UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

MATEUS MATIAS RIBEIRO

AVALIANDO MODELOS DE LLM PARA PERSONALIZAÇÃO DE CONSULTAS E AUMENTO DE RELEVÂNCIA NO E-COMMERCE

CAMPINA GRANDE - PB

2024

MATEUS MATIAS RIBEIRO

AVALIANDO MODELOS DE LLM PARA PERSONALIZAÇÃO DE CONSULTAS E AUMENTO DE RELEVÂNCIA NO E-COMMERCE

Trabalho de Conclusão de Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador : Fábio Jorge Almeida Morais

CAMPINA GRANDE - PB

2024

MATEUS MATIAS RIBEIRO

AVALIANDO MODELOS DE LLM PARA PERSONALIZAÇÃO DE CONSULTAS E AUMENTO DE RELEVÂNCIA NO E-COMMERCE

Trabalho de Conclusão de Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Fábio Jorge Almeida Morais
Orientador – UASC/CEEI/UFCG

Carlos Eduardo Santos Pires
Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro
Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 15 de Maio de 2024.

CAMPINA GRANDE - PB

RESUMO

A capacidade de fornecer resultados de pesquisa personalizados e relevantes em um ambiente de e-commerce altamente competitivo é crucial para a satisfação do cliente e o sucesso das lojas online. Neste trabalho, exploramos um método para melhorar a experiência de pesquisa no e-commerce usando modelos de aprendizado profundo para personalizar as consultas do usuário e melhorar a relevância dos itens retornados. O modelo de aprendizado de máquina apresentado foi projetado como uma prova de conceito para avaliar sua capacidade de entender o contexto e a intenção por trás das consultas de pesquisa do usuário e adaptá-las de forma inteligente antes de serem submetidas ao mecanismo de pesquisa. O modelo reescreve a consulta original para priorizar os produtos de interesse do cliente, descobrindo a intenção subjacente do usuário e o contexto da pesquisa. Além disso, também propomos um modelo classificador que é responsável por selecionar consultas passíveis de serem reescritas antes de usar o modelo de reescrita. Esta abordagem permite melhorar os resultados da pesquisa para destacar produtos de interesse, melhorando significativamente a relevância e a eficácia da pesquisa.

EVALUATING LLM MODELS FOR PERSONALIZING QUERIES AND INCREASING RELEVANCE IN E-COMMERCE

ABSTRACT

The ability to provide personalized and relevant search results in a highly competitive e-commerce environment is crucial for customer satisfaction and the success of online stores. In this work, we explore a method to enhance the search experience in e-commerce using deep learning models to personalize user queries and improve the relevance of the returned items. The presented machine learning model was designed as a proof of concept to assess its ability to understand the context and intention behind user search queries, and to intelligently adapt them before being submitted to the search engine. The model rewrites the original query to prioritize customer's interest products by uncovering the underlying user intention and search context. Additionally, we also propose a classifier model that is responsible for selecting rewritable queries before using the rewriter model. This approach allows search results to be improved to highlight products of interest, significantly improving the relevance and effectiveness of the search.

Evaluating LLM Models for Personalizing Queries and Increasing Relevance in E-commerce

Mateus Matias Ribeiro
Curso de Ciência da Computação - CCC
Unidade Acadêmica de Sistemas e Computação - UASC
Universidade Federal de Campina Grande - UFCG
Campina Grande, Paraíba - Brasil
mateus.ribeiro@ccc.ufcg.edu.br

Fabio Morais
Curso de Ciência da Computação - CCC
Unidade Acadêmica de Sistemas e Computação - UASC
Universidade Federal de Campina Grande - UFCG
Campina Grande, Paraíba - Brasil
fabio@computacao.ufcg.edu.br

***Abstract*—The ability to provide personalized and relevant search results in a highly competitive e-commerce environment is crucial for customer satisfaction and the success of online stores. In this work, we explore a method to enhance the search experience in e-commerce using deep learning models to personalize user queries and improve the relevance of the returned items. The presented machine learning model was designed as a proof of concept to assess its ability to understand the context and intention behind user search queries, and to intelligently adapt them before being submitted to the search engine. The model rewrites the original query to prioritize customer's interest products by uncovering the underlying user intention and search context. Additionally, we also propose a classifier model that is responsible for selecting rewritable queries before using the rewriter model. This approach allows search results to be improved to highlight products of interest, significantly improving the relevance and effectiveness of the search.**

***Index Terms*—Deep learning, LLM, User intent, Search context, Search engine, Search relevance, Context understanding, Query rewriting, Customer interest products**

I. Introduction

E-commerce has become an integral part of our lives. With the increasing competition in this sector, customer satisfaction is crucial for the success of any online store. By focusing on their consumers' needs and preferences, online stores may develop strong relationships and loyalty, resulting in higher buy conversion rates. One of the most effective ways to achieve this is by providing personalized and relevant search results for users.

