for current correction models, which, when trained in a single language, falter in the presence of multilingual inputs. Thus, spelling correction models must be adept at understanding and processing multiple languages. However, mainstream datasets like MSC, SIGHAN have scarcely included this linguistic phenomenon, underscoring an acute need for a comprehensive dataset that can aid in the evolution of spell correction models capable of handling such linguistic diversity.

To catalyze progress in Query Spelling Correction (QSC), we present a novel benchmark named **QSpell 250K**, a comprehensive Large-scale dataset. The volume of QSpell 250K is four to ten times that of its predecessors, amassing a total of 250,000 meticulously annotated queries. Besides, the queries are cleaned for personal data. Remarkably, over 12% of the queries in QSpell 250K feature code-switching (contains both English and Chinese.), mirroring the linguistic intricacies encountered in real-world contexts. The dataset predominantly comprises queries sourced from an actual search engine, capturing an extensive spectrum of subjects and newly coined internet phenomena. QSpell 250K encompasses five primary categories of errors: phonetic, orthographic, scrambled, omitted, and superfluous characters. The specific error types within QSpell 250K are enumerated in Table 1.

The main contributions of our benchmark are summarized as follows:

- We providea large scale Chinese Search Query Spelling Correction benchmark (**QSpell 250K**) derived from search engines, addressing the gap in the field of Chinese query correction. To ensure the high quality of QSpell 250K, we conduct three rounds of validation, enhancing its reliability and accuracy.

- We conduct a comprehensive study on recent state-of-the-art models, contributing to the advancement of the spell correction domain. By evaluating and analyzing these models within the context of our benchmark, we provide valuable insights and guidance for researchers and practitioners working in the field of spelling correction.

2 Related work

2.1 Datasets for Query Spelling Correction

The field of query spelling correction has garnered considerable interest following the Microsoft Speller Challenge. During this competition, an extensive public dataset comprising 5,892 spell-corrected queries, extracted from the TREC archives, was unveiled for training purposes. Subsequently, qSpell (Ganjisaffar et al., 2011) contributed an additional training set encompassing 6,000 queries. Augmenting the publicly accessible corpora, Webis released a substantial dataset of 54,772 queries, with a notable 16% containing spelling errors. Existing query correction models are predominantly evaluated using these three datasets. However, they are tailored exclusively for English, presenting challenges in assessing models designed for Chinese spell correction. In the realm of Chinese, the MCSCSet (Jiang et al., 2022) offers a repository for short text spell correction, albeit limited to the medical field and featuring a narrow range of error types. To address this gap, we have developed a comprehensive, multi-faceted benchmark tailored for query spell correction.

2.2 Approaches for Query Spelling Correction

A query corrector is essential for enhancing the relevance of web searches within search engines (Li et al., 2006; Ahmad and Kondrak, 2005; Gao et al., 2010). Initial studies on Query Spelling Correction (QSC) typically framed the issue within the context of a noisy channel model (Chen et al., 2007; Duan et al., 2012; Sun et al., 2012). Subsequent

Language	Category	Typos	Text	Translation
Chinese	Phonetic	小金菊怎么治咳嗽	小金桔怎么治咳嗽	How to cure cough with kumquat
	Visual	淮剧连花庵全集	淮剧莲花庵全集	The Lotus Ann of Huaiju Drama
	Order	岳云鹏声相	岳云鹏相声	Yue Yunpeng's comedy
	Missing	王者耀刘备	王者荣耀刘备	Arena Of Valor Liu Bei
	Redundant	飞天茅茅台鉴定方法	飞天茅台鉴定方法	Feitian Moutai identification method
English	Phonetic	iphome如何看海拔	iphone如何看海拔	How to watch elevation on iphone
	Visual	vaccum seal 怎么用	vacuum seal怎么用	How to use vacuum seal
	Order	levaes英语怎么读	leaves英语怎么读	How to pronounce leaves in English
	Missing	假面骑士amzons	假面骑士amazons	Kamen Rider amazons
	Redundant	windowss10电脑屏幕	windows10电脑屏幕	windows 10 computer screen

