Retrieval Augmented Spelling Correction for E-Commerce Applications

Xuan Guo
Amazon.com Inc
xuangu@amazon.com

Rohit Patki
Amazon.com Inc
patkir@amazon.com

Dante Everaert
Amazon.com Inc
danteev@amazon.com

Christopher Potts
Stanford University
cgpotts@stanford.edu

Abstract

The rapid introduction of new brand names into everyday language poses a unique challenge for e-commerce spelling correction services, which must distinguish genuine misspellings from novel brand names that use unconventional spelling. We seek to address this challenge via Retrieval Augmented Generation (RAG). On this approach, product names are retrieved from a catalog and incorporated into the context used by a large language model (LLM) that has been fine-tuned to do contextual spelling correction. Through quantitative evaluation and qualitative error analyses, we find improvements in spelling correction utilizing the RAG framework beyond a stand-alone LLM. We also demonstrate the value of additional finetuning of the LLM to incorporate retrieved context.

1 Introduction

New brand names are continuously introduced, and many of them use unconventional spelling to create specific associations while still ensuring that the brand is unique and memorable. Prominent examples include "playgro" vs. "playground", "biotanicals" vs. "botanicals", and "hygeeni" vs. "hygiene". Such cases pose a significant challenge for e-commerce spelling correction services, which are prone to over-correcting such terms, especially completely novel ones.

In this paper, we seek to address this challenge by leveraging Retrieval Augmented Generation (RAG). On this approach, the user's query is passed to a retrieval module that seeks to find relevant items from a product catalog. The retrieved items are then incorporated into a prompt to a large language model (LLM) that predicts a correct spelling for the user's query.

We report on a wide range of experiments with different retrieval models and LLMs. We find that the RAG-based approach consistently leads to large performance improvements with only minor latency increases. In addition, we explore methods for fine-tuning the LLM to make better use of retrieved contexts, and we find that this leads to very substantial improvements, with the largest gains coming from queries that contain brand names.

2 Related Work

Retrieval Augmentation Retrieval has proven highly effective across a wide range of knowledge intensive tasks. Early works such as DrQA (Chen et al., 2017) and Dense Passage Retrieval (DPR; Karpukhin et al. 2020) laid important groundwork by integrating retrieval with neural models to enhance question-answering accuracy and facilitate knowledge access. REALM (Guu et al., 2020) introduced unsupervised retrieval for language model pre-training, and Lewis et al. (2020) combined retrieval with generation to tackle knowledge intensive tasks. RETRO (Borgeaud et al., 2022) scaled these ideas by accessing a vast database of tokens for contextual relevance across large datasets. Khattab et al. (2021) weakly supervise the ColBERT neural retrieval model to improve performance on open-domain question answering. Overall, these approaches seem to help systems provide up-to-date knowledge and reduce hallucinations (Asai et al., 2023). Despite these advances, the specific challenges of contextual spelling correction in e-commerce settings, particularly for dynamically evolving brand names, remain underexplored.

Retrieval Augmented Fine-Tuning Retrieval Augmented Fine-Tuning (RAFT) adapts models by fine-tuning them for specific tasks like question answering, with a strong focus on handling irrelevant documents to boost accuracy (Zhang et al., 2024). Similarly, Atlas (Izacard et al., 2023) demonstrates the effectiveness of retrieval-augmented models for few-shot learning by integrating retrieved content during both pre-training and fine-tuning phases.

While other models like RPT (Rubin and Berant, 2023) and REPLUG (Shi et al., 2023) also combine retrieval with generative capabilities, they either focus on black-box integration or training from scratch without fine-tuning on external context. Our approach, akin to RAFT, fine-tunes the LLM with task-specific data, distinguishing between relevant and irrelevant contexts, especially for complex non-standard brand names. This underscores the value of targeted retrieval in specialized tasks like e-commerce spelling correction.

Contextualized Spelling Correction Related studies in character-level and multilingual spelling correction highlight the importance of context and fine-grained adjustments. For example, Zhang et al. (2023) show that using multiple teacher models improves spelling correction across diverse languages, underscoring the value of context-specific training. Huang et al. (2023) introduce structural interventions at the subword level to improve spelling correction, particularly for morphologically complex terms. Unlike these methods, which are tailored to specific contexts, our RAG-based approach generalizes to evolving and unconventional brand names by retrieving up-to-date context on demand, enhancing adaptability and robustness in real-world e-commerce applications.

