# TREC 2023: h2oloo in the Product Search Challenge

**Jheng-Hong Yang**
University of Waterloo
jheng-hong.yang@uwaterloo.ca

**Jimmy Lin**
University of Waterloo
jimmylin@uwaterloo.ca

## Abstract

This paper presents the submitted runs for the TREC 2023 Product Search track, offering insights into our multi-stage retrieval systems designed for both end-to-end retrieval and reranking tasks. In our approach, we employed a sparse first-stage ranker that leveraged textual information, complemented by a dense first-stage ranker tailored for processing visual data. Additionally, we evaluate the effectiveness of utilizing a large-language model within the context of product search, shedding light on its capabilities and contributions to improving retrieval performance. Our findings contribute to the ongoing discourse on enhancing product search techniques, showcasing the potential of combining various retrieval strategies and advanced language models for enhanced search accuracy.

## 1 Overview

**Disclaimer.** This notebook paper serves as a preliminary document to outline the context and toolkits utilized during our participation in the Product Search track at TREC 2023. It is important to note that the content presented here may undergo revisions and updates in the final proceedings. Therefore, we advise readers to use this notebook as a reference cautiously. Should you have any recommendations, inquiries, or require additional information, please do not hesitate to reach out to the authors. Your feedback and queries are greatly valued, and we are committed to providing accurate and up-to-date information to assist you in your understanding of our participation in the TREC 2023 Product Search track.

In our participation in the TREC 2023 Product Search track, we deployed a range of retrieval and ranking systems, each tailored to harness the strengths of specific algorithms and models. These systems were meticulously designed to enhance retrieval effectiveness, taking advantage of both traditional information retrieval methods and state-of-the-art neural network models. Below, we provide an overview of the systems we submitted:

- **BM25**: For this system, we leveraged the widely-used BM25 implementation with default parameters available in the Anserini toolkit.[1] BM25 is a classic probabilistic retrieval model that computes the relevance of documents based on term frequencies and document length normalization. It served as one of our foundational approaches to baseline retrieval.
- **SPLADE++**: SPLADE++ represents a significant advancement in neural text-based retrieval.[2] In our system, we employed a checkpoint of the SPLADE++ model to encode the query and create an index. SPLADE++ is known for its robust performance in capturing relevance and context in text-based retrieval tasks. This system contributed to our text-based retrieval strategies.

[1] https://github.com/castorini/anserini
[2] https://huggingface.co/naver/splade-cocondenser-ensembledistil

Product Search track at 32th Text REtrieval Conference (TREC 2023).

- **CLIP**: Our image-based retrieval system relied on the powerful OpenCLIP framework, utilizing checkpoints pretrained on the LAION dataset.[3] CLIP is a vision-language model that excels in understanding the semantic relationships between images and text. By incorporating CLIP, we aimed to enhance our image-based retrieval strategies, recognizing the importance of visual content in product search.
- **LLMRank**: In our pursuit of improved ranking and relevance, we integrated the RankGPT library to create the LLMRank system.[4] Specifically, we employed the gpt-3.5-turbo model to rerank the top-30 candidates. GPT-3.5-turbo is an advanced language model capable of generating human-like text and understanding complex queries, making it a valuable asset for reranking and fine-tuning our retrieval results.

Each of these systems was carefully selected and configured to address specific aspects of the retrieval task. By combining traditional and modern retrieval methods, as well as leveraging the strengths of neural network models, we aimed to achieve a comprehensive and effective approach to product search. Our submitted systems represent a strategic fusion of established techniques and cutting-edge technologies to optimize retrieval performance in the TREC 2023 Product Search track.

## 2 Results

The presented table provides a comprehensive view of the experimental results from the TREC 2023 Product Search track, where various retrieval and fusion strategies were evaluated to enhance retrieval effectiveness. These strategies encompassed a wide range of approaches, from basic baselines to sophisticated models that leverage both text and image data. In this section, we will delve into the key findings and insights derived from these results.

**Baselines.** The initial comparison between the two baseline retrieval strategies, "Simple" and "Full," sets the stage for our analysis. The "Simple" strategy, while achieving a moderate nDCG@10 score of 0.213, faces challenges in maintaining precision with larger result sets, as indicated by its R@1K score of 0.596. On the other hand, the "Full" baseline, which likely incorporates more features, demonstrates improved performance with an nDCG@10 score of 0.248 and a higher R@1K score of 0.663. This contrast emphasizes the importance of a strong baseline, as it provides a reference point for evaluating the effectiveness of more advanced strategies.

**Retriever: Text.** Moving on to the text-based retrieval strategies, several insights emerge. The "BM25 (T)" strategy, which leverages the BM25 algorithm on title text, performs well with an nDCG@10 score of 0.256 and a reasonable R@1K score of 0.607, indicating its effectiveness in harnessing title information. In contrast, "BM25 (D)," which applies BM25 to description text, lags behind in terms of nDCG@10 (0.154) and R@1K (0.427) scores, suggesting that description text alone may be less informative or discriminative for retrieval. Interestingly, the "BM25 (C)" strategy, which uses BM25 on category text, demonstrates the least effectiveness among text-based approaches with an nDCG@10 score of 0.020 and an R@1K score of 0.263. This implies that category information alone may not suffice for accurate retrieval, reaffirming the importance of leveraging richer text sources. The "BM25 (T+D)" strategy, which combines BM25 scores from title and description text, outperforms using these sources individually with an nDCG@10 score of 0.212 and an R@1K score of 0.593. This highlights the significance of synergy between different text sources and the potential for improving retrieval effectiveness by combining them. The SPLADE++ algorithm also plays a crucial role in text-based retrieval. "SPLADE++ (T)" excels in capturing relevance in both top results and larger result sets, with an nDCG@10 score of 0.267 and an R@1K score of 0.648. "SPLADE++ (D)" performs slightly lower but still effectively with description text, suggesting the versatility of SPLADE++ in different text contexts. Conversely, "SPLADE++ (C)," which relies solely on category text, demonstrates the least effectiveness among the SPLADE++ strategies. This reinforces the idea that category information alone may not be as discriminative as title or description text in retrieval tasks. The "SPLADE++ (T+D)" strategy, combining SPLADE++ scores from title and description text, emerges as the top performer among text-based strategies, with an nDCG@10 score of 0.282 and an R@1K score of 0.672. This underlines the advantage of leveraging both title and description information using SPLADE++ for robust retrieval.