Table 1: Examples of different types of edits in QSpell 250K that involve both Chinese and English languages.

approaches have employed Statistical Machine Translation-based models to address the contextual limitations inherent in error modeling(Hasan et al., 2015). In our study, we classify spelling correction models into three principal categories according to their architectural framework. Decoder-only models (Zhang et al., 2023b), represented by the pre-trained GPT2, are adaptable for sequence generation tasks through fine-tuning. Encoder-Decoder models (Pande et al., 2022; Zhang et al., 2023a; Kuznetsov and Urdiales, 2021),, such as T5 (Kakkar et al., 2023), are adept at encoding queries and subsequently generating the correct targets. Text edit models (Mallinson et al., 2022), like KSTEM (Ye et al., 2023), reconceptualize the sequence generation challenge as a sequence tagging task, with the objective of diminishing latency. Although these models have demonstrated enhanced performance on the MSC dataset, there is still an absence of a rigorous benchmark for QSC.

3 Chinese Search Query Spell Correction Benchmark

3.1 Task Definition

Given an incorrect query $\boldsymbol{x} = \{x_1, x_2, \ldots, x_i\}$, and a correct query $\boldsymbol{y} = \{y_1, y_2, \ldots, y_j\}$, the Chinese search query spelling correction task can be defined as $f : \boldsymbol{x} \rightarrow \boldsymbol{y}$, where f denotes the model to automatically convert the query $\boldsymbol{x}$ to another query $\boldsymbol{y}$. It should be noted that the length of sentences $\boldsymbol{x}$ and $\boldsymbol{y}$ may not be equal, reflecting the presence of missing or redundant errors in real-world scenarios.

3.2 Query Sampling

The first stage of the construction of QSpell 250K is to collect error query candidates. In a real-world search engines, the proportion of error queries is relatively small. In other words, if we do random sampling, most of the queries we get are correct.

To build a large-scale query spelling correction benchmark, we sample 5,000,000 queries with 2 up to 40 characters from the query log in an industrial web search engine. Queries that are excessively lengthy or brief are filtered out. These 5,000,000 queries do not constitute the final dataset for annotation. Rather, they represent the initial set of raw data that requires filtering and screening.

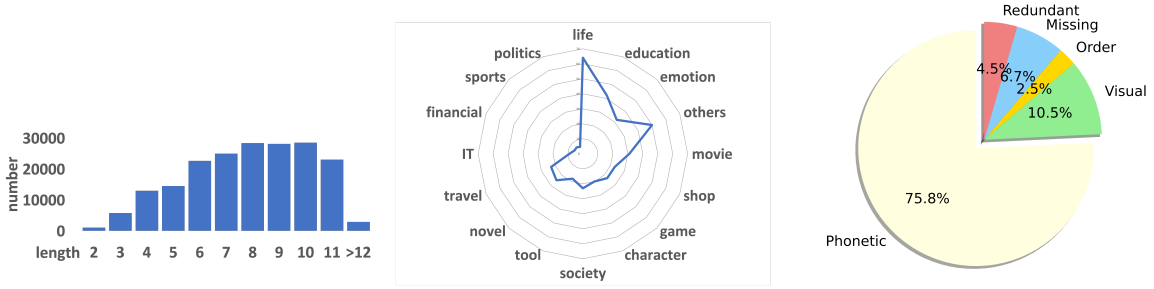
- Step 1. We collect the query log from January 2023 to December 2023 and compute the query search frequency and click rate (query click number / query search frequency). These data were collected from an industrial web search engine.

- Step 2. We remove queries that include personal information, toxic topics. In addition, we further filter out queries with more than 40 or less than 2 characters.

- Step 3. We remove queries with the top search frequency and click rate. A simple assumption is that high-frequency queries are less likely to contain errors. In addition, if the user cannot find a satisfactory result, the click action will not occur.

- Step 4. We select candidates to be annotated after corpus matching and perplexity (PPL) value filtering, which is calculated by the language model [2]. A query exhibiting a lower PPL score typically signifies a higher likelihood of occurrence according to the language model, suggesting a more coherent and grammatically aligned construction with the anticipated linguistic patterns. Hence, it can be inferred that queries with lower perplexity are generally characterized by fewer spelling errors.

