Bridging the Gap: Generative Retrieval via Query-to-Multi-Span Framework for Effective E-commerce Search

Anonymous ACL submission

Abstract

Generative retrieval introduces a groundbreaking paradigm to document retrieval by directly generating the identifier of a pertinent document in response to a specific query. This paradigm has demonstrated considerable benefits and potential, particularly in representation and generalization capabilities, within the context of large language models. However, it faces significant challenges in E-commerce search scenarios, including the complexity of generating detailed item titles from brief queries, the presence of noise in item titles with weak language order, issues with long-tail queries, and the interpretability of results. To address these challenges, we have developed an innovative framework for E-commerce search, called generative retrieval via query-to-multi-span. This framework is designed to effectively learn and align an autoregressive model with target data, subsequently generating the final item through constraint-based beam search. By employing multi-span identifiers to represent raw item titles and transforming the task of generating titles from queries into the task of generating multi-span identifiers from queries, we aim to simplify the generation process. The framework further aligns with human preferences using click data and employs a constrained search method to identify key spans for retrieving the final item, thereby enhancing result interpretability. Our extensive experiments show that this framework achieves competitive performance on a real-world dataset, and online A/B tests demonstrate the superiority and effectiveness in improving conversion gains.

1 Introduction

Deep semantic retrieval models (Zhang et al., 2020; Devlin et al., 2018; Qiu et al., 2022; Khattab and Zaharia, 2020; Zhang et al., 2021; Li et al., 2023b; Wang et al., 2023; Li et al., 2023a), have achieved significant success in online E-commerce retrieval and recommendation systems. Traditional methods rely on the dual-encoder to learn the dense representations of queries and items. They use the dot-product similarity to measure the relevance between the query and candidate items, but lack fine-grained interactions, leading to sub-optimal performance.

Recently, a new paradigm, generative retrieval (Wang et al., 2022; Tay et al., 2022; Tang et al., 2023; Yuan et al., 2024; Bevilacqua et al., 2022; Rajput et al., 2024a; Zhou et al., 2023), has been proposed in the recommendation field and question-answer fields. These models advocate generating identifiers of target passages/items directly through the autoregressive language models. Existing work could be divided into two categories based on identifier types: 1) Numeric-based (Wang et al., 2022; Zhuang et al., 2022; Rajput et al., 2024a; Yuan et al., 2024), they assign numeric identifiers in various ways, e.g., atomic, naive, and semantic. 2) lexical identifier-based methods (Bevilacqua et al., 2022; Lee et al., 2023; Li et al., 2023d) using the n-grams, title, and URLs as the document identifiers. They could leverage the knowledge of PLMs to decode identifiers, exploring the benefit of pre-trained vocabulary space. The lexical identifier-based methods show potential in terms of interpretability and generalization capabilities, especially in the era of large language models. Thus, we continue to explore along these lines of methods in this paper.

In the field of E-commerce, there exist several crucial challenges. Firstly, the task of query-to-title(query2title) generation poses difficulties. Specifically, product titles tend to be lengthy on average, whereas user-entered query words are typically short. Attempting to directly generate lengthy titles can result in significant hallucination issues. While some efforts have been made to utilize pre-trained semantic IDs as document identifiers (Wang et al., 2022; Tay et al., 2022; Yuan et al., 2024) to simplify the task into query-to-semanticID and re-

duce complexity, this approach heavily relies on external document representations, deviating significantly from the language itself and necessitating additional calibration, thereby diminishing result interpretability. Additionally, some research efforts focus on leveraging features inherent to the text itself (Zhang et al., 2024). However, these approaches encounter limitations in e-commerce contexts, including excessively fine-grained term granularity and an over-reliance on the model's fitting ability for term segmentation, which impedes their full effectiveness.

Secondly, noise in item titles and weak language order (i.e., keyword stacking) are prevalent issues. In actual product websites, merchants usually provide item titles that contain noise and redundant information. Moreover, the semantic order is predominantly local rather than globally coherent. Essential information such as brand words, attribute words, and categories is often present in the text without regard for position.

Thirdly, long-tail query challenges are apparent. Unlike in traditional question-and-answer domains, E-commerce faces a severe sample imbalance between queries and items. While some long-tail queries have limited associated products, head queries are linked to a vast array of items. In the age of deep semantics, one-to-many mapping issues can be mitigated through spatial clustering. However, in generative paradigms, such relationships manifest diversely, posing ongoing challenges in resolving them effectively.

