Enhancing E-Commerce with a RAG-Powered Conversational Recommender System

Danilo Xavier Valle[1], **Hilário Tomaz Alves de Oliveira**[1]

[1]Programa de Pós-graduação em Computação Aplicada (PPComp)
Instituto Federal do Espírito Santo (IFES) – Campus Serra

daniloxvalle@gmail.com, hilario.oliveira@ifes.edu.br

Abstract. *Conversational recommender systems have emerged as a promising approach to enhancing user experience in e-commerce by enabling interactive and personalized product discovery. This paper proposes a conversational recommender system for e-commerce that employs retrieval-augmented generation (RAG) to improve product recommendations based on natural language queries. Experiments were conducted using the ESCI-S dataset, an enriched version of the Amazon ESCI shopping queries dataset, to evaluate embedding models and large language models. The goal of this study is to assess the effectiveness of an RAG-based conversational recommender system and to identify optimal configurations for enhanced performance in e-commerce applications.*

1. Introduction

The natural language processing (NLP) field has been impacted by unprecedented advances in the development of large language models (LLMs) [Naveed et al. 2023]. These state-of-the-art models have improved human language processing and generation, enabling the development of conversational systems that facilitate more intuitive and effective human-machine interactions [Feng et al. 2023].

The e-commerce industry is one of the most important sectors of the economy [Iglesias-Pradas and Acquila-Natale 2023]. With the proliferation of online shopping, e-commerce platforms have become an essential part of people's lives. However, the vast number of products available on these platforms can make it difficult for users to find what they are looking for. Users often have to search through hundreds of products to find the one that best suits their needs.

Classical recommender systems in e-commerce face limitations due to their reliance on historical data and predefined algorithms, which struggle to adapt to rapidly changing user behaviors and preferences [Roy and Dutta 2022]. The "cold start" problem further compromises performance, as systems require substantial data to provide accurate recommendations, often delivering irrelevant suggestions for new users [Patro et al. 2023]. Additionally, these systems focus heavily on algorithmic accuracy but give little attention to user interaction. This limitation means that while the recommendations might be technically well-founded, they fail to engage users in a meaningful and interactive manner. This lack of interactive engagement makes the suggestions feel impersonal or out of touch with the user's immediate needs and preferences. Choice overload, also known as "Overchoice", is a cognitive process in which people have difficulty making a decision when faced with many options [Hu et al. 2023]. In the context of e-commerce, choice overload can lead to a poor user experience, as users may become

overwhelmed by the number of products available and struggle to find the one that best suits their needs.

A conversational agent in e-commerce has the potential to transform user interactions with online platforms by offering personalized product recommendations through natural language interactions [Benita et al. 2024]. Unlike conventional search and filter interfaces, which require manual product searches, conversational agents may help users find desired products by understanding their preferences and providing tailored recommendations.

This work aims to develop and evaluate a retrieval-augmented generation (RAG) based approach for product recommendations in e-commerce environments. The potential of RAG architecture in recommender systems still needs to be further investigated [Deldjoo et al. 2024], and this work seeks to contribute to filling this gap. Experiments were conducted to assess the performance across two stages of the RAG architecture and to compare the effectiveness of different embedding models and LLMs. Furthermore, it intends to offer benchmarks for selecting the most suitable technologies to enhance system efficiency and effectiveness.

The main contributions of this research include the development of a RAG-powered conversational recommender system specifically designed for e-commerce product discovery through natural language interactions. Additionally, multiple embedding models and LLMs within the RAG architecture were evaluated to identify optimal configurations for product recommendations. The research provides benchmarks and configuration guidelines to enhance system efficiency in real-world e-commerce scenarios, offering insights for practical implementation.

2. Related work

In the domain of conversational recommender systems (CRS), Jannach et al. [Jannach et al. 2021] provide a survey highlighting the shift from traditional one-shot recommender systems to interactive, dialogue-based approaches. Their work underscores the ability of CRS to enhance preference elicitation and user engagement through natural language interactions, leveraging advancements in natural language processing and chatbot technologies.