[2] https://github.com/xu-song/bert-as-language-model



(a) Distribution of text length (b) Distribution of topic (c) Distribution of error type

Figure 2: Feature distribution of QSpell 250K including text length, topic and error type. Additionally, the distribution of original query frequency is 10% hot, 30% torso, and 60% long-tail. This is because the higher the frequency of a query, the lower the likelihood of it containing errors.

Dataset	Volume	Error Ratio	Lang	Error Type	Length	Field
MSC	5,892	19%	English	4	Short	Web
qSpell	6,000	16%	English	4	Short	Web
Webis	54,772	16%	English	4	Short	Web
SINGHAN13	700/1,000	20%	Chinese	2	Long	Specific
SINGHAN14	3,437/1,062	75%	Chinese	2	Long	Specific
SINGHAN15	2,339/1,100	64%	Chinese	2	Long	Specific
QSpell 250K	200,000/50,000	51%	Chinese,English	4	Short	Web

Table 2: The comparison of QSpell 250K and existing spell correction datasets. QSpell 250K, both in terms of data volume and data characteristics, provides an excellent complement to existing datasets.

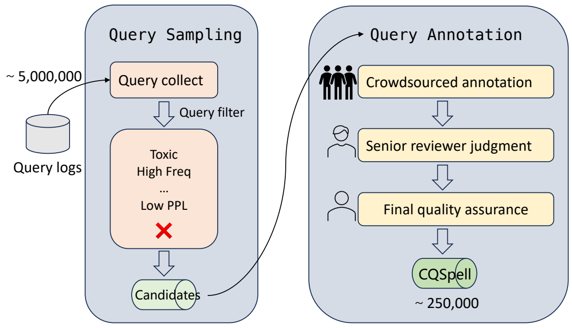


Figure 3: The annotation process of QSpell 250K benchmark.

3.3 Query Annotation

After automatically filtering the data, we manually annotated the remaining data, referred to as QSpell 250K. To encourage high-quality marking, we assign each query to three random annotators for independent annotation. Their submissions are then aggregated and sent to a random senior reviewer as the final judge. In addition, annotators can use any tool they want to support their work, such as search engines. Besides, to avoid persistent labeling mistakes, the annotation process is conducted in batches, and slight adjustments to the annotation standard are allowed at this stage. In this way, we can detect problems in the actual labeling process.

- Step 1. Crowdsourced annotation. During the annotation phase, the annotators are required to first examine whether the word itself contains any errors. If there are no errors, they proceed to assess the word's contextual appropriateness. Additionally, annotations utilize web search engines such as Google, Baidu, and references like Wikipedia to cross-validate the judgments. By adopting this approach, the annotations will not be biased to fit into one specific context.

- Step 2. Senior reviewer judgment. After a crowdsourced annotation of a batch is completed, it is sent to senior reviewer to judge whether it meets our annotation standard. This process repeats until the annotation accuracy rate reaches 90%.

- Step 3. Final quality assurance. Each batch of annotated queries that pass the first round of verification is sent to quality inspector for a second round of verification. The quality inspector randomly check 30% of the queries and send unqualified queries back to senior reviewers along with the reasons for rejection. The quality inspector possesses a solid educational background and is proficient in using various search tools.

3.4 Analysis and Comparison

In this section, we introduce the features of QSpell 250K from multiple perspectives and compare them with existing datasets.

Basic feature We show a comparison of our basic features with the existing datasets in Table 2. QSpell 250K comprises 250,000 queries, of which 50% are misspelled. It is evident that both the volume of our data and the proportion of errors exceed those of previous datasets by more than fourfold, signifying that QSpell 250K presents a more challenging task. QSpell 250K encompasses both Chinese and English, featuring a code-switching characteristic that previous datasets did not possess.

Topic distribution In a real-world application, user input often covers a variety of topics. For the convenience of analysis, we divide our data into 16 topics. Figure 2(b) depicts the proportion of our dataset for each topic. These topics are determined by annotators during the annotation process. Due to the diversity of topic distribution, our dataset also poses new challenges to the task of Query Spelling Correction.