Lastly, the interpretability of results is a critical concern. The ability to interpret search results provides valuable insight for enhancing user experience. Unfortunately, deep semantic methods often fall short in this aspect.

To alleviate the above problems, we introduce a novel framework for E-commerce search, called generative retrieval via query-to-multi-span (GenR-Q2MS). This framework comprises four key stages: 1) Task re-definition stage; 2) Supervised fine-tuning stage; 3) Preference optimization stage; and 4) Inference stage based on constraint beam-search. The architecture's cornerstone lies in its initial task redefinition, where we reconstruct item titles through linguistic reorganization while preserving core information. By segmenting titles into semantically coherent spans and reformulating the task as query-to-multi-span generation, we achieve three critical improvements: (1) reducing generation complexity by shortening output sequences, (2) mitigating information redundancy through localized span modeling that aligns with e-commerce titles' inherent structure, and (3) amplifying training data utility via multi-span sampling from single query-title pairs. The supervised fine-tuning stage aims to learn the knowledge of the E-commerce field and reduce illusions. The stage of preference optimization is to align with human preference data to produce more significant and human-standard-compliant results. The constraint beam search could prevent the generation of invalid identifiers (i.e., span not occurring in any items). We conducted comprehensive experiments using an industrial dataset collected from user interactions on an e-commerce platform, and the results of offline and online demonstrate the effectiveness of the proposed.

The contributions of this paper can be summarized as follows:

- We formally characterize the structural mismatch between conventional query-to-title generation and e-commerce semantics (title redundancy, local ordering). To bridge this gap, we innovatively propose a task redefinition paradigm centered on query-to-multi-span generation, which is composed of segmentation, reconstruction and aggregation.
- We propose a novel framework, generative retrieval via query-to-multi-span (GenR-Q2MS), that provides a complete pipeline for training, aligning, and inference, meanwhile enhancing result interpretability.
- We conduct extensive experiments on a real-world dataset. Offline and online experimental results demonstrate that GenR-Q2MS achieves significant improvements, especially on long-tail queries, compared to generative retrieval and dense retrieval baselines.

2 Method

While numerous efforts (Wang et al., 2022; Tay et al., 2022; Yuan et al., 2024; Zhang et al., 2024) have been made to enhance generative retrieval in e-commerce, significant challenges remain. First, generating product titles from queries is complicated by length disparities, leading to hallucinations and reduced interpretability due to reliance on semantic IDs. Second, product titles often suffer from noise and lack coherence, with key elements poorly positioned. Third, long-tail queries create

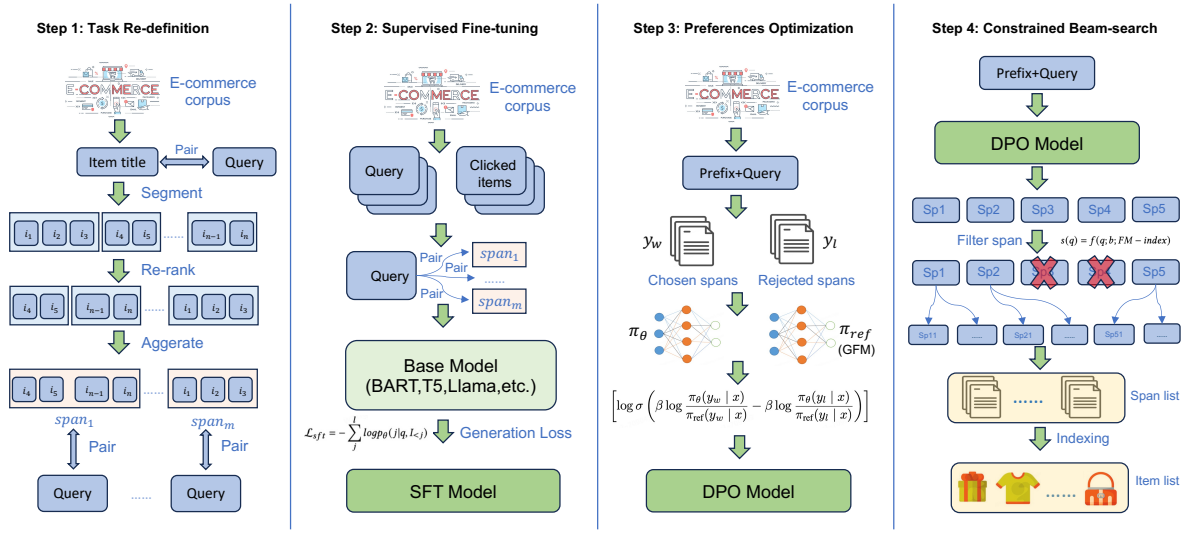

Figure 1: The framework of GenR-Q2MS. It comprises four key stages: 1) Task re-definition stage; 2) Supervised fine-tuning stage; 3) Preference optimization stage; and 4) Inference stage based on constrained beam-search.

