CatalogRAG: Retrieval-Guided LLM Prediction for Multilingual E-commerce Product Attributes

Bryan Zhang
bryzhang@amazon.com
Amazon
USA

Suleiman Khan
suleimkh@amazon.com
Amazon
USA

Stephan Walter
sstwa@amazon.de
Amazon
Germany

Abstract

E-commerce stores increasingly use Large Language Models (LLMs) to enhance catalog data quality through automated regeneration. A critical challenge is accurately predicting missing structured attribute values across multilingual product catalogs, where LLM performance varies significantly by language. While existing approaches leverage general knowledge through prompt engineering and external retrieval, more effective and accurate signals for attribute prediction can exist within the catalog ecosystem itself - similar products often share consistent patterns and structural relationships, and may have the missing attributes filled. Therefore, this paper introduces `CatalogRAG`, a novel retrieval-augmented system that strategically leverages existing product catalog entries to guide LLM predictions for missing attributes. Our approach introduces a multi-stage retrieval framework that progressively refines the search space based on product type, uses textual similarity, glance views and brand relationships to identify the most relevant attribute-filled examples for LLM prediction guidance. Experiments on test sets across three major e-commerce stores in different languages (US, DE, FR) demonstrate substantial improvements in catalog data quality, achieving up to 34% increase in recall and 0.8% in precision for attribute value prediction. At catalog entry level, it also achieves up to +43.32% increase in completeness and up to +2.83% in correctness.

CCS Concepts

• **Computing methodologies** → *Natural language processing*; • **Information systems** → *Information retrieval.*

Keywords

E-commerce, Multilingual Catalog Enrichment, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Few-shot Learning, Knowledge Retrieval

ACM Reference Format:
Bryan Zhang, Suleiman Khan, and Stephan Walter. 2025. CatalogRAG: Retrieval-Guided LLM Prediction for Multilingual E-commerce Product Attributes. In *Proceedings of LLM4ECommerce Workshop at the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (LLM4ECommerce Workshop at KDD'25).* ACM, New York, NY, USA, 7 pages. https://doi.org/

1 Introduction

Product catalogs are the backbone of e-commerce stores, serving as crucial resources for customers, sellers, and internal teams. They play a pivotal role in enhancing user experience, facilitating product discovery, and driving sales. The use of Large Language Models (LLMs) to regenerate and improve these catalogs has gained significant traction. A typical catalog product entry primarily consists of two textual parts: unstructured attributes (UAs) such as product titles, and structured attributes (SAs) like *color* and *material*. One major challenge in this process is accurately predicting missing values for the SAs of the catalog entries. Our observations indicate that, on average, nearly half of the relevant SA values are missing (empty) for a given entry across product types, highlighting the widespread nature of incomplete information in e-commerce product catalogs.

Predicting missing structured attributes (SA) in multilingual e-commerce catalogs presents several significant challenges: (1) Given the global nature of e-commerce, catalogs often span multiple languages to serve worldwide stores and enable multi-lingual product discovery [2, 9, 11, 13, 16, 17]. While LLMs typically excel in widely-used languages like English, they may exhibit reduced performance in less common languages or those with limited training data, leading to inconsistent attribute prediction quality across different stores. (2) Although it is common to incorporate all available product information such as titles and descriptions to predict attributes, not all missing attributes can be easily predicted or inferred solely from this information and static metadata, even with the model's latent knowledge—particularly as new products and attributes emerge constantly in worldwide e-commerce. (3) Furthermore, while recent efforts have explored Retrieval-Augmented Generation (RAG) to address the inherent limitations of LLMs in accessing specialized knowledge, external knowledge sources cannot effectively capture the specific conventions and norms that exist within product types, stores, and seller practices in constantly growing catalogs.