Error type To enhance the coherence of the dataset with real-world scenarios, we incorporate these errors into QSpell 250K. Figure 2(c) shows the proportion of each error type. From the figure, it can be observed that approximately 75.8% of the errors are phonetically similar errors. This phenomenon may be attributable to the phonetic tendencies inherent in the Chinese language.

4 Evaluation

4.1 Datasets Processing

We randomly split QSpell 250K into a training set (200K), and a test set (50K) with a ratio of 10:1. In order to better fit the actual application scenarios of error correction and objectively measure the effect of the model, QSpell 250K contains both correct queries and error queries, the ratio is close to 1:1. If all the data in the training set need to be corrected, then the model will assume by default that all the input data are wrong.

4.2 Benchmark Models

Large Language Models (LLMs), such as ChatGPT and GPT-4 (Brown et al., 2020; OpenAI, 2023), have brought about a revolution in natural language processing, showcasing strong zero-shot and few-shot generalization capabilities. In this paper, we aim to evaluate the effectiveness of ChatGPT as a zero-shot learner for spelling correction. Specifically, we utilize the gpt-4-turbo model in Chat mode. To explore the efficacy of large language models in query spelling correction, we conduct supervised instruction tuning on the Qwen2.5 with size from 0.5B to 7B (Yang et al., 2024). Additionally, to more clearly present the performance metrics of existing datasets, we have also documented the results of state-of-the-art (SOTA) models (Sun et al., 2024).

4.3 Benchmark Metrics

We utilize Precision, Recall, and F1 Score as our evaluation metrics (Hasan et al., 2015; Ye et al., 2023). For each query q within the set Q, the spell correction approach predicts a result $G(q)$. For queries that require no correction, the corrector simply outputs the original query. Subsequently, we compare the model-generated results with the standard corrections $S(q)$ provided by the corpus.

4.4 Parameter Settings

Our experiments are conducted with Pytorch. For hyperparameter tuning, the learning rate is set to 3e-6, the max sequence length is set to 512, the up is 0.02 and the linear decay is 1.0. All experiments are conducted on the NVIDIA Tesla H100 with 80GB memory. For each model, we obtained the average from five experiments. This approach ensures a fairer comparison and mitigates the impact of random events. The prompt we used is as follows: `As a query spelling error correction model, your task is to automatically detect and correct query spelling errors in the query. If the query does not contain errors, output the original query. The input query is: {}. The output query is: {}`

4.5 Benchmark Experiments

Table 3 reveals the main results of our experiments. From the experimental results we have the following observation: 1) QSpell-250K demonstrates superior practicality compared to Webis and

Model Type	Model	QSpell 250K			SIGHAN			Webis		
		P	R	F1	P	R	F1	P	R	F1
Prompt	GPT4	0.7409	0.3146	0.4416	0.6395	0.4060	0.4967	0.3896	0.4418	0.4140
	Qwen2.5 7B	0.4906	0.2980	0.3708	0.3868	0.2328	0.2906	0.2155	0.3581	0.2691
FT	7B	0.8391	0.5750	0.6824	0.7835	0.3455	0.4795	0.5096	0.4867	0.4979
	3B	0.7380	0.3984	0.5175	0.3337	0.1565	0.2131	0.3725	0.3778	0.3751
	1.5B	0.7015	0.3266	0.4457	0.2018	0.0861	0.1207	0.2734	0.3540	0.3085
	0.5B	0.6184	0.3109	0.4138	0.1718	0.0546	0.0829	0.2406	0.2539	0.2471
SOTA	BERT	-	-	-	0.7803	0.7873	0.7880	-	-	-

Table 3: The performance of baselines on QSpell 250K, SIGHAN and Webis. For each model, we obtained the average from five experiments.