Chat-Rec [Gao et al. 2023] is a framework that leverages LLMs to build conversational recommender systems. By transforming user profiles and past interactions into prompts, it facilitates natural dialogues that deliver more personalized and context-aware recommendations. Its primary advantage is the capability to capture user preferences and link users to products through in-context learning.

Nguyen et al. [Nguyen et al. 2024] developed an Artificial Intelligence (AI) chatbot for tourist recommendations in Vietnam, addressing the challenge of limited destination information for foreign visitors. Their system integrates machine learning algorithms to classify question topics, predict user intent, and generate contextually relevant responses. The chatbot was deployed as a mobile application, enabling users to access location information and hotel prices, as well as participate in interactive question-and-answer (Q&A) sessions.

RecAI [Lian et al. 2024] is a toolkit designed to leverage LLMs to develop recom-

mender systems that emulate human-like interactions. It is built on several key components, each addressing various real-world applications through different techniques. For instance, engineers can use the recommender AI agent framework to evolve their industrial recommender systems into conversational interfaces, and researchers can utilize the chat-rec framework to develop conversational recommender systems.

Exploring knowledge-enhanced recommendations, Xiao et al. [Xiao et al. 2024] present a novel e-commerce recommendation system using a RAG-based framework, which integrates customer review data as an external knowledge source to enhance recommendation scalability and personalization. This work demonstrates the efficacy of RAG in processing review data for user-friendly, context-aware recommendations.

In the context of customer service applications, Benita et al. [Benita et al. 2024] explore the implementation of RAG-based chatbot systems for real-time customer support in e-commerce, combining retrieval and generative models to deliver accurate, contextually relevant responses to diverse customer queries. Their RAG-based approach enhances scalability and user satisfaction by addressing the limitations of traditional chatbots, including the ability to handle complex user queries.

This work advances the field by systematically evaluating a RAG-based approach in the context of conversational recommender systems for e-commerce, with a particular focus on integrating product metadata from the ESCI-S dataset. In contrast to previous research, which primarily uses customer reviews [Xiao et al. 2024] or targets customer service applications [Benita et al. 2024], our approach leverages detailed product descriptions and semantic matching to enhance recommendation accuracy. By comparing multiple embedding models and LLMs within the RAG pipeline, we provide a benchmark for optimizing system performance. This analysis addresses existing gaps in the literature concerning the practical implementation of RAG-based CRS in e-commerce scenarios. Furthermore, this work contributes to the underexplored application of RAG in recommender systems, as highlighted by [Deldjoo et al. 2024], by providing insights into effective configurations for enhancing system efficiency.

3. Method

Retrieval-augmented generation (RAG) is a technique that improves LLM performance by incorporating external knowledge sources into the generation process [Lewis et al. 2020]. The retrieved data is used to enrich the model's input, adding contextual and factual grounding to the output. This approach mitigates common limitations of traditional LLMs, such as hallucinations and outdated information, while preserving their flexibility and fluency [Jing et al. 2024]. RAG has demonstrated strong potential in applications like question answering, conversational AI, and content generation, particularly in scenarios requiring up-to-date and domain-specific information [Fan et al. 2024].

This work proposes an approach for an e-commerce recommender system using a RAG pipeline, as depicted in Figure 1. Each product description is transformed into a vector representation using an embedding model, capturing its semantic essence for query matching. These vectors are stored in a vector database to enable product retrieval. When a user query is received, the system generates a vector representation of the query using the same embedding model. The vector database is then queried to retrieve the most relevant product vectors based on their similarity to the query vector. The retrieved

products are passed to a LLM, which generates a recommendation that includes the most relevant products.