imbalance issues, as generative models struggle to capture diverse relational dynamics. To address these challenges, we propose a novel framework for e-commerce search, called generative retrieval via query-to-multi-span (GenR-Q2MS), as illustrated in Figure 1. This framework includes four key stages: 1) Task re-definition; 2) Supervised fine-tuning; 3) Preference optimization; and 4) Inference using constraint beam-search. The specific details of each stage will be discussed in the following sections.

2.1 Task Re-definition

To address the aforementioned issue, we propose a method that utilizes multi-segment identifiers to reconstruct product titles, transforming the complex task of matching queries with long titles into an association with multiple relevant text segments. Specifically, this process is composed of four steps: segmentation, re-ranking, aggregation, and matching.

Initially, the segmentation step divides the title into multiple semantically meaningful phrases or terms. Subsequently, the reordering step rearranges these segmented terms according to a predefined sorting rule. Following this, the aggregation step combines the reordered terms into text segments (spans) of similar length. Finally, in the matching step, the generated segments are scored to evaluate their degree of correspondence with the title, thereby determining which products to recall.

Based on the task re-definition, we can achieve three significant advantages. Firstly, transforming the task from generating long texts to match titles into generating short spans to match titles significantly reduces task complexity. This reduction in the number of tokens generated not only shortens inference time but also effectively decreases the likelihood of hallucinations. Secondly, by segmenting e-commerce titles into multiple s, we effectively address the issue of information redundancy, aligning more closely with the characteristic of e-commerce titles being locally ordered yet globally unrelated. Thirdly, in terms of sample construction, whereas a single query-title pair originally could only form one sample, after segmentation, it can generate multiple samples, thereby greatly enriching the dataset's sample volume. This strategy, when applied to large model applications in search and recommendation systems, can significantly enhance both the efficiency and effectiveness of the task.

The details of the 4 steps are presented in the following sections.

2.1.1 Segmentation

Assuming that there is a training sample pair $< query, item >$, and the item's title consists of n tokens, i.e., $[i_1, i_2, \cdots, i_n]$. We first adopt a self-developed tokenization tool similar to Jieba, but it considers terms related to title, brand, and category, processing them at the granularity of n-grams $\{[i_1, i_2, i_3], [i_4, i_5], \cdots, [i_{n-1}, i_n]\}$.

2.1.2 Re-ranking

Considering the position insensitivity in n-grams, we can re-rank terms using various methods such as

Figure 2: The processing of titles.

term frequency, part of speech, lexicographical order, and neural network scoring. For simplicity and ease of implementation, we adopt lexicographical order in this paper. The Rerank process is denoted as $\{[i_4, i_5], [i_{n-1}, i_n], \cdots, [i_1, i_2, i_3]\}$.

2.1.3 Aggregation

After that, then we aggregate them into spans with a length threshold L. If the length exceeds L, the last n-gram is shifted to the next span. This method suits the stacked information in e-commerce titles, where order within spans is significant.

$$\left\{ \underbrace{[i_4, i_5, i_{n-1}, i_n]}_{span_1}, \cdots, \underbrace{[i_1, i_2, i_3]}_{span_m}, \right\} \quad (1)$$

where each span has a corresponding length l.

To illustrate the process more specifically, we use an example from e-commerce product titles, as shown in Figure 2:

Given the title: "Safeguard Antibacterial Hand Wash 100ml Old Packaging," we begin by sorting the n-grams in lexicographical order, in accordance with Equation 1. Next, we aggregate multiple consecutive n-grams into spans. Each span has a predefined length threshold (L). If the aggregated length exceeds (L), the last n-gram is moved to the next span.

It is important to note that this strategy is tailored to our specific business context. For other domains, different methods that better suit their unique scenarios can be employed.

2.1.4 Matching

Following the above reconstruction, we have not only preserved the original information but also augmented the shared information between products at the span level, significantly reducing the noise and diversity introduced by the word order.