Contextualized spell-checking methods (Song et al., 2023; Wang et al., 2023) emphasize the value of integrating external knowledge to handle domain-specific terms, demonstrating the benefits of user-specific data for improved spelling accuracy. Unlike these methods, which use prompt conditioning and attention mechanisms to incorporate context, our approach employs RAG to meet the unique demands of an evolving e-commerce catalog. By leveraging RAG, we dynamically integrate new and modified brand names directly from the catalog, enabling adaptation to non-standard lexicons unique to brand names. This retrieval-based method helps address the continuously updated nature of brand catalogs in ways that prompt-tuning or attention-based techniques alone may not fully resolve, particularly for ambiguous or unconventional spellings.

3 Approach

We now present our approach to spelling correction, which leverages Retrieval Augmented Generation (RAG) and optionally includes a phase of fine-tuning the LLM to make better use of context.

3.1 Language Models

We experiment primarily with Mistral-7B (Jiang et al., 2023) (specifically, `open-mistral-7b` v0.1) and Claude-3-sonnet (`claude-3-sonnet-20240229` v1:0). We chose Mistral-7B because it is highly effective for its size (see Section 4.3), and we chose Claude-3-sonnet as a representative of a much larger class of LLMs. We would expect to see similar results for other LLMs, even larger and more capable ones, because our central challenge is making accurate predictions about novel brand names.

3.2 Retrieval Models

We evaluate three main retrieval methods:

1. **BM25** (Robertson et al., 2009) is a traditional, time-tested n-gram-based retrieval model. We expect it to excel where exact matching suffices but may struggle where fuzzy or semantic matching is called for.

2. **Fuzzy BM25** combines traditional BM25 with fuzzy matching. This allows for minor spelling errors that are common in e-commerce queries. For instance, it can match "air fryer cusinart" to both "air fryer" and "cuisinart".

3. **ColBERT** (Khattab and Zaharia, 2020; Santhanam et al., 2022a,b) is a neural retrieval model that represents queries and documents with token-level vectors. We expect this model to excel at retrieving semantically relevant terms.

In all cases, we index our product catalog. For ColBERT, this is done using a pretrained ColBERTv2 model checkpoint released by the ColBERT team.[1]

3.3 Prompt Design

The following is the LLM prompt template that we use where the retrieved items are provided as context:

```
### Instruction:
Provide spelling correction for given query
if necessary, referring to the provided context
if it's relevant.
### Context:
{context}
### Query:
{input}
### Correction:
```

[1] huggingface.co/colbert-ir/colbertv2.0

Retrieved items are formatted as a comma-separated list and placed in the `context` field. We use the top 3 or 4 retrieved items, with the precise number controlled by the maximum prompt length we permit in our fine-tuning runs. (Future work could explore the effects of including more context items at inference time.)

3.4 LLM Fine-Tuning

Our approach includes an optional step of fine-tuning the LLM to make more effective use of context. We consider two variants.

For **Basic Fine-Tuning**, the training dataset consists of 50K `<input query, label query>` pairs derived from search logs to create a training dataset. Here, `label query` reflects user-validated corrections. For this fine-tuning, we remove the "referring to the provided context if it's relevant" wording from `Instruction` of the prompt in Section 3.3, with the `context` field also removed and the desired output appended to the end followed by ### on a newline.

For **Contextual Fine-Tuning**, we begin with the same dataset of 50K pairs, but now we retrieve context for each `<input query>`, forming `<input query, context, label query>` triples. This added context, derived through the same retrieval mechanism used in RAG, helps teach the model how to make use of the `context` field. The prompt has the same format as the one in Section 3.3 but with the desired output appended to the end followed by ### on a newline.

For both variants, the fine-tuning process is simply additional language model training using strings formatted from our prompt templates. Further evaluation details, including metrics, are covered in Section 4.

4 Experiments

4.1 Evaluation Data

Our evaluation dataset is sourced from search logs collected between 2021 and 2023. Each data point consists of an `<input query, label query>` pair. The `input query` refers to the user's original search query, while the `label query` is obtained from annotators. We conducted stratified sampling to arrive at a 10K `input query` set, which was designed to promote the diversity of the query population, particularly with regard to the presence of misspellings (roughly 1/4) and brand names (roughly 1/3 cases with a brand name). To ensure label quality, we applied a "2+1" annotation method. That is, two annotators initially labeled each query; if they disagreed, an auditor made the final determination. This process included specific instructions to guide annotators on the e-commerce context. We use this dataset to evaluate all experiments in this paper.