For effective attribute prediction using LLM in e-commerce catalogs, retrieved information needs to capture specific structural relationships and conventions that exist within product listings, especially where attributes follow store-specific patterns. Examining product listings in e-commerce catalogs, we observe a common pattern: while individual products may have missing attribute values, similar product entries within the same category often have this information filled. Moreover, such a focused set of similar entries

LLM4ECommerce Workshop at KDD'25, Toronto, ON, Canada

https://doi.org/

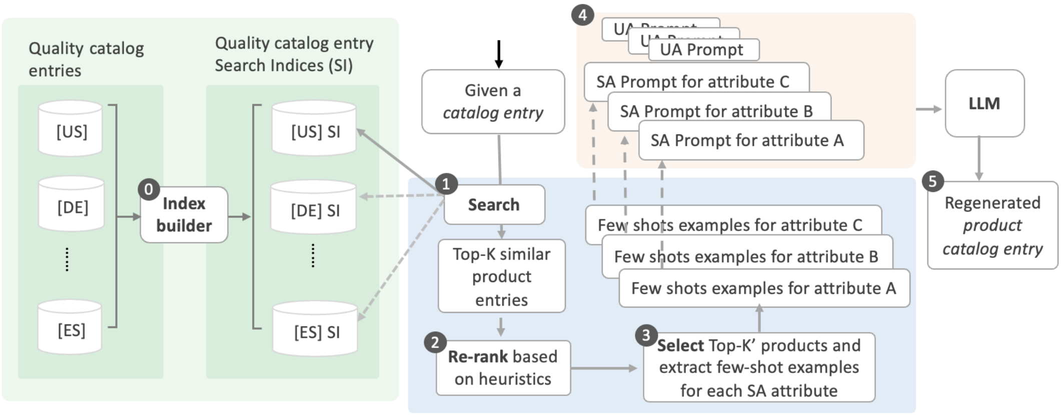
Quality catalog entries
[US]
[DE]
[ES]
0 Index builder
Quality catalog entry Search Indices (SI)
[US] SI
[DE] SI
[ES] SI
Given a catalog entry
1 Search
Top-K similar product entries
2 Re-rank based on heuristics
3 Select Top-K' products and extract few-shot examples for each SA attribute
Few shots examples for attribute C
Few shots examples for attribute B
Few shots examples for attribute A
4
UA Prompt
UA Prompt
UA Prompt
SA Prompt for attribute C
SA Prompt for attribute B
SA Prompt for attribute A
LLM
5 Regenerated product catalog entry

Figure 1: Overview of CatalogRAG: The system builds search indices for quality catalog entries in advance (0). At runtime, given a catalog entry with missing structured attributes, the system retrieves similar products for a given catalog entry from the respective index (1), applies heuristic-based reranking (2), selects relevant few-shot examples for each structured attribute (3), incorporates them into attribute-specific prompts (4), and generates predictions using LLM (5).

naturally encodes business rules, category conventions, and brand-specific patterns through their complete listings. These patterns, when strategically leveraged, can guide LLM predictions more effectively than external knowledge sources, particularly in multilingual environments where maintaining language/store-specific conventions is crucial.

Therefore, in this paper, we introduce `CatalogRAG`, a system that strategically leverages existing product entries to guide LLM predictions for missing structured attributes in e-commerce catalogs. Our system implements a multi-level filtering strategy to identify and utilize implicit patterns from similar products within the store of same language. This approach not only can improve prediction performance but also ensure consistency with existing catalog patterns while adapting to language-specific nuances. Our proposed `CatalogRAG` includes a term-based and probabilistic text-based retrieval mechanism that efficiently identifies similar products based on product title similarities within identical product categories. This is followed by an novel heuristics-based reranking mechanism that utilizes brand and glance view to optimize the search results by balancing brand consistency with overall similarity scores. The system then performs selective sampling, extracting up to three highly relevant few-shot examples per missing attribute from the reranked list of the identified similar products. These examples are carefully curated and integrated into attribute-specific prompts for contextually-informed predictions using LLM.

`CatalogRAG` consistently outperforms the baseline prompting approach across three major e-commerce stores (US, DE, FR). At the structured attribute level, it achieves improvements of up to 34% in recall and 0.8% in precision. At the catalog entry level, it achieves substantial improvements of up to +43.32% in *completeness* and up to +2.83% in *correctness*[1].

Contributions: We have two contributions in this paper: (1) A novel pattern-aware approach to RAG that strategically leverages similar products within the catalog ecosystem, demonstrating that internal catalog patterns can be more effective than general knowledge sources for structured attribute prediction, particularly in multilingual environments. (2) A catalog-aware framework within `CatalogRAG` for identifying and utilizing implicit patterns through strategic few-shot examples, combining product type constraints, title similarity, glance views and brand alignment to guide LLM predictions.