Each new span contains a portion of the effective information, representing a perspective. This is because we have simplified the original query2title task into a parallel query2multi-span task, meaning one training sample has become m samples, as follows:

$$\begin{array}{c} < query, item > \\ \downarrow \\ < query, span_1 >, \cdots, < query, span_m > \end{array}$$

2.2 Supervised Fine-tuning

Due to the general pre-trained model lacking e-commerce domain knowledge, we perform supervised fine-tuning (SFT) on specific data via the click pairs between the query and item. Specifically, for each training sample, the objective is to minimize the sum of the negative log-likelihoods of the tokens $\{i_1, \cdot, i_j, \cdot, i_l\}$ in a target identifier I (span), whose length is l. The generation loss is formulated as,

$$\mathcal{L}_{sft} = -\sum_{span}^{m} \sum_{j}^{l} \log p_\theta(j|q, I_{<j}) \quad (2)$$

where $I_{<j} = \{i_1, i_2, \cdots, i_j\}$, p_θ is the SFT model.

2.3 Preferences Optimization

Although the supervised fine-tuning model has achieved tremendous success, the outcomes it generates remain uncontrollable, unstable, and do not align with human preference requirements. To alleviate this problem, existing works attempt to align preferences with reinforcement learning from human feedback (RLHF). However, this pipeline may be too complex and often unstable. Fortunately, recent work DPO (Rafailov et al., 2024) derives a simple approach for policy optimization using preferences directly. Given a query, the preference data $\mathcal{D} = \{(x, y_w, y_l)\}$ contains the query x, chosen span y_w, and rejected span y_l, and the objective of DPO is denoted as:

$$\mathcal{L}_{DPO} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x))} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)} \right) \right] \quad (3)$$

where β is a parameter controlling the deviation from the base reference policy π_{ref}.

It's crucial to highlight that the construction of preference data is closely tied to business metrics. By employing a learning-to-rank approach, preference pairs such as <exposed but not clicked,

clicked> and <random negative, clicked> are created, which enhances the visibility of products that are more likely to convert.

2.4 Constrained Beam-search

During the inference process, given a query text, the trained autoregressive language model SFT/DPO model could generate predicted identifiers in an autoregressive manner with constrained beam search, which adopts the FM-index (Ferragina and Manzini, 2000) to identify the set of possible next tokens, avoiding invalid identifiers without in all item title.

More precisely, after a single decoding pass, we get a set of n-grams along with their autoregressively computed probabilities according to the model LM and then retrieve their FM-index scores via normalized index frequencies. The constrained beam-search score is the sum of the model score and FM-index score, formulated as

$$s(q) = f(q; b; \text{FM-index}) \tag{4}$$

where b is the beam size for beam search. Subsequently, we obtained a refined probability distribution, and by employing various ranking strategies such as top@p and top@k, we generated a set of n-grams, which are the next-step inputs. This process continued until the generation phase was completed. During this computation process, if an out-of-vocabulary (OOV) n-gram is encountered, its feature mapping (FM) score will be assigned negative infinity. As a result, it will be filtered out of the selection process. This approach falls under the retrieval-augmented paradigm(RAG), which effectively reduces the rate of hallucination, thereby enhancing the efficacy and accuracy of the inference process. More details could refer to the original paper SEAL (Bevilacqua et al., 2022).

Leveraging the above constrained beam-search, we efficiently harvest a batch of potent spans. Subsequently, we employ the FM-index for the swift identification of items that closely correspond to these segments. Importantly, the FM-index operates independently of span positioning, thus ensuring comprehensive retrieval of all relevant items, a feature that is in harmony with the objectives set forth by the task redefinition module.

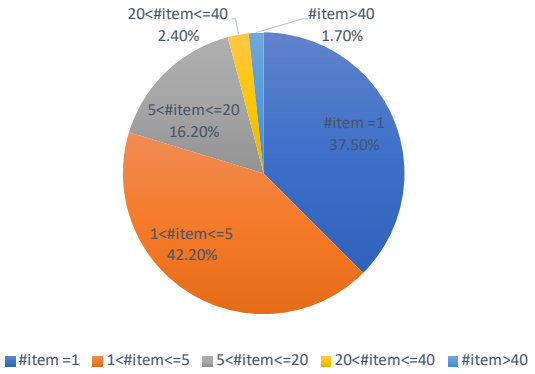

Figure 3: The distribution of percentages across different queries.