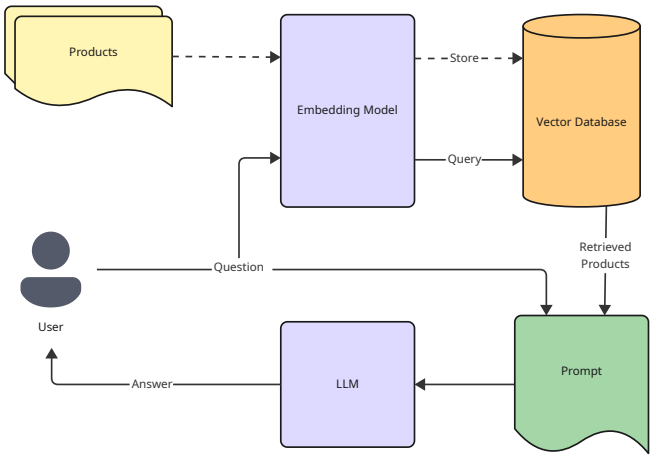

Figure 1. RAG pipeline for e-commerce conversational recommender systems.

3.1. Dataset

To assess the performance of the proposed RAG pipeline, we adopted the Amazon ESCI Dataset[1], which is a shopping queries dataset released by Amazon in 2022 [Reddy et al. 2022]. It is a comprehensive collection of search queries designed to facilitate research on semantic matching between queries and products. The dataset comprises 2.6 million queries, each linked to a set of products and their corresponding relevance labels. The relevance labels are categorized into four classes: Exact (E), Substitute (S), Complement (C), and Irrelevant (I). An example of a query-product relevance pair is shown in Table 1.

Table 1. Example of query and product relevance from the ESCI dataset.

Query	Product ASIN	ESCI label	Score
tsp of grenetine without sweetener	B01N0SLA5L	E	1.0
	B07DK5C8QG	E	1.0
	B000121A7O	E	1.0
	B00CWTZ8UY	S	0.1
	B01JI1EV4C	S	0.1
	B000F3N7AC	C	0.01

The Extended Metadata for Amazon ESCI (ESCI-S)[2] is a complementary dataset that enhances the original Amazon ESCI dataset by offering additional information about the products. This complementary dataset includes metadata such as product categories, prices, images, and other attributes. The extended metadata enables researchers to derive subsets of the dataset based on specific criteria, such as product categories. An example of a product from the ESCI-S dataset is shown in Table 2, with some columns omitted and the description truncated for the sake of simplicity.

[1] https://github.com/amazon-science/esci-data
[2] https://github.com/shuttie/esci-s

Table 2. Example of product extracted from the ESCI-S dataset.

ASIN	Title	Category	Description
B000R4K1D0	Blue Diamond Almonds Nut Thin Crackers Crisps, Smokehouse, 4.25 Oz	Grocery & Gourmet Food	Gone are the days when gluten-free meant boring. Nut Thins are a crunchy cracker loaded with nutritious almonds and baked to perfection. Wonderful as an appetizer and ideal for snacking. With 3 grams of protein per serving, they're a great afternoon snack...

The ESCI-S dataset is a large-scale resource containing 1.66 million products, which is significantly more than the number typically found on standard e-commerce platforms. The "Grocery & Gourmet Food" category within this dataset is particularly interesting, as it can simulate a real-world supermarket e-commerce scenario. This category comprises 30,367 products, representing a substantial subset of the overall dataset.

To ensure the quality and relevance of the dataset for our experiments, we applied several filtering criteria. First, we selected only product descriptions with a length between 1,500 and 2,000 characters, resulting in a subset of 3,121 products from the "Grocery & Gourmet Food" category. This range was chosen to focus on products with sufficiently detailed descriptions for semantic matching. Additionally, we filtered the queries to include only those that return between 2 and 20 products, ensuring each query is associated with a meaningful set of candidate products. After this filtering, the dataset comprises 1,061 distinct queries.

An enriched version of the original query dataset was created to better simulate real-world user interactions with conversational recommender systems. While the original ESCI dataset contains primarily keyword-based search queries, these terse formulations do not reflect how users naturally interact with conversational assistants. The enrichment process transformed these compact queries into more natural, conversational questions that a customer might ask when speaking to a shop assistant, as shown in Table 3. This transformation creates more realistic test data for evaluating the RAG pipeline's ability to handle natural language inputs, provides a richer semantic context for the embedding models to work with, and helps bridge the gap between traditional search-oriented e-commerce interactions and the conversational approach being explored in this study. The enriched queries were generated using the OpenAI-o3[3] model and maintain the same intent and product requirements as the original queries, but are expressed in a more human-like conversational manner.