The rest of this paper is organized as follows: Section 2 details our approach, including the retrieval framework selection and the multi-stage prediction process. Section 3 describes our experimental setup, including dataset characteristics, system implementation details, and evaluation metrics. In Section 4, we present and analyze our results. Finally, we conclude with a discussion of our findings and potential directions for future work.

2 Method: CatalogRAG system

2.1 System Overview

Our approach leverages the inherent patterns within the catalog ecosystem - where similar products within the same language store share consistent attribute structures, brand-specific conventions, and category-specific value formats. For example, running shoes from the same brand in the German store often follow similar pattern in describing their sport type, while fashion items within the same category in the French store share consistent terminology for materials and styles. This stems from a fundamental observation: similar products, particularly those sharing the same product type and brand within a language store, often exhibit consistent patterns in their attribute values. `CatalogRAG` combines efficient retrieval mechanisms with strategic example selection to identify these informative patterns for predicting missing structured attributes.

[1] Details on the catalog entry-level evaluation metrics is provided in Section 3.3

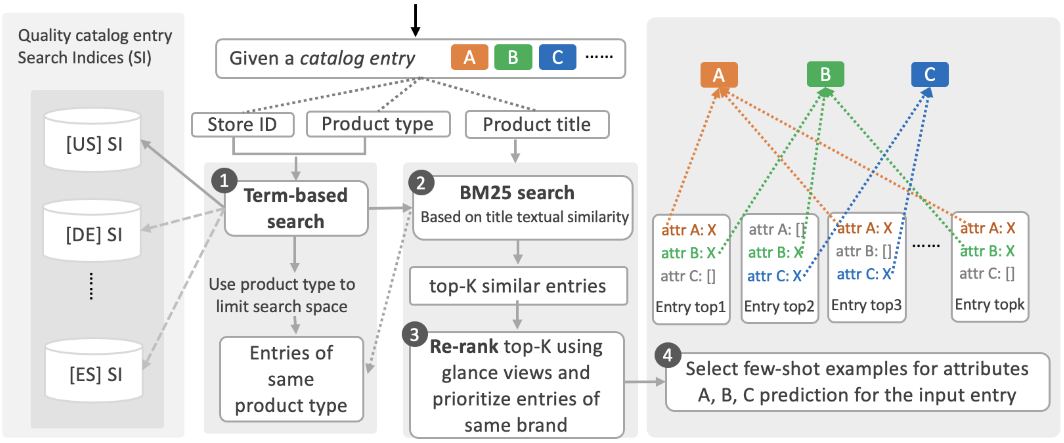
Quality catalog entry Search Indices (SI)
[US] SI
[DE] SI
[ES] SI
Given a catalog entry
A
B
C
Store ID
Product type
Product title
1
Term-based search
Use product type to limit search space
Entries of same product type
2
BM25 search
Based on title textual similarity
top-K similar entries
3
Re-rank top-K using glance views and prioritize entries of same brand
4
Select few-shot examples for attributes A, B, C prediction for the input entry
attr A: X
attr B: X
attr C: []
Entry top1
attr A: []
attr B: X
attr C: X
Entry top2
attr A: X
attr B: []
attr C: X
Entry top3
attr A: X
attr B: X
attr C: []
Entry topk

Figure 2: Multi-Stage Retrieval Framework

As illustrated in Figure 1, CatalogRAG operates in two phases - offline index building and run-time attribute prediction, with the runtime workflow consisting of five key stages. First, the system builds store (language)-specific search indices (SI) from quality catalog entries (stage 0). At runtime, given a catalog entry with missing attributes, the system queries the relevant language's index to retrieve similar products (stage 1). The retrieved results undergo heuristic-based reranking to prioritize entries from the same brand and having higher glance views (stage 2). For each missing structured attribute, the system then selects relevant few-shot examples from these reranked entries (stage 3). These examples are incorporated into attribute-specific prompts (stage 4), which guide the LLM in generating predictions for the missing values (stage 5). By leveraging these patterns through strategically selected examples, CatalogRAG enables more accurate and consistent attribute prediction.