SIGHAN datasets. Our experiments with prompt-based LLMs reveal that QSpell-250K achieves better performance in error correction tasks. Additionally, with the increase in LLM model parameters, there is a smooth growth in performance on the QSpell 250K dataset without abrupt changes. This enhanced performance suggests that QSpell's samples better reflect real-world scenarios, as modern LLMs already possess strong error-correction capabilities. 2) The LLM performed good on QSpell 250K and Webis, but it showed poor results on the SIGHAN dataset. This may be attributed to the smaller sample size of SIGHAN, which makes it difficult for the LLM to transition to downstream tasks. Additionally, since Sighan is collected from a computer-based test of Chinese as a foreign language, it contains numerous rare error corrections, which also contribute to the suboptimal performance of LLMs on the Sighan dataset. 3) Off-the-shelf LLMs perform poorly in spell correction tasks and require fine-tuning. As the size of the model parameters increases, the performance of the LLM improves significantly. Overall, the experimental results indicate that the performance of the existing models on QSpell 250K falls short of our expectations, even with a substantial amount of training data.

4.6 Case Study

To verify the problems of existing models, we further analyze errors that cannot be handled in all baseline models.

Firstly, QSpell 250K requires more domain knowledge. For example, the correct query for "谁献计杀了蔡帽" (Who plotted to kill Cai Mao) should be "谁献计杀了蔡瑁" (Who plotted to kill Cai Mao). "蔡瑁"(Cai Mao) is a role name in the Romance of the Three Kingdoms, which is a famous Chinese novel.

Secondly, QSpell 250K requires greater context understanding. There are many queries with multiple error points in QSpell 250K. In such texts, the context of each error points contains at least one misspelled character, which brings noise information. For example, "成都半面的作法" is misspelled, and the correct query is "成都拌面的做法"(The method of ChengDu noodles served with soy sauce).

Thirdly, QSpell 250K requires multilingual understanding capabilities. For example, "windowss屏木翻转", contains Chinese and English errors. The correct query should be "windows屏幕翻转"(windows screen flip). To rectify such errors, the model must possess the capability to represent a multitude of languages effectively.

Overall, it is still very challenging to use existing models in a general application and correct these kinds of error.

5 Conclusion

In this paper, we present a Large-scale, naturalistic benchmark for Chinese Search Query Spelling Correction (QSpell 250K), which is collected from a real-world application. Compared with existing datasets like Microsoft Speller Challenge and SIGHAN, QSpell 250K supports more reliable evaluation due to the following features: 1) a variety of error patterns, 2) large scale, 3) code-switching. In addition, we conduct experiments on several representative spelling correction methods. The experiments have demonstrated that QSpell 250K is more challenging. At last, as shown by our experiments, the current Query Spelling Correction is not a "solved" problem and has much room for improvement. We hope our benchmark will benefit future research.

6 Limitations and Ethical Considerations

Data Collection for QSpell 250K: During the collection of QSpell 250K, we employ multiple methods aimed at ensuring user privacy, collecting only the users' search query information. Additionally, the data we gathered includes only Chinese and English. Since it does not encompass other languages, our experiments might not be easily generalizable to other search environments. Furthermore, the data originates from a Chinese search engine, representing a specific cultural and linguistic context, and does not reflect the global population.

Annotation of QSpell 250K: For annotating QSpell 250K, we utilized a mixed approach of crowdsourcing and senior reviewer annotations to ensure the quality of the annotations. During the annotation process, annotators could only see the queries and had no access to user information. Additionally, the annotators underwent multiple rounds of training to ensure the accuracy of the annotations. Although we made every effort to remove queries containing harmful intent during the annotation process, there may still be queries with potential risks remaining. In order to protect user privacy, we refrained from accessing the context of user queries.

In summary, we hope that QSpell 250K will foster development in the field of spell correction.

References

Farooq Ahmad and Grzegorz Kondrak. 2005. Learning a spelling error model from search query logs. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 955–962.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Qing Chen, Mu Li, and Ming Zhou. 2007. Improving query spelling correction using web search results. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 181–189.

Huizhong Duan, Yanen Li, ChengXiang Zhai, and Dan Roth. 2012. A discriminative model for query spelling correction with latent structural svm. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1511–1521.

Adam Fourney, Meredith Ringel Morris, and Ryen W White. 2017. Web search as a linguistic tool. In *Proceedings of the 26th International Conference on World Wide Web*, pages 549–557.