[3] `https://github.com/mlfoundations/open_clip`

[4] `https://github.com/sunnweiwei/RankGPT`

**Retriever: Image (CLIP).** In the realm of image-based retrieval strategies, the utilization of CLIP with different Vision Transformer (ViT) models offers valuable insights. The "ViT-B-32" model yields a modest nDCG@10 score of 0.043, indicating that image-based retrieval captures relevant items, albeit with less effectiveness when compared to text-based strategies alone. However, the "ViT-L-14" model outperforms "ViT-B-32" with an nDCG@10 score of 0.128, highlighting that larger and more complex image models have the potential to significantly enhance retrieval effectiveness in the context of product search. Furthermore, "ViT-G-14" pushes the boundaries further by achieving an nDCG@10 score of 0.168. This result underscores the scaling effect of model size, indicating that increasing the model's capacity can further improve performance in the product search task when utilizing image-based retrieval strategies. This result underscores the scaling effect of model size, indicating that increasing the model's capacity can further improve performance in the product search task when utilizing image-based retrieval strategies. These findings collectively illustrate that while image-based retrieval may not match the strength of text-based methods, it still provides substantial value. Moreover, the results emphasize that sophisticated image models and scaling strategies can be instrumental in unlocking the full potential of image-based retrieval techniques in product search.

**Fusion & Reranker Strategies.** The fusion and reranking strategies offer a fascinating dimension to the analysis. "sum(c, j)," which combines BM25 (T+D) and SPLADE++ (T+D), achieves an impressive nDCG@10 score of 0.296 and an R@1K score of 0.680, demonstrating the benefits of integrating text-based strategies. "sum(c, j, m)," which further incorporates ViT-G-14 into the fusion, maintains a high level of effectiveness with an nDCG@10 score of 0.272 and an R@1K score of 0.675. This extended fusion strategy outperforms text-only or image-only approaches, highlighting the synergy of combining different modalities. Finally, "sum(c, j, m) + GPT" introduces GPT into the strategy, resulting in a remarkable nDCG@10 score of 0.623 and an R@1K score of 0.929 on the evaluation dataset. This inclusion of GPT significantly improves retrieval effectiveness, underscoring the value of incorporating advanced language models for reranking. In conclusion, the experimental results provide valuable insights into the effectiveness of various retrieval and fusion strategies. They demonstrate the importance of leveraging different text sources, image representations, and advanced language models to enhance product search retrieval, ultimately offering a roadmap for future improvements in this domain.

| | | | Dev | | | Eval | | |
|---|---|---|---|---|---|---|---|---|
| Entry | Description | Run ID | nDCG@10 | R@10 | R@1K | nDCG@10 | R@10 | R@1K |
| | | | Baselines | | | | | |
| a | Simple | - | 0.213 | 0.132 | 0.596 | | | |
| b | Full | - | 0.248 | 0.158 | 0.663 | | | |
| | | | Retriever: Text | | | | | |
| c | BM25 (T) | - | 0.256 | 0.159 | 0.607 | | | |
| d | BM25 (D) | - | 0.154 | 0.090 | 0.427 | | | |
| e | BM25 (C) | - | 0.020 | 0.013 | 0.263 | | | |
| f | BM25 (T+D) | - | 0.212 | 0.131 | 0.593 | | | |
| g | SPLADE++ (T) | - | 0.267 | 0.165 | 0.648 | | | |
| h | SPLADE++ (D) | - | 0.177 | 0.103 | 0.456 | | | |
| i | SPLADE++ (C) | - | 0.023 | 0.015 | 0.308 | | | |
| j | SPLADE++ (T+D) | - | **0.282** | **0.175** | **0.672** | | | |
| | | | Retriever: Image (CLIP) | | | | | |
| k | ViT-B-32 | - | 0.043 | 0.025 | 0.223 | | | |
| l | ViT-L-14 | - | 0.128 | 0.073 | 0.420 | | | |
| m | ViT-G-14 | - | 0.168 | 0.098 | 0.471 | | | |
| | | | Fusion & Reranker | | | | | |
| n | sum(c, j) | `f_splade_bm25` | 0.296 | 0.184 | 0.680 | 0.751 | - | 0.923 |
| o | sum(c, j, m) | `f_splade_clip_bm25` | 0.272 | 0.169 | 0.675 | 0.733 | - | 0.929 |
| p | sum(c, j, m) + GPT | `f_gpt_rerank` | - | - | - | 0.623 | - | 0.929 |
| q | Full + GPT | `r_gpt3d5_turbo` | - | - | - | 0.595 | - | 0.727 |
| | | `median` | | | | 0.449 | | |
| | | `max` | | | | 0.875 | | |

Table 1: Results of TREC 2023 Product Search track experiments. T: title, D: description, C: category. The "sum" indicates min-max normalization + score fusion. The term "GPT" refers to GPT-3.5-Turbo.