3.2. Embedding models and LLMs evaluated

Several embedding models were evaluated to generate vector representations of product descriptions and user queries within the conversational recommender system. The selected models, summarized in Table 4, include both proprietary and open-source options, offering a range of configurations in terms of embedding dimensions, token limits, and multilingual support. Pricing information for OpenAI models was obtained from their

[3] https://openai.com/index/introducing-o3-and-o4-mini/

Table 3. Examples of enriched queries.

Query	Enriched query
0 cal water inhancer without sugar	Do you have sugar-free, zero-calorie water enhancers?
1 huba buba origanal not exspencive	Show me an inexpensive pack of original Hubba Bubba gum.
cherries without stems	Do you have stemless cherries?

official pricing page[4], for Google's models from the Gemini API documentation[5], and for open-source models from DeepInfra[6], a platform that provides API access to a variety of open-source machine learning models.

Table 4. Evaluated embedding models.

Model	Dimension	Max tokens	Model size(M)	Open source	Multi-lingual	Price $/Mtok
text-embedding-3-large	3072	8191	-	No	Yes	0.13
text-embedding-3-small	1536	8191	-	No	Yes	0.02
text-embedding-004	768	2048	1200	No	No	Free
Multiling. E5 Large	1024	512	560	Yes	Yes	0.01
Multiling. E5 Large Instruct	1024	512	560	Yes	Yes	0.01
all-mpnet-base-v2	768	384	420	Yes	No	0.005
all-MiniLM-L6-v2	384	512	23	Yes	No	0.005

As with the embedding models, several LLMs were evaluated for their ability to generate natural language responses within the recommender system. The selected LLMs, summarized in Table 5, were chosen to strike a balance between performance and cost, with model sizes ranging from 8 to 30 billion parameters. To establish a high-end benchmark, GPT-4.1 was included in the evaluation despite its significantly higher computational cost. Pricing for OpenAI models was obtained from the official pricing page, while costs for open-source models were retrieved via DeepInfra.

Table 5. Evaluated LLMs.

Language model	Model size	Max tokens	Open source	Input price $/Mtok	Output price $/Mtok
GPT-4.1	-	1M	No	2.00	8.00
GPT-4o-mini	-	128K	No	0.15	0.60
Qwen3-30B-A3B	30B	32K	Yes	0.10	0.30
Gemma-3-27b-it	27B	128K	Yes	0.10	0.20
Mistral-Small-24B-Instruct	24B	32K	Yes	0.07	0.14
Llama-4-Scout-17B-16E-Instruct	17B	10M	Yes	0.08	0.30
Llama-3.1-8B-Instruct-Turbo	8B	128K	Yes	0.02	0.03

The open-source PostgreSQL database, combined with the pgvector extension[7],

[4] https://platform.openai.com/docs/pricing
[5] https://ai.google.dev/gemini-api/docs/pricing
[6] https://deepinfra.com/
[7] https://github.com/pgvector/pgvector

was adopted as the vector database due to its flexibility and strong community support. The pgvector extension enables efficient storage and similarity search of embeddings directly within the database. This choice was motivated by the cost-effectiveness and accessibility of open-source solutions, offering robust functionality without reliance on proprietary systems.

The prompt template, shown in Listing 1, guides the LLM in generating concise and relevant product recommendations based on the product context retrieved through the RAG pipeline. It enforces strict adherence to the provided product information, requires explicit inclusion of each product's ASIN and name, and restricts the response to a maximum of four products and forty words to prevent overly long outputs and excessive product listings.