LLM (no finetune)	Size	F1
Flan UL2	20B	13.0
mT0	13B	19.3
Flan T5	11B	20.0
mGPT	13B	21.7
Mistral	7B	28.1
Claude-3-sonnet	70B*	34.7
Mixtral	47B	57.4

Table 1: Results for zero-shot LLMs. The largest models achieve the best results, and Mistral-7B achieves excellent results for its size. *The size given for Claude-3-sonnet is a guess based on the comparative performance of the model relative to others of known size.

4.2 Metrics

Our primary metric for evaluating spelling correction quality is the F1 score, which is the harmonic mean of precision and recall. Precision reflects the proportion of model-predicted corrections that match the gold standard annotations, while recall measures the proportion of required corrections that the model identifies correctly. We rely on exact match criteria, where two strings are considered equal after punctuation removal. All F1 scores are reported as percentages for clarity.

4.3 Zero-Shot LLM Performance

Table 1 reports baseline results for a range of different LLMs used without any retrieval. The prompt template used for this is the same one as in Section 3.3 without the `context` field and its mentions in the `Instruction` section. The top-performing model by a large margin is Mixtral-47B, followed by Claude-3-sonnet (≈70B, estimated). The much smaller Mistral-7B model (Jiang et al., 2023) is reasonably competitive, though, and it represents a better balance of costs and performance, especially for high-volume services like spelling correction.

4.4 RAG with a Frozen LLM

Table 2 provides our primary results. We consider Mistral-7B and Clause-3-sonnet as the base LLMs. For each LLM, we evaluate our three different re-

Retriever	LM	Doc index	F1
BM25	Mistral-7B	572K	25.4
Fuzzy BM25	Mistral-7B	60K	31.7
Fuzzy BM25	Mistral-7B	572K	34.6
ColBERT	Mistral-7B	60K	35.8
ColBERT	Mistral-7B	572K	35.9
BM25	Claude-3-sonnet	572K	26.2
Fuzzy BM25	Claude-3-sonnet	60K	30.4
Fuzzy BM25	Claude-3-sonnet	572K	34.8
ColBERT	Claude-3-sonnet	60K	29.8
ColBERT	Claude-3-sonnet	572K	39.3

Table 2: Results for RAG with frozen LLMs. The best configurations use ColBERT and a larger document index. Both LLMs are substantially improved by RAG.

trieval models. We also explore the role of the size of the product catalog by considering two different pools of documents: one with 572K documents (132K unique brands) and one with 60K documents (29K unique brands).

The overall best-performing setting is with ColBERT indexing 572K documents and providing retrieval results to Claude-3-sonnet. This configuration results in 39.3 F1 vs. 34.7 for Clause-3-sonnet used without retrieval (Table 1): a 4.6 point increase. Strikingly, the next best system is one that uses Mistral-7B, again with ColBERT indexing the larger document collection. This configuration achieves 35.9 vs. 28.1 for Mistral-7B used without retrieval: a 7.8 point increase.

Table 3 provides a more qualitative analysis that reveals nuanced interactions between retrieved context and LLM responses. When relevant context is retrieved (Examples 1–4), the LLM provides accurate corrections aligned with expected outputs. Conversely, in instances where context is absent (Examples 5–6), incorrect (Examples 7–8), or misleading (Example 9), the LLM's performance varied, highlighting the complex balance between the LLM's parameterized knowledge and the information retrieved. These examples illustrate that, while retrievers enrich context, they can also introduce noise or irrelevant data that might detract from accuracy if not carefully managed.

4.5 LLM Fine-Tuning

Table 4 summarizes our experiments involving LLM fine-tuning, using both Basic and Contextual variants of this method (Section 3.4). For these experiments, we adopt Mistral-7B as our base LLM. To facilitate comparisons, the top row in the first section of the table is repeated from Table 1, and the top rows from the middle and bottom sections are repeated from Table 2. We include the precision/recall breakdown here to support further analysis of the trade-offs.

Across all three panels we see substantial gains from fine-tuning, with similar precision/recall ratios across all settings. The largest gains come from Contextual Fine-Tuning. The best performing configuration uses ColBERT and Contextual Fine-Tuning, leading to 70.1 F1, a 34.2 point increase over the system that employs only RAG with ColBERT. Thus, the overall message is very clear: if it is feasible to fine-tune the LLM with context, that is likely to lead to very substantial performance improvements.