3 Experiments

3.1 Datasets and Metrics

We collect search logs of user clicks and purchases from an online E-commerce website, where the size of the dataset is 2.8 billion. Within the practical confines of our business operations, we prioritize set-based metrics due to their relevance to our objectives. Traditional ranking metrics such as NDCG and AUC, while standard for evaluating ranking quality, are not the main focus during the retrieval phase. Therefore, We choose the standard retrieval quality metric Recall@K to measure the results based on the full corpus, where $K \in \{500, 1000\}$ respectively. To examine the model's performance on long-tail queries with fine granularity, we divided the original queries into five groups based on the word-level click count. As shown in Figure 3, queries with less than 5 clicks per account for 80%, indicating a significant long-tail effect.

3.2 Baselines

In the industrial field, there are two foundational paradigms, dense retrieval and generative retrieval. Therefore, we conduct separate experimental comparisons for each paradigm.

- *Dense retrieval.* This paradigm is the most widely used work and makes a great success. The representative work is DSSM (Huang et al., 2013) and the variant version with a pre-trained model based on Bert (Devlin et al., 2018). Without loss of generality, we select **RSR** (Qiu et al., 2022) as the representative of the backbone of Bert, which had been deployed in the online system, severing hundreds of millions of users.

- *Generative retrieval*. This paradigm is an emerging and promising work. Based on different identifiers, it can be divided into two main categories, numerical-based method and lexical-based method. The state-of-the-art work numerical-based is **TIGER** (Rajput et al., 2024b), which utilizes semantic codes generated by residual quantization (RQ) as identifiers. In this paper, we first use the two-tower (RSR) product of the item's embedding and then construct the semantic ID of a given item by RQ. The most relevant lexical-based model is **SEAL** (Bevilacqua et al., 2022) uses arbitrary n-grams in documents as identifiers, and retrieves documents under the constraint of a pre-built FM-indexer. What's more, GenR-Q2MS is easily extensible and could be adapted in various aligning via LTR learning (Zhou et al., 2023; Qiu et al., 2022; Li et al., 2023c).

3.3 Implementation Details

To ensure a fair comparison among different methods, we keep the vocabulary size, the dimension of query/item, and parameters of PQ the same as (Li et al., 2023a). Specifically, we set the dimension as 128, batch size as 350, n-list of IVF-PQ as 32768, nprobe as 1, and the indexing construction is used in the Faiss ANNS library[1]. The default temperature τ of softmax is 1/30. The Adam optimizer is employed with an initial learning rate of 5e-5, and 6e-5 for RSR/SFT and DPO respectively. The default value of beam-search size is set to 100. The base model of SEAL, TIGER, and ours are all BART-large[2] (Lewis et al., 2019). For the TIGER model, the parameter is set to RQ3x12, which consists of three residual layers, each encoding 4096 codebooks, enabling the representation of a product scale in the tens of billions.

3.4 Experiment Results

The experimental results are shown in Table 1. We can conclude that the proposed framework achieves a significant improvement over dense retrieval and generative retrieval. Specifically,

- Compared with GenR-Q2MS + SFT, SEAL + SFT leads to a performance decline in various metrics, showing that the straightforward query2title task is ineffective. This result is consistent with previous analysis. Compared with TIGER, the lexical-based GenR-Q2MS* makes a great improvement, indicating that it describes a better semantic match. However, the numerical pattern has a semantic gap that requires additional alignment.

- Compared with RSR, GenR-Q2MS, and variable versions perform better in terms of different long-tail queries, especially in the #item=1. This phenomenon indicates that generative paradigms have increased generalization capabilities compared to traditional paradigms. Additionally, it is observed that the performance of the generative method varies significantly across different types of queries. For example, under head queries, the suboptimal performance of the generative method is more pronounced, which may be associated with one-to-many map learning.

- The main goal of our experiments was to improve recall for mid- and long-tail queries. From Table 1, we can see that after applying SFT (GenR-Q2MS + SFT vs RSR), the model's recall for mid- and long-tail queries improved significantly compared to KNN. However, the performance for head queries was relatively poor, affecting overall usability. To address this issue, we employed DPO (Differentiable Prompt Optimization) for pairwise learning and adjusted the dataset accordingly (hard-to-easy negatives ratio: 1:3). The results show that after using DPO (GenR-Q2MS + SFT + DPO (w/ cons) vs RSR), recall for mid- and long-tail queries saw a substantial increase, while the performance difference for head queries compared to KNN became smaller (especially in recall@1000, where the difference is quite small). When comparing DPO to SFT (GenR-Q2MS + SFT + DPO (w/ cons) vs GenR-Q2MS + SFT), although the metrics for mid- and long-tail queries slightly declined, the performance for head queries improved. In summary, DPO has brought overall benefits.