Listing 1. LLM prompt template for product recommendations.

```
Product context:
----------------
{context}
----------------
You are an online sales assistant. Use only the information provided in the product
    context to answer the customer's question. If there are suitable products, promote
    them by explicitly mentioning the product ID and the product name in your response.
    The product ID must appear in brackets, for example, [1234], immediately before the
    product name. Always include the product ID whenever you refer to a product. Include
     up to four relevant products without numbering them. If no product meets the
    customer's needs, inform them politely and suggest alternative solutions. If a
    product is irrelevant to the question, ignore it. Demonstrate expertise, give clear
    and concise information, and use a positive, engaging tone to encourage purchase.
Customer question:
----------------
{question}
----------------
Respond briefly, using at most 40 words.
```

4. Experimental Methodology

Two sets of experiments were conducted to evaluate the effectiveness of the proposed RAG-based conversational recommender system. The first experiment focused on evaluating the performance of embedding models in the context of e-commerce product retrieval. The goal was to identify which models most effectively capture the semantic relationship between user queries and product descriptions. Two standard information retrieval metrics, Normalized Discounted Cumulative Gain (nDCG) and Recall, were used for evaluation.

The nDCG is a practical metric because it considers the position of relevant items in the ranked retrieval results, thereby reflecting the quality of the ranking. The Product ESCI scores, as illustrated in Table 1, were used to compute the nDCG. Recall measures the proportion of relevant products successfully retrieved, making it crucial for assessing the system's ability to retrieve all relevant products. High recall is particularly important in a RAG pipeline, where the LLM component is typically more resource-intensive than the embedding model. Effective retrieval reduces the number of products passed to the generation step, thereby lowering the associated computational costs. Both metrics were computed for the top-k retrieved products, with k ranging from 1 to 20.

The second experiment assessed the performance of different LLMs in generating product recommendations in a conversational context. The objective was to determine which LLMs most effectively transform retrieved products into relevant and concise

recommendations. For this evaluation, retrieval results from the *multilingual-e5-large-instruct* model, identified as the top-performing open-source embedding model in the first experiment, were used. Product recommendations were generated based on the top 20 retrieved items.

To evaluate the quality of the product recommendations generated by the LLMs, we adopted a metric called the product recommendation score. This metric is inspired by nDCG but adapted for scenarios where the order of recommended items is not critical. Instead of measuring ranking quality, it focuses solely on the relevance of the selected products. For each query, we first calculate an ideal score by summing the ESCI relevance scores of the top four products retrieved, those with the highest possible relevance values, since the system is instructed to recommend at most four items. Next, we compute the actual score by summing the ESCI scores of the four products recommended by the LLM. The normalized recommendation score is then obtained by dividing the actual score by the ideal score, resulting in a value between 0 and 1. This score indicates how closely the model's recommendations align with the most relevant possible selection. The overall product recommendation score for an LLM is defined as the average of the normalized scores across all evaluation queries.

5. Results

5.1. Experiment 1 - Embedding models performance

Figure 2 and Figure 3 show the nDCG and recall obtained in this first experiment. The results demonstrate a consistent performance ranking across all embedding models for both metrics. The proprietary model from OpenAI, text-embedding-3-large, set the benchmark for retrieval accuracy and ranking quality, followed closely by Gemini 004. The all-MiniLM-L6-v2 model, characterized by a much smaller model size and embedding dimension, ranked lowest in performance.

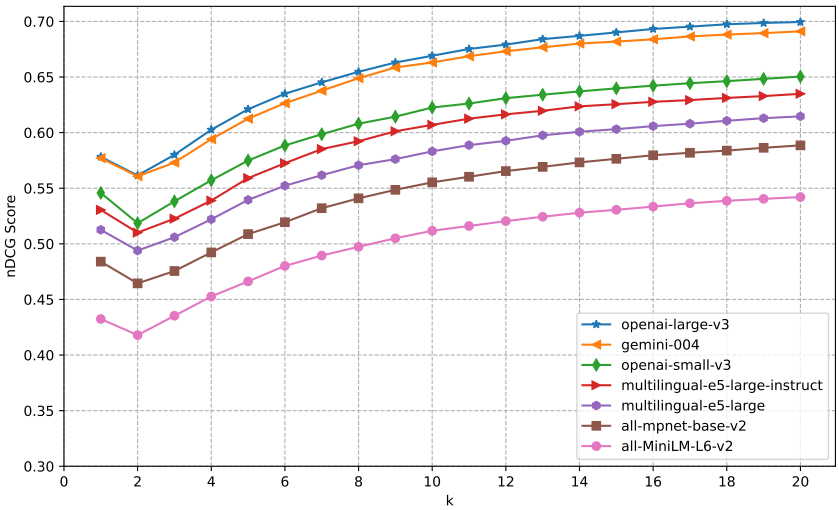

Figure 2. nDCG Scores for k values from 1 to 20.