However, the implementation of this personalization approach is a significant challenge. Achieving search relevance involves understanding the context and intention behind user search queries, a complex task given the multifaceted and often ambiguous nature of human language. Additionally, each user has unique and specific needs, making scale customization a difficult task, especially considering the diversity of segments in e-commerce today. Traditional machine learning approaches may not be sufficient to handle these challenges due to their inability to fully grasp the semantic context and user intention behind the query.

Deep Learning Language Models, Large Language Models (LLMs) in particular, have shown great promise in e-commerce. According to Huang et al. (2023), pre-trained LLMs can predict the similarity of behavior-based queries in e-commerce, highlighting the ability of LLMs to capture the semantics of user queries [4]. Therefore, based on evidence from current literature, LLMs offer a promising approach to improving the personalization and relevance of search results in e-commerce.

Current approaches focus on understanding user intent and context when converting confusing queries into more coherent searches. For example, Contrastive Learning Enhanced Query Rewriting (CLE-QR) [7] identifies alternative formulations based on user intent and context. Another example in the literature is the Query Understanding Enhanced Mechanism (QUEEN) [10], which employs query annotations from search pipelines to train a deep learning model for rewriting tasks. These studies demonstrate the possibility of improved search results through a more detailed understanding of user intent and context applied to query rewriting.

Despite the encouraging results of existing query rewriting techniques, the potential for LLMs in this task remains unexplored. Our study differs in that it focuses on query rewriting within a multi-tenant e-commerce platform and uses Brazilian Portuguese data. This new technique intends to explore the effectiveness of query rewriting in a diversified e-commerce environment, with a particular focus on the Brazilian market.

II. Query Rewriting in E-commerce

The need to rewrite queries in the e-commerce domain arises from the diversity of search behaviors and the inherent limitations of traditional search algorithms in interpreting user intent [5]. Users often approach search features with a high degree of uncertainty about the exact nomenclature or specifications of the product they want to purchase [6]. This uncertainty leads to the introduction of vague or ambiguous queries into search engines, which, without sophisticated interpretation mechanisms, can produce irrelevant results.

Typographical errors can further exacerbate the disconnect between user queries and desired search results Search engines that rely on text-to-text matching are ill-equipped to deal with these inconsistencies, resulting in a failure to present relevant products [3], thus degrading the user experience and purchase

conversion. Furthermore, the complexities of natural language, with its numerous syntactic and semantic nuances, present additional challenges in query interpretation. Users may inadvertently use industry jargon, colloquial terms, or highly specific phrases that don't align with the standardized product descriptions or metadata used by e-commerce platforms.

The increasing trend toward natural language queries requires a more advanced understanding of user intent and context to provide satisfactory search results. Therefore, query rewriting systems must evolve to accommodate not only keyword-based searches but also the complexities of question-based queries.

III. Related Works

Query rewriting is currently capable of improving the retrieval of relevant results for the user. For example, Sen Li [7] presents Contrastive Learning Enhanced Query Rewriting (CLE-QR), a solution capable of finding rewrites of historical queries, considering relevance and personalization, which showed significant gains in the observed metrics.

In a related study, Zuo et al [12]. proposed an end-to-end context-aware query rewriting model to enhance the understanding of users' true shopping intents in e-commerce search. Recognizing the limitations of existing query rewriting models that overlook users' history behaviors and focus solely on the instant search query, this study emphasizes the importance of modeling contextual information for a better query rewriting model.

Wang et al. presented QUEEN [10], a neural query rewriting solution for e-commerce search, with the goal of improving the buying experience by converting unclear consumer input queries into well-formed searches. By utilizing query annotations obtained from query processing pipelines, the Query Understanding Enhanced Mechanism (QUEEN) model, tackles the problem of large-scale query rewriting in e-commerce search engines.

These studies show the potential for enhanced search results by better understanding user intent and context, and they also emphasize the significance of query rewriting in the e-commerce area. However, these research do not investigate the context-understanding capacity of LLM models. Our study differs from those cited in the application of LLMs to rewrite queries in the e-commerce area, using queries in Portuguese and product data from a multi-tenant e-commerce platform.

Our methodology integrates real-time online assessments within the platform's API, allowing a practical and dynamic assessment of LLMs' performance in rewriting questions in a live e-commerce context. The goal of this approach is to improve search results, enhance user experience, and increase conversion rates in an increasingly diverse e-commerce landscape.