2.2 Multi-Stage Retrieval Framework

Retrieval Process: CatalogRAG employs a multi-stage retrieval framework to identify relevant pattern-aware examples for attribute prediction. The process combines term-based and text-based search strategies to efficiently navigate the catalog ecosystem while ensuring high-quality example selection as shown in Figure 2:

Given a catalog entry e with missing structured attributes $\mathcal{M} = \{A, B, C\}$ for prediction. The system first uses Store ID s to select the language-specific search index. The retrieval process begins with term-based search and uses Product Type p to constrain the search space $S_{p,s}$. This initial filtering step significantly reduces the computational overhead by focusing subsequent operations on a relevant subset of the catalog. The system then employs BM25 text-based search on product titles within $S_{p,s}$ to capture semantic and feature-level similarities beyond exact matches, returning a result of the ordered set R of top-k similar entries. This approach proves particularly effective in identifying products with related features, styles, or use cases.

Heuristic-based re-ranking: To enhance relevance, the system employs two heuristic features for re-ranking: glance views (GV) and brand alignment. The re-ranking process consists of two steps:

First, the top-k retrieved entries in R are sorted by glance views in descending order to obtain R'. Glance views typically refers to the number of times customers view a product's detail page, which indicates the popularity or visibility of a product in the catalog. Hence higher glance views of a catalog entry typically indicates better entry quality and more reliable attribute values.

Second, if brand b is available from the input entry e, entries in R' are reordered to obtain R'' as equation 1. This brand-aware reranking preserves the original ordering within each group while prioritizing entries from the same brand.

$$R'' = \{d_i \in R' \mid \text{brand}(d_i) = b\} \oplus \{d_j \in R' \mid \text{brand}(d_j) \neq b\} \quad (1)$$

Few-shot example construction: Finally, for each missing attribute $m \in \mathcal{M}$ in the given catalog entry e, the system selects up to n entries N from the re-ranked list R'' where the attribute m has filled value. Then we extract product title and its corresponding attribute-value pair as a few-shot example from each entry in N and incorporate those examples into attribute-specific prompts.

This multi-stage approach offers several key advantages. First, term-based filtering significantly reduces the search space by focusing on structurally and categorically relevant items, thereby improving both computational efficiency and matching accuracy. Second, the BM25 text-based search captures nuanced similarities in product features and descriptions. Third, the heuristics-based reranking, particularly through brand alignment and glance views, maintains consistency with brand-specific attribute patterns while prioritizing high-quality entries. Fourth, the flexible example selection mechanism adapts to the availability of relevant information, ensuring optimal use of available data. Furthermore, the use of language-specific indices allows CatalogRAG to adapt seamlessly to different linguistic and cultural contexts across various e-commerce stores.

BM25 retrieval advantages: We choose text-based BM25 as our retrieval framework for several reasons. First, given our large-scale product catalog, BM25's lexical matching effectively captures product similarities through surface-level textual patterns, which aligns well with our attribute prediction task. Consider these examples[2] from our system, where [*] denotes the attribute value:

```
Example 1
For attribute "shoes occasion"
Input catalog entry :
    <brand_1> <model_1> Women's Neutral Running Shoe-
    Fuschia/Yucca/Navy - 11
Retrieved top-3 similar catalog entries:
    <brand_1> Women's <model_1> Neutral Running Shoe-
    Oyster/Yucca/Pink - 5 Medium [Sports]

    <brand_1> Women's <model_1> Neutral Running Shoe-
    Black/Pink/Yucca - 8 Medium [Running]

   <brand_1> Women's <model_2> 5 Neutral Running Shoe-
    Ebony/Black/Yucca - 11 [Running]


Example 2
For attribute "shirt occasion"
Input catalog entry:
    <brand_2>, Illusion Neck Magnolia Blouse,
    2, Ivory
Retrieved top-3 similar catalog entries:
    <brand_2>, Passionflower Tie-Neck Chiffon
    Blouse, Lavender Multi, 14 [Vacation]

    <brand_2>, Passionflower Tie-Neck Chiffon
    Blouse, Lavender Multi, 0 [Vacation]

    <brand_2>, Chambray Wrap Blouse, 10,
    Indigo [Vacation]
```

As demonstrated in these examples, BM25 effectively retrieves products with strong symbolic similarities—matching brand names, model names and product types. These matches provide highly relevant context for attribute prediction, as products with similar titles often share similar attribute values.