3.5 Impact of Different Tasks

To investigate the effect of different tasks on performance, we conduct several tasks, i.e., query2title, title2query, and query2multi-span. The results are shown in Table 2. We can find that the performance of the query2title and title2query tasks are

[1]https://github.com/facebookresearch/faiss
[2]https://huggingface.co/facebook/bart-large

Table 1: Performance of different methods in various types of query. Based on the number of items under each query (#item), we divide the queries into five categories. The fewer the number, the more long-tailed the description. w/o cons denotes the without constrainted beam-search. The bolded values indicate the optimal values; the underlined values denote the suboptimal values.

Method	$\#item = 1$	$1 < \#item <= 5$	$5 < \#item <= 20$	$20 < \#item <= 40$	$\#item > 40$
Recall@500					
RSR	0.2900	0.2922	0.3083	**0.3025**	**0.2117**
SEAL + SFT	0.0180	0.0120	0.0133	0.0102	0.0039
TIGER	0.1470	0.1484	0.1561	0.1801	0.1377
GenR-Q2MS + SFT	**0.3760**	**0.3762**	0.3266	0.2850	0.1635
+ DPO(w/o cons)	0.3240	0.3344	0.3016	0.2662	0.1544
+ DPO (w/ cons)	0.3680	0.3672	**0.3289**	0.2918	0.1690
Recall@1000					
RSR	0.3100	0.3086	0.3306	0.3315	**0.2451**
SEAL + SFT	0.0240	0.0198	0.0179	0.0139	0.0061
TIGER	0.1920	0.1930	0.1998	0.2307	0.1906
GenR-Q2MS + SFT	0.4230	**0.4304**	0.3890	0.3515	0.2169
+ DPO (w/o cons)	0.3700	0.3803	0.3609	0.3273	0.2074
+ DPO (w/ cons)	**0.4310**	0.4273	**0.4001**	**0.3674**	0.2330

Table 2: The effect of different tasks for the performance.

Task	Recall@500	Recall@1000
query2title	0.0180	0.0240
title2query	0.0160	0.0232
query2multi-span (l=10, m=2)	0.3600	0.4070
query2multi-span (l=8, m=7)	**0.3680**	**0.4310**

extremely poor, while the query2multi-span task has significantly improved. This suggests that there is noise in the original data in the e-commerce field, which once again underscores the importance of task re-definition.

3.6 Impact of Beam Size

The beam size controls the quality and quantity of the generated results, impacting the model's performance. Here, we conduct additional experiments to explore the influence (parameters are l=10, m=2). The experimental results are shown in figure 4. As the size increases, the effect improves, but the gradient of improvement decreases. It is also found that the larger the beam size, the greater the irrelevance of the returned results. Specifically, as the beamsize increases, the number of generated spans also increases, resulting in a higher number of items after constrained generation. However, expanding the beamsize may introduce irrelevant spans, which in turn can lead to irrelevant product titles through fm-index. For examples:

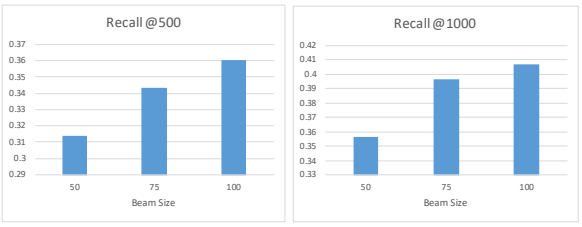

Figure 4: The performance of different beam sizes.

- Query: Phone, Span: Red, Case: Red phone case
- Three-fold product, Span: Three-fold, Case: Three-fold umbrella
- Query: 100ml hand sanitizer, Span: 100ml, Case: 100ml liquor

Therefore, in practical applications, a certain compromise must be made.

3.7 Online A/B Test

To evaluate the effectiveness of our model, we begin by introducing the funnel-shaped architecture of the online system. In e-commerce search engines, the system is typically divided into several stages: as shown in Figure 5: retrieval, truncation, pre-ranking, ranking, and mix-ranking. It is important to note that the online experiment employs a multi-channel recall mechanism, which includes inverted indexing, KNN, i2i, and our method. There

Table 3: Online performance of A/B tests. The improvements are averaged over a week in 2024. p-value is obtained by t-test over the online dense retrieval model.

Metric	UCVR	UV-value	recall exposure rate	pre-rank exposure rate	relevance score
Gain	+0.225%	+0.050%	+1.80%	+0.30%	+0.12%
p-value[b]	0.0276	0.8780	-	-	-

[b] Small p-value means statistically significant.