In the context of implementing a conversational recommender system, the cost of the embedding model is a pivotal consideration. The text-embedding-3-large model offers

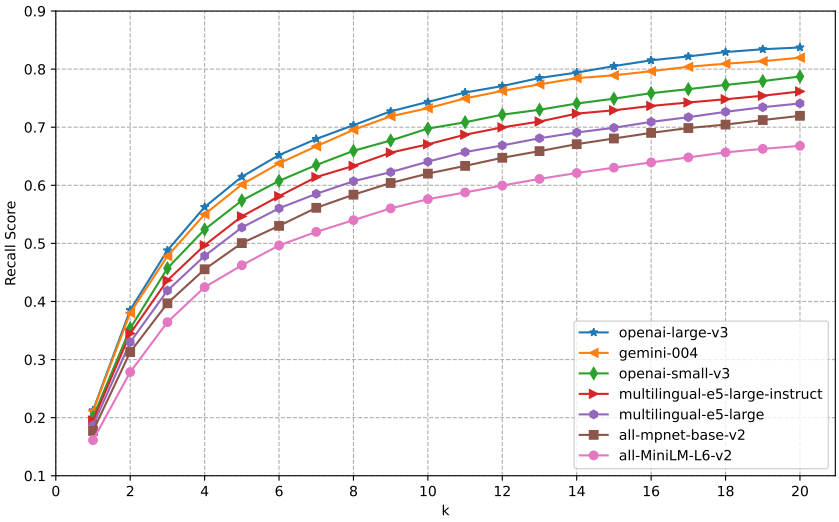

Figure 3. Recall Scores for k values from 1 to 20.

superior performance but incurs a significantly higher cost compared to other models tested. In contrast, the Gemini 004 model provides a comparable performance at no cost, as it is free to use; however, it lacks multilingual support, which may limit its suitability for certain applications.

A critical architectural consideration is whether to adopt proprietary or open-source models. Opting for an open-source alternative offers significantly more flexibility and data privacy, allowing the system to run on any chosen infrastructure, including on-premise. This flexibility can reduce operational costs. The multilingual-e5-large-instruct model offers a compelling open-source option with competitive results. It consistently outperformed its non-instruct counterpart, multilingual-e5-large, and delivered results comparable to OpenAI's text-embedding-3-small model, while featuring a smaller embedding dimension and a lower cost. This result highlights the potential of open-source models and the impact of instruction-tuning in enhancing embedding model capabilities.

5.2. Experiment 2 - Large language models performance

The second experiment assessed how well different LLMs generate product recommendations based on retrieval results from the proposed RAG pipeline. The goal was to measure each model's ability to produce accurate and relevant recommendations from a set of retrieved products. The graph in Figure 4 illustrates the product recommendation score, as described in Section 4, for the evaluated LLMs.

GPT-4.1 achieved the highest score of approximately 0.65, significantly outperforming other models, likely due to its large model size and extensive training data. This result establishes it as a benchmark for high-end performance, though its high cost may limit its usability in a recommender system. Examples of the recommendations generated by GPT-4.1 are provided in Table 6, illustrating its usage in real-world scenarios. It should be noted that product IDs shown in brackets are excluded from the final recommendations presented to users, as they serve only as internal references and for computing the product recommendation score.