Second, BM25 offers operational advantages. Unlike neural retrieval methods that require maintaining embedding models and managing vector similarity computations, BM25 provides a simpler, more maintainable production workflow. This simplicity translates to lower operational complexity and reduced infrastructure requirements while delivering the retrieval quality needed for our attribute prediction task.

3 Experimental setup

3.1 Dataset

Index corpus: For our experiments, we sampled from our production catalog about 30 million entries for US store, and about 20 million respectively for German (DE) and French (FR) stores, serving as the total search space for each store for similar product retrieval.

Test dataset: For evaluation, we constructed test sets of 3,000 catalog entries from each store by sampling across different product types, ensuring comprehensive coverage of the catalog's diversity. The test dataset consists of catalog entries with approximately half missing structured attributes having missing values that require prediction, making it suitable for evaluating our attribute completion task.

SA relevance tag: structured attributes vary in their significance within each product type. To reflect this hierarchy, we annotate each SA in the test datasets with a relevance tag which is either *VRel* ("very relevant") for high-priority attributes or *Rel* ("relevant") for standard attributes[3]. This classification, determined by business requirements, enables granular performance analysis across relevance groups and informs entry-level metric calculations (detailed in the metrics section).

3.2 LLM

We use `Mixtral-8x7B-Instruct`[4], a publicly available LLM, for both structured attribute (SA) prediction and evaluation tasks. For prediction, we employ attribute-specific prompts, where each SA is processed independently to maximize prediction accuracy. For evaluation, we utilize the same model with an optimized evaluation prompt that assesses the quality of predicted SA values and returns standardized evaluation labels.

3.3 Evaluation metrics

Evaluation metrics: With the labels from evaluator, we calculate the following evaluation metrics:

(1) Attribute-level **precision**, **recall** and **F1** are calculated for all SA predictions.

- **Precision (P)**: Measures the number of correctly generated attributes divided by the number of generated attributes.
- **Recall (R)**: Measures the number of generated attributes divided by the number of required attributes
- **F1**: The F1 score is calculated as $2PR/(P+R)$.

(2) Catalog entry-level **Completeness** and **Correctness**: we define the catalog entry-level *Completeness* and *Correctness* to assess the precision and recall of overall relevant SA quality respectively at catalog entry-level. The correctness (or completeness) of a given catalog entry is considered as high-quality when the precision (or recall) of all the "very relevant" SAs (SAs annotated with *VRel*) of a catalog entry meets the product type (PT)-specific thresholds determined by the business needs.

[2] We use <brand_x>, and <model_y> to anonymize brand and product model names

[3] The same attribute in catalog entries of different product types can be classified with different tags, because its significance vary in different product type categories.

[4] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1 (Apache 2.0 licensed)

3.4 Retrieval system and few-shot examples

We implement the term-based and OKAPI BM25 Probabilistic retrieval framework for similar product discovery using Lucene 8.11[5]. The system uses product title textual similarity between query catalog entry and entries in the index to rank the retrieved entries within the same product type (PT). For each query catalog entry, we first retrieve the top-100 similar entries from the respective entry index (1) based on title textual similarity (2) from the same PT search space, (3) we rerank the top-100 entries and prioritize the entries of the same brand as the query entry while maintaining the overall ranking (4) For each SA prediction, the system processes the reranked top-100 entries to select few-shot examples from the top rank.

Up to three entries with filled values for the target attribute are selected, where each example consists of the product title and its corresponding attribute-value pair. These examples are then integrated into the attribute-specific prompt. For this experiment, we ensure none of the chosen examples is same as the given entry.

3.5 Experimental configurations

We evaluate two distinct configurations for our experiments:
Base Prompt: Utilizes our iteratively improved prompt to predict missing values for given structured attributes. This prompt contains the entire product information of the given product catalog entry including both UA (e.g. title, description) and all the SAs, as well as general external knowledge such as the attribute and product definition, etc.