IV. Medodology

A. Query Customizer Solution

This section presents the flow described in this work. The pipeline considers a query set along with the respective items of interest. It analyzes the query's potential for rewriting using a Rewrite Candidate Classifier (RCC). Queries identified as candidates for rewriting are then sent to the LLM for the purpose of being rewritten. Ultimately, a rewritten query is produced for each query classified as improvable. The rewritten queries are then dispatched to the ecommerce store's Search API, which begins the process of creating updated search results. Figure 1 shows the overview of the proposed pipeline.

The modified results are then juxtaposed with the initial query results, allowing for comparative analysis to determine any changes in product positioning. This iterative method allows us to evaluate the effectiveness of query rewriting in increasing the relevance and accuracy of search results on the e-commerce platform, resulting in a systematic framework to improve user experience and engagement.

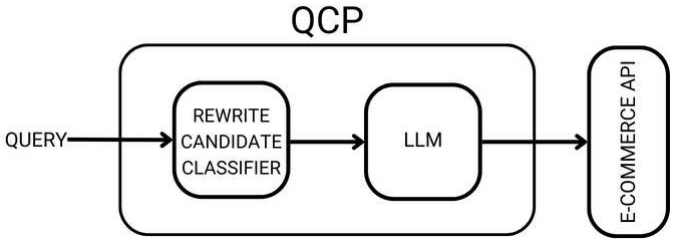

Fig. 1. Query Customizer Pipeline

The decision to use LLMs for query rewriting in e-commerce queries is supported by research demonstrating the effectiveness of LLMs in information retrieval tasks [11]. We use an LLM to act as a reviewer taking inspiration from the query review strategies established in Amemiya et al. study [2], which determines deletion of terms, replacement of terms, reformulation, and correction of typos, as identified as ways of improving queries and added an additional rule for lower context queries, term addition.

B. Data

The data used in this study came from a multi-tenant e-commerce platform and included a comprehensive dataset of about 1 million samples. These examples include queries, search engine results and products selected by consumers on the platform, these products are considered the product of interest and their position is used to validate the quality of the query. The data covered a one-month period, providing a snapshot of user interactions and search behavior. Specifically, data from five tenants on the e-commerce platform were used for the analysis, ensuring a diverse representation of user behaviors.

C. Experimental Setup

To improve the effectiveness of LLM-based query rewriting, a two-step pipeline was developed that combines RCC and LLM-based techniques. This technique prioritized rewriteable queries over non-rewriteable questions in order to increase the accuracy and relevancy of ecommerce search results. The RCC

successfully extracted non-rewritable queries from the dataset; its design will be discussed in (V).

The selected queries were then rewritten using LLM using OpenAI's GPT 3.5 Turbo. We employed the rewriting rules to improve queries described by Amemiya et al. [2], which determine term deletion, term replacement, reformulation, and correction of typing errors. As well as, an additional rule, term addition, for queries with less context.

Following the rewriting process, the modified queries are sent to the e-commerce store's search API, and the search engine results pages are saved. These results are then compared to the original query results, which allows a comparison analysis to evaluate the pipeline.

D. Evaluation

Comparing the product position on the new results page with the original position categorized the results as Success (improved position), Neutral (unchanged position) or Failure (worsened position). An analysis of a randomized subset of 2,000 queries, used for validation and model comparison in (VI), revealed a distribution of 81.8% successful results, 3.2% neutral, and 15% unsuccessful. In an upcoming part, we will use Mean Reciprocal Rank (MRR) to evaluate and compare the performance of different models once they are incorporated into the pipeline. This metric will show the efficacy of each model in terms of ranking relevance, allowing for a thorough evaluation of its impact on overall system performance.

V. Non-improvable Queries

It is crucial to understand why some queries fail to improve or may produce less relevant results compared to the original results obtained. An analysis of data from multiple tenants reveals a recurring pattern among most queries that showed no improvement. These patterns fall into three distinct categories: first, information-insufficient queries, where our method cannot derive meaningful insights from single or extremely brief words and unrelated terms; second, searches using specific codes, such as product IDs or internal standards; and third, exact product names, such as searches for a specific brand or model.

Given these constraints, it has become imperative to devise a strategy to filter information intended for rewriting, thus addressing only those cases that present a genuine problem. Thus, we build a Rewrite Candidate Classifier based on the patterns above.

A. Rewrite Candidate Classifier

To improve the query rewriting process and filter non-rewritable queries, a Bert Classifier based on the BERTimbau Base [8] was tuned over 5 epochs. This classifier was trained using 2,400 manually labeled queries for each of 5 retailers representing different segments (e.g. clothing, pharmacy). Labels indicated whether a query could be rewritten (1) or not (0). The main objective of this classifier is to identify queries that do not require rewriting, thus optimizing resource utilization and improving query processing efficiency.