The mid-tier performers include models with sizes ranging from 17B to 30B pa-

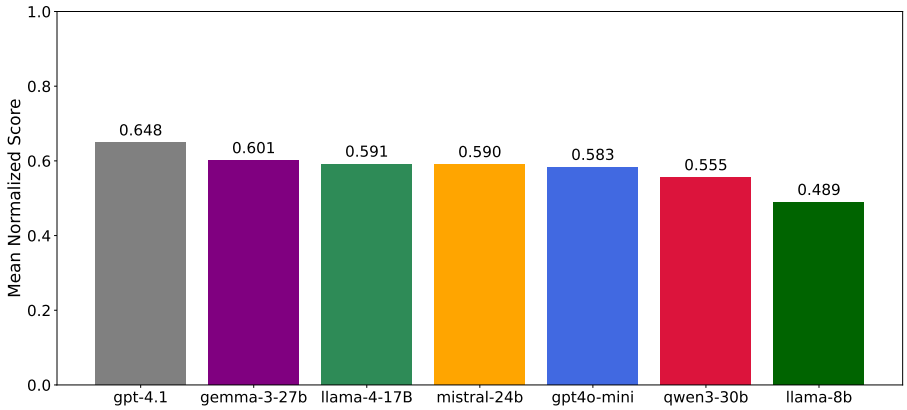

Figure 4. LLM Product Recommendation Score.

rameters. The Gemma 3 27B, Llama 4 17B, Mistral 24B, and GPT-4o-mini achieved similar performance, scoring around 0.6. These models, while not as powerful as GPT-4.1, demonstrated reasonable recommendation capabilities. Despite its 30B parameters, Qwen3-30B scored lower at about 0.55, suggesting that size does not guarantee better performance. Meanwhile, the smaller Llama 3.1 8B scored around 0.50, indicating reduced accuracy in recommendations compared to larger models.

Table 6. Examples of generated product recommendations.

Enriched query	**LLM Answer**
Do you have sugar-free, zero-calorie water enhancers?	Absolutely! We have several sugar-free, zero-calorie water enhancers, including [B078PR2QSD] Jelly Belly - Water Enhancer, [B01A2L9PHU] Stur Liquid Water Enhancer, [B087XMB16P] Twinings Cold Infuse, and [B07FDQ88L3] Jelly Belly Drink Mix - Variety Pack.
Do you have black cumin seed powder?	We currently do not have black cumin seed powder, but we offer whole black cumin seeds such as [B094S8BMPZ] Black Cumin Seeds, [B074N319BK] Organic Black Cumin Seed Whole, and [B00L2FLETS] Kevala Organic Raw Black Cumin Seeds. You can grind them at home!

The product recommendation scores, ranging from approximately 0.5 to 0.65, indicate that even top-performing models faced challenges in generating precise product recommendations, likely due to the complexity of the Amazon ESCI dataset, which was designed for keyword-based search queries rather than natural language interactions.

From a cost-benefit perspective, the open-source mid-tier models strike a compelling balance since their input costs range from $0.07 to $0.10 per million tokens (Mtok) and output costs from $0.14 to $0.30 per Mtok, roughly one-tenth to one-twentieth of GPT-4.1's $2.00/$8.00 rates, while still delivering nearly 90% of its recommendation quality. When compared to GPT-4o-mini, which costs $0.15/$0.60 per Mtok, the open-source models offer a more cost-effective solution with a comparable performance.

Conversely, smaller models such as Llama-3.1 8B reduce inference costs even further (to $0.02/$0.03 per Mtok), but at the expense of a steeper 20% performance decline,

making them suitable only when the cost outweighs precision. These trade-offs highlight that model selection should be guided by the target service-level agreement. High performance and better experiences may justify GPT-4.1, whereas cost-sensitive or high-volume scenarios benefit more from the latest open-source alternatives.

6. Conclusions and Future Work

This work presented a retrieval-augmented generation (RAG)-powered conversational recommender system designed to enhance product discovery in e-commerce through natural language interactions. By leveraging the ESCI-S dataset, enriched with detailed product metadata and conversationally reformulated queries, the system simulated real-world user interactions in online retail settings. The combination of embedding models and LLMs within the RAG architecture demonstrated potential in delivering accurate and contextually relevant product recommendations, addressing the limitations of traditional recommender systems, such as a lack of personalization and cold start problems.