Base Prompt + CatalogRAG: Builds upon the previous configuration by adding few-shot examples and explicit instructions on using the few-shot examples. We added the following instruction in our prompt to inform the LLM of usage of the few-shot examples:

```
The examples below are the titles of similar products
along with the values of the attribute you are going
to predict. You can use these examples as reference to
help with your prediction.
```

4 Results and analysis

Store	Structured attr.			Catalog entry	
	Precision	**Recall**	**F1**	**Cor.**	**Compl.**
FR	+0.88%	+34.41%	+20.89%	+0.98%	+43.32%
DE	+0.74%	+4.00%	+2.29%	+2.83%	+2.36%
US	-0.16%	+5.08%	+2.99%	+0.37%	+2.94%

Table 1: Base prompt with CatalogRAG improvements (%) over base prompt alone: Attribute-level (precision, recall, F1) and catalog entry-level (completeness, correctness) metrics

CatalogRAG demonstrates consistent improvements over the base prompt alone across different metrics and stores as shown in Table 1. At the structured attribute level, we observe significant recall improvements across all stores (up to +34.41% for FR), while maintaining or slightly improving precision (up to +0.88% for FR, with a minor decrease of -0.16% for US). The FR shows particularly strong improvements, especially in recall metrics, indicating that LLM can benefit significantly from CatalogRAG for non-English predictions. Meanwhile, DE and US demonstrate consistent gains.

Catalog entry-level metrics of *completeness* and *correctness* also shown consistent improvements on our test datasets. The FR store shows the most substantial gains in *completeness* with a +43.32% improvement, indicating that CatalogRAG can enhance the overall attribute coverage of catalog entries. This is accompanied by a modest but positive increase in *correctness* (+0.98%), suggesting that the improved coverage does not come at the expense of precision. The DE store demonstrates balanced improvements in both metrics, with a +2.83% increase in *correctness* and a +2.36% gain in *completeness.* For the US store, while the improvements are more modest, we still observe positive gains with +0.37% in *correctness* and +2.94% in *completeness.* These results indicate that even in English-language catalogs, where the base LLM performance is typically stronger, CatalogRAG can still provide improvements in overall catalog quality.

	Group	ΔPrecision%	ΔRecall%	ΔF1%	SA%
US	VRel	-0.14%	+3.58%	+1.83%	59.80%
	Rel	-0.11%	+9.62%	+5.76%	40.20%
DE	VRel	+0.57%	+2.39%	+1.56%	60.36%
	Rel	+1.16%	+5.54%	+4.05%	39.64%
FR	VRel	+0.91%	+28.27%	+15.86%	61.18%
	Rel	+0.79%	+49.35%	+32.67%	38.82%

Table 2: Precision, Recall, and F1 score improvements when using CatalogRAG compared to base prompt across marketplaces, grouped by attribute relevance (VRel: Very Relevant, Rel: Relevant)

The effectiveness of CatalogRAG varies across attribute relevance groups, as detailed in Table 2. *Very Relevant (VRel)* attributes constitute the majority (~60%) of structured attributes across stores, while *Relevant (Rel)* attributes make up the remaining ~40%. As Table 2 shows, column SA% represents the percentage of structured attributes for each relevance group within the test dataset for each store. Both groups show substantial improvements, with *Rel* attributes demonstrating stronger gains overall. For *VRel* attributes, FR shows the highest recall improvement (+28.27%), also with substantial gains in DE and US, accompanied by slight precision improvements in European stores and a minimal decrease in US. *Rel* attributes show even more substantial improvements, with recall gains up to +49.35% (FR) and consistent precision improvements across most stores. The stronger performance in *Rel* attributes might indicate that these attributes benefit more from the contextual information provided by similar product examples, possibly because they are more standardized or follow more consistent patterns within product categories.

CatalogRAG's few-shot example coverage at the Product Type Attribute level varies across test datasets: 33% (DE), 43% (US) and

[5] https://lucene.apache.org/core/ (Apache 2.0 licensed)

37% (FR) of the empty structured attributes (SAs) in the test dataset have up to three relevant few-shot examples available for prompt regeneration. Notably, these SAs with few-shot examples represent products from more than 92% of the catalog entries in the test dataset across the stores.