The classifier aims to prioritize queries that benefit from the rewriting process by accurately classifying queries based on their rewriting potential. This ensures a more focused and effective query optimization strategy and minimizes worsening situations, such as unnecessary rewriting of product IDs, specific brands, etc. The Rewrite Candidate Classifier performance evaluation produced the results shown in Table I.

TABLE I
Classification Report

Class	Precision	Recall	F1-score	Support
0	0.93	0.96	0.94	346
1	0.97	0.95	0.96	254
Accuracy			0.95	
Macro avg			0.95	
Weighted avg			0.95	

These results demonstrate the effectiveness of the Rewrite Candidate Classifier in accurately distinguishing between rewritable and non-rewritable queries, achieving a high level of precision, recall, and overall accuracy in query classification.

VI. Models Comparison and Performance Analysis

Based on the classifier model results, we used rewritable queries to compare two LLM models as pipeline rewriters: GPT-4 [1] and an improved version of LLama2 [9]. While the other two models were employed in their original form, the refined LLama2 model was trained using the revised queries from the prior experiment.

We used the same pipeline and dataset as mentioned in the previous section to guarantee an impartial and exhaustive examination. The pipeline was designed to work in two steps: first, it used the RCC to find and eliminate queries that could not be rewritten, and then it applied an LLM-based query rewriting model.

Table II shows the MRR metrics obtained by the models. The results show that GPT-4 performed better than the fine-tuned LLama2 model, with an MRR@20 score of 0.88 against 0.86 for the latter. This shows that, given the context and dataset, GPT-4 might be more successful in producing accurate and pertinent query rewrites.

TABLE II
Comparison of Models based on MRR@20

Model	MRR@20
LLama2 (Fine-tuned)	0.86
GPT-4	0.88

It's essential to recall that the LLama2 model, which was improved by using the rewritten queries from the prior experiment, did not outperform the GPT-4 model in terms of performance. Due to their innate capacity to catch and generalize language patterns, big language models like GPT-4 have the potential to perform exceptionally well in query

rewriting jobs with little to no fine-tuning. However, it is essential to consider the operational costs associated with each approach, despite the potential performance gap, the local operational cost of using the LLama2 model is expected to be lower than that of the GPT-4 model.

VII. Solutions by Tenant

Upon analyzing the results, we observed significant differences in performance across different segments within the query rewriting pipeline. This section focuses on comparing two specific segments: the Clothing Segment and the Pharmacy Segment, highlighting the distinct outcomes and challenges encountered in each.

A. Clothing Segment

The Clothing Segment demonstrated favorable outcomes within the query rewriting pipeline, with a significant improvement in search results. Users in this segment typically search for generic terms, brand names, or product categories, allowing for effective query rewriting and enhanced search relevance. The general nature of clothing queries and the flexibility in interpreting similar products contributed to the success of the rewriting process in this segment.

Figure 2 shows the percentage of products better and worse positioned after the rewrite process. Around 80% of the queries returned product rankings with the item of interest better positioned, showcasing the effectiveness of the query rewriting approach in enhancing product visibility and relevance within the search results.

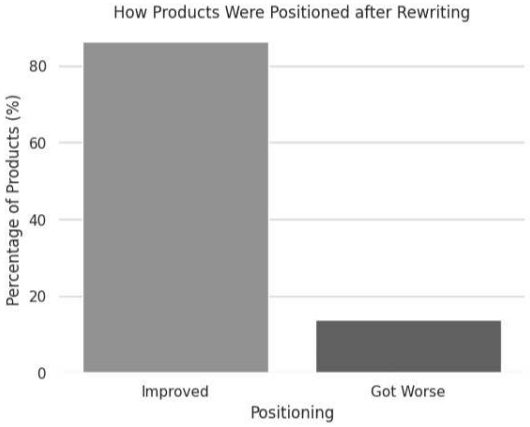

Fig. 2. Clothing Segment

B. Pharmacy Segment

In contrast, the Pharmacy Segment presented unique challenges within the query rewriting pipeline, with a significant proportion of queries resulting in worsened outcomes. Analysis revealed that Pharmacy users tend to search for specific medication names, including dosage and active ingredients, making query rewriting less effective. This effectiveness can be observed in Figure 3 that shows the percentage of products better and worse positioned after the rewrite process, where only 64,25% of products of interest have seen improvements in their ranking, although slightly less positively than the clothing segment.