Yasser Ganjisaffar, Andrea Zilio, Sara Javanmardi, Inci Cetindil, Manik Sikka, Sandeep Katumalla, Narges Khatib, Chen Li, and Cristina Lopes. 2011. qspell: Spelling correction of web search queries using ranking models and iterative correction. In *Spelling Alteration for Web Search Workshop*, page 15.

Jianfeng Gao, Chris Quirk, et al. 2010. A large scale ranker-based system for search query spelling correction. In *The 23rd International Conference on Computational Linguistics*.

Hongyu Gong, Yuchen Li, Suma Bhat, and Pramod Viswanath. 2019. Context-sensitive malicious spelling error correction. In *The World Wide Web Conference*, pages 2771–2777.

Jai Gupta, Zhen Qin, Michael Bendersky, and Donald Metzler. 2019. Personalized online spell correction for personal search. In *The World Wide Web Conference*, pages 2785–2791.

Matthias Hagen, Martin Potthast, Marcel Gohsen, Anja Rathgeber, and Benno Stein. 2017. A large-scale query spelling correction corpus. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1261–1264.

Saša Hasan, Carmen Heger, and Saab Mansour. 2015. Spelling correction of user search queries through statistical machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 451–460.

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. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4084–4088.

Vishal Kakkar, Chinmay Sharma, Madhura Pande, and Surender Kumar. 2023. Search query spell correction with weak supervision in e-commerce. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 687–694.

Alex Kuznetsov and Hector Urdiales. 2021. Spelling correction with denoising transformer. *arXiv preprint arXiv:2105.05977*.

Mu Li, Muhua Zhu, Yang Zhang, and Ming Zhou. 2006. Exploring distributional similarity based models for query spelling correction. In *Proceedings of the 21st*

International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pages 1025–1032.

Yanen Li. 2020. Query spelling correction. *Query Understanding for Search Engines*, pages 103–127.

Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Edit5: Semi-autoregressive text editing with t5 warm-start. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2126–2138.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Madhura Pande, Vishal Kakkar, Manish Bansal, Surender Kumar, Chinmay Sharma, Himanshu Malhotra, and Praneet Mehta. 2022. Learning-to-spell: Weak supervision based query correction in e-commerce search with small strong labels. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3431–3440.

Sanat Sharma, Josep Valls-Vargas, Tracy Holloway King, Francois Guerin, and Chirag Arora. 2023. Contextual multilingual spellchecker for user queries. *arXiv preprint arXiv:2305.01082*.

Changxuan Sun, Linlin She, and Xuesong Lu. 2024. Two issues with chinese spelling correction and a refinement solution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–204.

Xu Sun, Anshumali Shrivastava, and Ping Li. 2012. Fast multi-task learning for query spelling correction. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 285–294.

Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen. 2015. Introduction to sighan 2015 bake-off for chinese spelling check. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, pages 32–37.

Kuansan Wang and Jan Pedersen. 2011. Review of msr-bing web scale speller challenge. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1339–1340.

Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee. 2013. Chinese spelling check evaluation at sighan bake-off 2013. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages 35–42.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Fan Yang, Alireza Bagheri Garakani, Yifei Teng, Yan Gao, Jia Liu, Jingyuan Deng, and Yi Sun. 2022. Spelling correction using phonetics in E-commerce search. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 63–67, Dublin, Ireland. Association for Computational Linguistics.

Dezhi Ye, Bowen Tian, Jiabin Fan, Jie Liu, Tianhua Zhou, Xiang Chen, Mingming Li, and Jin Ma. 2023. Improving query correction using pre-train language model in search engines. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2999–3008.

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 *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 126–132.

Jingfen Zhang, Xuan Guo, Sravan Bodapati, and Christopher Potts. 2023a. Multi-teacher distillation for multilingual spelling correction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 142–151.

Xiaowu Zhang, Xiaotian Zhang, Cheng Yang, Hang Yan, and Xipeng Qiu. 2023b. Does correction remain an problem for large language models? *arXiv preprint arXiv:2308.01776*.