5 Related work

Attribute value prediction in e-commerce has traditionally been approached as an information extraction problem - extracting attribute values from existing text, evolving from rule-based methods to neural approaches.

Early extraction-based methods rely on rule-based systems that use handcrafted patterns and domain-specific heuristics to find attribute values in product text [3]. When an attribute value is missing, these methods attempt to locate it in other parts of product information such as titles and descriptions. However, this approach struggles with scalability and adaptability across constantly growing multilingual product categories in modern e-commerce. As neural extraction methods emerge, they offer more flexibility and better handling of natural language variations: Previous study [21] treats the challenge as a Named Entity Recognition (NER) task, enabling more flexible identification of attribute values within product descriptions. Parallel developments see the emergence of sequence tagging models [22], which improves the ability to capture contextual relationships in product descriptions and specifications. However, these methods remain fundamentally limited by their extraction nature - they can typically only identify values explicitly mentioned in the product text.

A significant paradigm shift occurs with the introduction of Google's MAVEQA system [23], which reformulates the extraction task as a question-answering problem, where each attribute becomes a question to be answered from the product's textual information. This novel approach allows for more natural interaction with product data and improves the handling of complex attribute relationships. The recent SAGE model [14] introduces a generative approach to the task, enabling the inference of implicit values and demonstrating capability in zero-shot predictions for previously unseen product-attribute combinations. With the advent of Large Language Models (LLMs), the scope of attribute prediction expands beyond pure extraction. LLMs can be prompted with more product information and attribute definitions to generate predictions, potentially inferring values even when not explicitly stated. However, not all missing attributes can be predicted or inferred solely from product information and static metadata, even with the model's latent knowledge. This has led to the exploration of Retrieval-Augmented Generation (RAG) to incorporate external knowledge sources due to the inherent limitations of LLMs in accessing specialized knowledge [4, 8, 15, 20]. While RAG has shown success in general question-answering tasks [5, 7, 19] and various studies [1, 6, 10, 12, 18], external knowledge sources are suboptimal for attribute prediction in e-commerce for two key reasons: (1) external sources cannot effectively capture the specific catalog conventions and norms that exist within product types, stores, and seller practices, and (2) they struggle to keep pace with the constant emergence of new products and attributes in worldwide e-commerce.

CatalogRAG takes a novel approach by leveraging the catalog ecosystem itself as the source of relevant information. Instead of relying on external knowledge bases, we retrieve similar products from within the catalog and transform this information into few-shot examples to guide LLM predictions. This approach naturally captures product type-specific patterns, store-specific conventions, and brand relationships. To our knowledge, this is the first work to propose a retrieval-augmented approach that utilizes internal catalog patterns for multilingual attribute value prediction in e-commerce.

6 Conclusion

In this paper, we introduce CatalogRAG, a novel pattern-aware retrieval-augmented generation system for predicting missing structured attributes in e-commerce catalogs. Our approach addresses the challenge of accurately predicting missing attribute values, a problem affecting nearly half of the relevant attributes across product types. By strategically leveraging similar products within the catalog ecosystem, CatalogRAG provides LLMs with contextually relevant examples for attribute prediction. Evaluation on experimentation catalog samples across three major language stores (US, DE, FR) demonstrated significant improvements in catalog data quality. CatalogRAG achieved substantial gains in attribute-level metrics, with recall improvements of up to 34.41% and precision improvements of up to 0.88%. At the catalog entry level, we observe increases up to 43.32% in completeness and up to 2.83% in correctness.

For future research directions, we propose integrating successfully enhanced, high-quality catalog entries back into the search indices. This direction could create a self-improving ecosystem where the system continuously learns from its own successes, potentially leading to compounding improvements over time. Additionally, investigating adaptive example selection techniques that dynamically adjust based on product category complexity could further optimize both performance and computational efficiency.