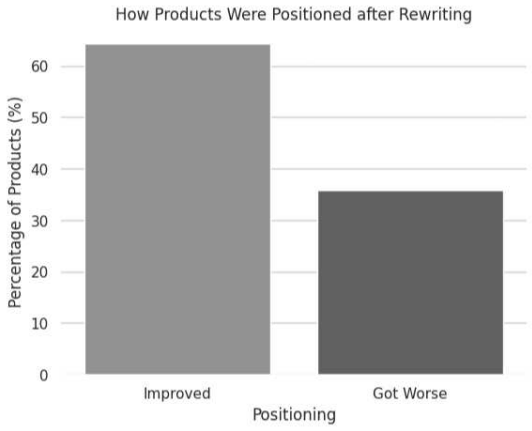

Fig. 3. Pharmacy Segment

The specificity of medication names and the critical nature of accurate information in this domain posed challenges for the rewriting process. Furthermore, the variability in medication quantities and the potential for different active ingredients in similar products added complexity to the rewriting task. Notably, the model's ability to rewrite medicines by active ingredients, such as transforming "Dorflex 10 comprimidos" to "Dipirona 300 10 comprimidos," highlighted the intricacies of rewriting pharmaceutical queries accurately.

The findings underscore the importance of using strategies tailored to individual segments, such as the pharmaceutical segment that deals with information of a critical and extremely specific nature, rather than a one-size-fits-all approach. Strategies such as special prompt adjustments or Retrieval-Augmented Generation (RAG) that combines an external source of knowledge can improve the performance of LLMs on this task. There is a clear need to conduct a thorough study to understand the nuances of each segment before applying the query rewriting approach.

VIII. Conclusions

In this paper, we examine how deep learning models—particularly LLMs—work when it comes to query rewriting in e-commerce search engines. Our research findings show that applying LLMs, such OpenAI's GPT-3.5 Turbo, may dramatically improve user query relevance and accuracy, producing superior search results.

We found that 81.8% of sample queries from a multi-tenant e-commerce platform were successful, with 3.2% of the results remaining neutral and 15% indicating a worsening of the situation. These results illustrate how LLMs could improve online shoppers' search experiences.

Additionally, we compared other models, such as GPT-4 and an improved LLama2 model. We found that GPT performed more effectively than the latter regarding MRR@20, indicating that it was better at providing query rewrites. The results

emphasize the importance of choosing the best LLM model for a given dataset and environment while performing query rewriting jobs.

A detailed examination of the Pharmacy and Clothing segments further highlighted the necessity for customized approaches when rewriting queries. Due to the generic character of the queries, the clothes segment showed promising results; nevertheless, the pharmacy segment offered particular difficulties with drug names, particular active components, etc. Future paths for research could go into more LLM improvements targeted to particular industry domains, incorporating techniques like RAG to ensure query and catalog consistency.

Finally, our research supports the increasing literature encouraging deep learning models in query rewriting to improve search relevancy and e-commerce customization.

REFERENCES

[1] Josh Achiam et al. "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774* (2023).

[2] Yuki Amemiya et al. "How do users revise zero-hit product search queries?" In: *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*. Springer. 2021, pp. 185–192.

[3] Lijuan Chen et al. "Toward the understanding of deep text matching models for information retrieval". In: *arXiv preprint arXiv:2108.07081* (2021).

[4] Yupin Huang et al. "Behavior-driven query similarity prediction based on pre-trained language models for e-commerce search". In: (2023).

[5] Bernard J Jansen and Caroline M Eastman. "Limitations of advanced searching techniques on web search engines". In: *Journal of Electronic Resources in Law Libraries* 1.1 (2006), pp. 55–81.

[6] Youngsoo Kim and Ramayya Krishnan. "On product-level uncertainty and online purchase behavior: An empirical analysis". In: *Management Science* 61.10 (2015), pp. 2449–2467.

[7] Sen Li et al. "Query Rewriting in TaoBao Search". In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2022, pp. 3262–3271.

[8] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. "BERTimbau: pretrained BERT models for Brazilian Portuguese". In: *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*. Springer. 2020, pp. 403–417.

[9] Hugo Touvron et al. "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint arXiv:2307.09288* (2023).

[10] Yaxuan Wang et al. "Queen: Neural query rewriting in e-commerce". In: (2021).

[11] Yutao Zhu et al. "Large language models for information retrieval: A survey". In: *arXiv preprint arXiv:2308.07107* (2023).

[12] Simiao Zuo et al. "Context-Aware Query Rewriting for Improving Users' Search Experience on E-commerce Websites". In: *arXiv preprint arXiv:2209.07584* (2022).