Our primary goal is to improve performance on brands that are novel from the perspective of the LLM. Table 5 seeks to quantify the extent to which Contextual Fine-Tuning marks progress in this area, as compared to Basic Fine-Tuning. In both cases, we train on the same set of examples (Section 3.4). The table shows that Contextual Fine-Tuning leads to a 6.9 point increase in overall F1 and a 16.7 point increase in queries containing brands. While brands remain very challenging, our approach certainly alleviates the challenge.

These gains are supported by qualitative analysis as well. For example, in cases like "snowflake necklace for women", where the context includes varied necklaces ("swarovski snowflake necklace for women, efytal necklace for women, birthstone necklace for women, baguette necklace for women pavori"), Contextual Fine-Tuning helps the model produce the correct label "swarovski snowflake necklace for women," ensuring that it makes thoughtful corrections rather than echoing the context.

Ideally, we would be able to home in on a set of definitely new brands and measure performance on them. However, this set is challenging to define, since multiple sources of knowledge are in play, including the pretrained knowledge parameterized in LLM. However, we are able to identify a set of 525 brands that are in our evaluation set but absent from our fine-tuning dataset. For this set, we actually get an F1 score of 78.4 using Contextual Fine-Tuning and a ColBERT retriever. For future work, we will snapshot new brands by LLM release date as the cutoff, to present a more controlled experiment.

#	User query	Retrieved items	RAG generation	Ground truth
1	sumbrella umbrella outdoor patio	abba patio umbrella, patio umbrella sunbrella, sunbrella umbrellas, hampton bay patio umbrella	sunbrella umbrella outdoor patio	sunbrella umbrella outdoor patio
2	niwax ne tx direct	nikwax tx direct,tx direct spray on,tx direct	nikwax ne tx direct	nikwax ne tx direct
3	dubman brush	denman brush natural hair, denman brush, denman nylon brush, denman brush	denman brush	denman brush
4	salamin boots	mens salomon boots, salomon snowboard boots, salomon womens winter boots	salomon boots	salomon boots
5	tumeric soap bars	bali soap bars, camay soap bars, dettol soap bars, himalaya soap bars	turmeric soap bars	turmeric soap bars
6	doom eternal puns	doom eternal juguete,doom,doom eternal	doom eternal puns	doom eternal pins
7	tonkatsu ramen noodles	tonkatsu ramen noodles, tonkatsu ramen bowl, buldak ramen noodles, immi ramen noodles	tonkotsu ramen noodles	tonkotsu ramen noodles
8	correlle	correlle cafe red bowl, correlle plates white, correlle , correlle red bowl	correlle	corelle
9	laroche b5	guy laroche ,flydigi b5, laroche , laroche set	laroche b5	la roche b5

Table 3: Qualitative evaluation of RAG-generated spelling corrections across various examples. In 1–4, correctly spelled retrieved items lead to accurate corrections. In 5–6, the retrieved items do not involve misspelled spans of the input query, leading RAG generation to rely on the LLM's internal knowledge. In 7, the LLM generates the correct spelling despite misspelled retrieved items, while in 8, the model is misled by the incorrect retrieval. Finally, in 9, "la roche" and "laroche" are both real brands. The retriever does not correctly consider the context "b5" to distinguish the brands ("b5" is a specific item that is only associated with brand "la roche").

Retriever	LLM	Precision	Recall	F1
None	Mistral-7B	30.3	26.2	28.1
	Mistral-7B with Basic Fine-Tuning	70.3	59.0	64.1
Fuzzy BM25	Mistral-7B	42.8	29.0	34.6
	Mistral-7B with Basic Fine-Tuning	49.7	40.3	44.5
	Mistral-7B with Contextual Fine-Tuning	77.4	59.5	67.3
ColBERT	Mistral-7B	43.1	30.8	35.9
	Mistral-7B with Basic Fine-Tuning	52.3	42.1	46.6
	Mistral-7B with Contextual Fine-Tuning	77.6	65.5	71.0

Table 4: Performance comparison across different retrieval configurations and fine-tuning setups. All experiments used an indexed pool of 572K documents (132K unique brands). The highest F1 score of 71.0 was achieved with ColBERT in RAG, demonstrating the added benefit of Contextual Fine-Tuning. These results indicate that fine-tuning with context-specific instructions is extremely effective.