References

[1] Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context Examples Selection for Machine Translation. arXiv:2212.02437 [cs.CL] https://arxiv.org/abs/2212.02437

[2] Tianchi Bi, Liang Yao, Baosong Yang, Haibo Zhang, Weihua Luo, and Boxing Chen. 2020. Constraint Translation Candidates: A Bridge between Neural Query Translation and Cross-lingual Information Retrieval. arXiv:2010.13658 [cs.CL]

[3] Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. Domain Adaptation of Rule-Based Annotators for Named-Entity Recognition Tasks. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Cambridge, MA, 1002–1012. https://www.aclweb.org/anthology/D10-1098

[4] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 [cs.CL] https://arxiv.org/abs/2312.10997

[5] Gauthier Guinet, Behrooz Omidvar-Tehrani, Anoop Deoras, and Laurent Callot. 2024. Automated Evaluation of Retrieval-Augmented Language Models with Task-Specific Exam Generation. arXiv:2405.13622 [cs.CL] https://arxiv.org/abs/2405.13622

[6] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. arXiv:2002.08909 [cs.CL] https://arxiv.org/abs/2002.08909

[7] Jennifer Hsia, Afreen Shaikh, Zhiruo Wang, and Graham Neubig. 2024. RAGGED: Towards Informed Design of Retrieval Augmented Generation Systems. arXiv:2403.09040 [cs.CL] https://arxiv.org/abs/2403.09040

[8] Wenqi Jiang, Shuai Zhang, Boran Han, Jie Wang, Bernie Wang, and Tim Kraska. 2024. PipeRAG: Fast Retrieval-Augmented Generation via Algorithm-System Co-design. arXiv:2403.05676 [cs.CL] https://arxiv.org/abs/2403.05676

[9] Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual Information Retrieval with BERT. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*. European Language Resources Association, Marseille, France, 26–31. https://aclanthology.org/2020.clssts-1.5

[10] Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. arXiv:2305.06983 [cs.CL] https://arxiv.org/abs/2305.06983

[11] Mike Lowndes and Aditya Vasudevan. 2021. Market Guide for Digital Commerce Search. (2021).

[12] Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023. SAIL: Search-Augmented Instruction Learning. arXiv:2305.15225 [cs.CL] https://arxiv.org/abs/2305.15225

[13] Jian-Yun Nie. 2010. Cross-language information retrieval. *Synthesis Lectures on Human Language Technologies* 3, 1 (2010), 1–125.

[14] Athanasios N Nikolakopoulos, Swati Kaul, Siva Karthik Gade, Bella Dubrov, Umit Batur, and Suleiman Ali Khan. 2023. SAGE: Structured Attribute Value Generation for Billion-Scale Product Catalogs. *arXiv preprint arXiv:2309.05920* (2023).

[15] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-Context Retrieval-Augmented Language Models. arXiv:2302.00083 [cs.CL] https://arxiv.org/abs/2302.00083

[16] Andreas Rücklé, Krishnkant Swarnkar, and Iryna Gurevych. 2019. Improved Cross-Lingual Question Retrieval for Community Question Answering. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 3179–3186. doi:10.1145/3308558.3313502

[17] Shadi Saleh and Pavel Pecina. 2020. Document Translation vs. Query Translation for Cross-Lingual Information Retrieval in the Medical Domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 6849–6860. doi:10.18653/v1/2020.acl-main.613

[18] 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 [cs.CL] https://arxiv.org/abs/2301.12652

[19] Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. *Transactions of the Association for Computational Linguistics* 11 (01 2023), 1–17. doi:10.1162/tacl_a_00530 arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00530/2067834/tacl_a_00530.pdf

[20] Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. DRAGIN: Dynamic Retrieval Augmented Generation based on the Information Needs of Large Language Models. arXiv:2403.10081 [cs.CL] https://arxiv.org/abs/2403.10081

[21] Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A Unified Generative Framework for Various NER Subtasks. doi:10.48550/ARXIV.2106.01223

[22] Jun Yan, Nasser Zalmout, Yan Liang, Christan Grant, Xiang Ren, and Xin Luna Dong. 2021. AdaTag: Multi-attribute value extraction from product profiles with adaptive decoding. In *ACL-IJCNLP 2021*. https://www.amazon.science/publications/adatag-multi-attribute-value-extraction-from-product-profiles-with-adaptive-decoding

[23] Li Yang, Qifan Wang, Zac Yu, Anand Kulkarni, Sumit Sanghai, Bin Shu, Jon Elsas, and Bhargav Kanagal. 2022. MAVE: A Product Dataset for Multi-Source Attribute Value Extraction. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 1256–1265. doi:10.1145/3488560.3498377