Experimental evaluations provided insights into the performance and cost-effectiveness of the models. Proprietary embedding models, such as OpenAI's text-embedding-3-large, set the benchmark for retrieval accuracy. In contrast, open-source alternatives like multilingual-e5-large-instruct offer competitive results with greater flexibility and lower costs. Among LLMs, GPT-4.1 achieved the highest recommendation accuracy; however, mid-tier open-source models (Gemma 3 27B, Llama 4 17B, Mistral 24B) delivered nearly 90% of its quality at a fraction of the cost, making them viable for scalable applications.

Looking ahead, future work will focus on assessing the quality of generated responses, evaluating LLM performance not only based on their ability to cite relevant retrieved products but also on the overall quality of answers. Moreover, hybrid strategies that integrate RAG with other recommender system methodologies, such as collaborative filtering and content-based filtering, will be investigated to enhance personalization. Additional research directions include the development of a conversational recommender system in a production environment, with a focus on scalability and integration with existing e-commerce infrastructures.

References

Benita, J., Tej, K. V. C., Kumar, E. V., Subbarao, G. V., and Venkatesh, C. (2024). Implementation of retrieval-augmented generation (rag) in chatbot systems for enhanced real-time customer support in e-commerce. In *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pages 1381–1388. IEEE.

Deldjoo, Y., He, Z., McAuley, J., Korikov, A., Sanner, S., Ramisa, A., Vidal, R., Sathiamoorthy, M., Kasirzadeh, A., and Milano, S. (2024). A review of modern recommender systems using generative models (gen-recsys). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6448–6458.

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., and Li, Q. (2024). A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Feng, Y., Liu, S., Xue, Z., Cai, Q., Hu, L., Jiang, P., Gai, K., and Sun, F. (2023). A large language model enhanced conversational recommender system. *arXiv preprint arXiv:2308.06212*.

Gao, Y., Sheng, T., Xiang, Y., Xiong, Y., Wang, H., and Zhang, J. (2023). Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524*.

Hu, X., Turel, O., Chen, W., Shi, J., and He, Q. (2023). The effect of trait-state anxiety on choice overload: the mediating role of choice difficulty. *DECISION*, 50.

Iglesias-Pradas, S. and Acquila-Natale, E. (2023). The future of e-commerce: Overview and prospects of multichannel and omnichannel retail. *Journal of Theoretical and Applied Electronic Commerce Research*, 18(1):656–667.

Jannach, D., Manzoor, A., Cai, W., and Chen, L. (2021). A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)*, 54(5):1–36.

Jing, Z., Su, Y., Han, Y., Yuan, B., Xu, H., Liu, C., Chen, K., and Zhang, M. (2024). When large language models meet vector databases: A survey. *arXiv preprint arXiv:2402.01763*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Lian, J., Lei, Y., Huang, X., Yao, J., Xu, W., and Xie, X. (2024). Recai: Leveraging large language models for next-generation recommender systems. In *Companion Proceedings of the ACM on Web Conference 2024*, WWW '24, page 1031–1034, New York, NY, USA. Association for Computing Machinery.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Nguyen, H., Tran, T., Nham, P., Nguyen, N., and Duy Anh, L. (2024). Ai chatbot for tourist recommendations: A case study in vietnam. *Applied Computer Systems*, 28:232–244.

Patro, S. G. K., Mishra, B. K., Panda, S. K., Kumar, R., Long, H. V., and Taniar, D. (2023). Cold start aware hybrid recommender system approach for e-commerce users. *Soft Computing*, 27(4):2071–2091.

Reddy, C. K., Màrquez, L., Valero, F., Rao, N., Zaragoza, H., Bandyopadhyay, S., Biswas, A., Xing, A., and Subbian, K. (2022). Shopping queries dataset: A large-scale esci benchmark for improving product search. *arXiv preprint arXiv:2206.06588*.

Roy, D. and Dutta, M. (2022). A systematic review and research perspective on recommender systems. *Journal of Big Data*, 9(1):59.

Xiao, G., Wu, J., and Tseng, S.-P. (2024). A novel e-commerce recommendation system based on rag and pretrained large model. In *2024 International Conference on Orange Technology (ICOT)*, pages 1–4.