	F1	
	All queries	Brands
Basic Fine-Tuning	64.1	44.1
RAG + Contextual Fine-Tuning	71.0	60.8

Table 5: Performance improvements from Contextual Fine-Tuning and RAG as compared to Basic Fine-Tuning without RAG, broken down by overall performance and performance on queries containing brands. The LLM is Mistral-7B and the retriever used for RAG is ColBERT over the 572K document collection.

4.6 Latency Considerations

Incorporating RAG leads to a slight latency increase; however, it remains within acceptable limits for real-time applications. Using the Mistral-7B model as a baseline, retrieval from a pool of 60K candidate documents adds only 2.42% to the overall generation time, while expanding the pool to 572K documents results in a 2.79% increase. These changes are minimal, and they enable substantial gains in accuracy.

5 Conclusion

In this paper, we introduced a fine-tuned Retrieval Augmented Generation (RAG) framework tailored for e-commerce spelling correction, specifically addressing the complexities posed by brand names and other non-standard lexicons. We showed that this approach is highly effective even with a frozen retriever and frozen large language model (Table 2). In addition, we showed that fine-tuning the LLM with retrieved context leads to even larger gains (Table 4), particularly for spelling corrections involving evolving brand names (Table 5). These results underscore the value of incorporating retrieval and allowing the model to dynamically adapt to context in a way that standalone LLMs or RAG with frozen components cannot achieve.

Our qualitative analysis further revealed challenges inherent in using real-world data, such as the presence of misspellings in indexed documents, which can mislead the LLM during generation (Table 3). This highlights the importance of ensuring the quality of retrieved contexts. Practical improvements include refining the contextual data through heuristic signals, like user interactions and engagement metrics, to enhance relevance and accuracy. Another promising avenue is to diversify the styles and noise levels within the retrieved context to bolster the model's robustness.

For long-term directions, we propose exploring mechanisms that enable LLMs to assess and selectively integrate context based on relevance and quality. Such advancements could pave the way for smarter, more context-aware LLMs that distinguish valuable insights from noise, ultimately enhancing their adaptability in real-world applications. Additionally, evaluating models with context on emerging entities could provide a more dynamic measure of RAG's effectiveness as new content enters the dataset. These lines of inquiry contribute insights into optimizing LLMs within RAG frameworks, driving advancements in the broader field of adaptive language models and their application in context-sensitive domains.

6 Ethics Statements

This study uses anonymized, user-generated data to enhance the model's ability to do contextual spelling correction in e-commerce. We acknowledge that user-generated data may reflect inherent biases, such as regional or demographic linguistic preferences, which could affect spelling correction accuracy for certain user groups. We are committed to monitoring these issues and improving the fairness of the model over time, aiming to make spelling correction equitable, inclusive, and accurate. Future efforts will focus on refining our methodology to address these concerns, especially as the model encounters new data and evolves to handle a broader range of brand-specific terminology and user inputs.

References

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 41–46, Toronto, Canada. Association for Computational Linguistics.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240, Baltimore, Maryland. PMLR.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879,

Vancouver, Canada. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, pages 3929–3938, Vienna, Austria.

Jing Huang, Zhengxuan Wu, Kyle Mahowald, and Christopher Potts. 2023. Inducing character-level structure in subword-based language models with type-level interchange intervention training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12163–12180, Toronto, Canada. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. arXiv:2310.06825.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. Relevance-guided supervision for OpenQA with ColBERT. *Transactions of the Association for Computational Linguistics*, 9:929–944.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, New York, NY. Association for Computing Machinery.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Ohad Rubin and Jonathan Berant. 2023. Retrieval-pretrained transformer: Long-range language modeling with self-retrieval. arXiv:2306.13421.

Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. 2022a. PLAID: An efficient engine for late interaction retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1747–1756, New York, NY, USA. Association for Computing Machinery.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022b. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. REPLUG: Retrieval-augmented black-box language models. arXiv:2301.12652.

Gan Song, Zelin Wu, Golan Pundak, Angad Chandorkar, Kandarp Joshi, Xavier Velez, Diamantino Caseiro, Ben Haynor, Weiran Wang, Nikhil Siddhartha, et al. 2023. Contextual spelling correction with large language models. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Xiaoqiang Wang, Yanqing Liu, Jinyu Li, and Sheng Zhao. 2023. Improving contextual spelling correction by external acoustics attention and semantic aware data augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jingfen Zhang, Xuan Guo, Sravan Bodapati, and Christopher Potts. 2023. 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, Singapore. Association for Computational Linguistics.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. 2024. RAFT: Adapting language model to domain specific RAG. arXiv:2403.10131.