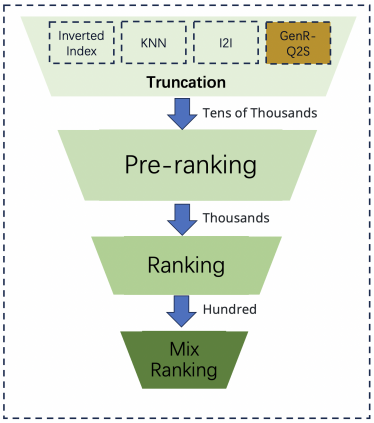

Figure 5: The funnel-shaped architecture of the online system.

may be overlap among the results from these different recall paths. Therefore, to achieve performance improvement through the addition of a new recall source, it is essential to ensure that it provides a significant incremental contribution. In the e-commerce context, the recall stage primarily focuses on three key metrics: exposure, relevance, and efficiency.

The exposure metric primarily measures the visibility rate of candidates in the pre-ranking and coarse ranking stages. We assess these metrics by calculating the percentage of candidates that progress through each stage relative to the total number of candidates. Relevance metrics are evaluated using a teacher model within the relevance module, specifically employing a cross-encoder large model to assess the relevance of search results. Efficiency metrics are gauged through the user conversion rate (UCVR) which is consistent with previous work (Cheng et al., 2024; Wang et al., 2024; Li et al., 2023b), reflecting the actual conversion impact on users utilizing the search engine.

After a comprehensive one-week monitoring period, the results are presented in Table 3. The data indicate that the proposed model achieved a 0.225% increase in UCVR (p-value=0.0276), suggesting that the model can provide higher-quality candidates, thereby enhancing conversion rates. Additionally, the proportion of candidates advancing through the pre-ranking stage increased by 1.8%, and through the coarse ranking stage by 0.3%, indicating that the system can offer more effective product candidates in these stages. The relevance metric also improved by 0.12%, demonstrating significant performance enhancements in the relevance of generated products, indicating higher quality in the recommended products.

These results collectively validate the effectiveness of our approach, illustrating that by optimizing system architecture and model design, the overall performance of search engines in e-commerce scenarios can be significantly enhanced.

4 Conclusion

This paper introduces an innovative generative retrieval framework via query-to-multi-span tailored for E-commerce search. The framework is crafted to adeptly train an autoregressive model in line with the target data and leverage constrained beam-search to produce the ultimate item selection. To cater to the E-commerce domain, we reconstruct the raw item titles and employ multi-span as identifiers, thereby converting the query2title task into a query2multi-span task, which simplifies the generation process. During inference, a constrained beam-search approach is utilized to pinpoint crucial spans, meaning as well as the interpretability of the retrieved items. Comprehensive testing on a real-world dataset shows that our framework markedly outperforms contemporary generative retrieval and dense retrieval in long-tail queries. The A/B test demonstrates that the model has brought about substantial conversion gains.

In future work, we aim to harness the power of large language models to bolster representation and generation capabilities for the base model and formulate an improved learning-to-rank scheme to amplify the pertinence of the generated outcomes.

5 Limitation

- **Real-time Online Service**: The current natural language processing systems in e-commerce scenarios require several hundred milliseconds for inference, which is insufficient for delivering a seamless user experience. To achieve genuine real-time online services, we must explore additional techniques and strategies to accelerate the inference process, such as utilizing more efficient algorithms, optimizing model structures, and leveraging hardware acceleration.

- **More Precise Multi-Spans**: In e-commerce scenarios, the generated spans often contain significant semantic noise due to the presence of adjectives and determiner words. These words are typically short and appear alone or in combination with other words to form short phrases, resulting in spans that are semantically unrelated or redundant. We are currently attempting to alleviate this issue by pairing longer "l" values with brand words and other terms, and preliminary evaluation results indicate that this method can enhance the semantic relevance and usability of spans. However, to further improve the quality and relevance of spans, we will combine the characteristics of the e-commerce scenario, using models or other methods to generate spans that include product words, attribute words, and brand words, and filter out irrelevant spans to reduce noise and improve accuracy.

- **Model Generalization Improvement**: Due to the vast and constantly evolving range of products in e-commerce scenarios, models struggle to cover all possible combinations of products and attributes, leading to poor generalization performance on new products. To address this challenge, we intend to employ incremental learning (Incremental Learning) or online learning (Online Learning) techniques to continuously update and refine the model, enabling it to adapt to new product and attribute combinations and enhance generalization performance.

References

Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. *Advances in Neural Information Processing Systems*, 35:31668–31683.

Peng Cheng, Huimu Wang, Jinyuan Zhao, Yihao Wang, Enqiang Xu, Yu Zhao, Zhuojian Xiao, Songlin Wang, Guoyu Tang, Lin Liu, et al. 2024. Modrl-ta: A multi-objective deep reinforcement learning framework for traffic allocation in e-commerce search. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3694–3698.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paolo Ferragina and Giovanni Manzini. 2000. Opportunistic data structures with applications. In *Proceedings 41st annual symposium on foundations of computer science*, pages 390–398. IEEE.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338.

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

Sunkyung Lee, Minjin Choi, and Jongwuk Lee. 2023. Glen: Generative retrieval via lexical index learning. *arXiv preprint arXiv:2311.03057*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Mingming Li, Chunyuan Yuan, Binbin Wang, Jingwei Zhuo, Songlin Wang, Lin Liu, and Sulong Xu. 2023a. Learning query-aware embedding index for improving e-commerce dense retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3265–3269.

Mingming Li, Chunyuan Yuan, Huimu Wang, Peng Wang, Jingwei Zhuo, Binbin Wang, Lin Liu, and Sulong Xu. 2023b. Adaptive hyper-parameter learning for deep semantic retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 775–782.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2023c. Learning to rank in generative retrieval. *arXiv preprint arXiv:2306.15222*.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2023d. Multiview identifiers enhanced generative retrieval. *arXiv preprint arXiv:2305.16675*.

Yiming Qiu, Chenyu Zhao, Han Zhang, Jingwei Zhuo, Tianhao Li, Xiaowei Zhang, Songlin Wang, Sulong Xu, Bo Long, and Wen-Yun Yang. 2022. Pre-training tasks for user intent detection and embedding retrieval in e-commerce search. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4424–4428.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. 2024a. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. 2024b. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36.

Yubao Tang, Ruqing Zhang, Jiafeng Guo, Jiangui Chen, Zuowei Zhu, Shuaiqiang Wang, Dawei Yin, and Xueqi Cheng. 2023. Semantic-enhanced differentiable search index inspired by learning strategies. *arXiv preprint arXiv:2305.15115*.

Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al. 2022. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, 35:21831–21843.

Binbin Wang, Mingming Li, Zhixiong Zeng, Jingwei Zhuo, Songlin Wang, Sulong Xu, Bo Long, and Weipeng Yan. 2023. Learning multi-stage multi-grained semantic embeddings for e-commerce search. In *Companion Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 411–415. ACM.

Huimu Wang, Mingming Li, Dadong Miao, Songlin Wang, Guoyu Tang, Lin Liu, Sulong Xu, and Jinghe Hu. 2024. A preference-oriented diversity model based on mutual-information in re-ranking for e-commerce search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2895–2899.

Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, et al. 2022. A neural corpus indexer for document retrieval. *Advances in Neural Information Processing Systems*, 35:25600–25614.

Peiwen Yuan, Xinglin Wang, Shaoxiong Feng, Boyuan Pan, Yiwei Li, Heda Wang, Xupeng Miao, and Kan Li. 2024. Generative dense retrieval: Memory can be a burden. *arXiv preprint arXiv:2401.10487*.

Han Zhang, Hongwei Shen, Yiming Qiu, Yunjiang Jiang, Songlin Wang, Sulong Xu, Yun Xiao, Bo Long, and Wen-Yun Yang. 2021. Joint learning of deep retrieval model and product quantization based embedding index. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1718–1722.

Han Zhang, Songlin Wang, Kang Zhang, Zhiling Tang, Yunjiang Jiang, Yun Xiao, Weipeng Yan, and Wen-Yun Yang. 2020. Towards personalized and semantic retrieval: An end-to-end solution for e-commerce search via embedding learning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2407–2416.

Peitian Zhang, Zheng Liu, Yujia Zhou, Zhicheng Dou, Fangchao Liu, and Zhao Cao. 2024. Generative retrieval via term set generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 458–468.

Yujia Zhou, Zhicheng Dou, and Ji-Rong Wen. 2023. Enhancing generative retrieval with reinforcement learning from relevance feedback. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Shengyao Zhuang, Houxing Ren, Linjun Shou, Jian Pei, Ming Gong, Guido Zuccon, and Daxin Jiang. 2022. Bridging the gap between indexing and retrieval for differentiable search index with query generation. *arXiv preprint arXiv:2206.10128*.