Enhancing Related Searches Recommendation system by leveraging LLM Approaches

Hung Nguyen
eBay Inc
San Jose, CA, USA
hungnguyen@ebay.com

Jayanth Yetukuri
eBay Inc
San Jose, CA, USA
jyetukuri@ebay.com

Phuong Ha Nguyen
eBay Inc
San Jose, CA, USA
phuongha.ntu@gmail.com

Lizzie Liang
eBay Inc
San Jose, CA, USA
yliang@brandeis.edu

Ishita Khan
eBay Inc
San Jose, CA, USA
ishikhan@ebay.com

Jiang yu
eBay Inc
San Jose, CA, USA
jyu6@ebay.com

Zhe Wu
eBay Inc
San Jose, CA, USA
zhewu1@ebay.com

Abstract

Leveraging user historical search data and collaborative filtering is a common approach for generating suggested search queries on e-commerce platforms. However, this method often results in a limited number of suggestions, particularly for non-head queries, due to the reliance on identifying similar behaviors across users, and for these queries, the number of users exhibiting such behaviors is often small. In this paper, we study user engagement behavior and propose a method to expand suggestion lists, focusing on, but not limited to, non-head queries, by integrating large language model (LLM) techniques with traditional collaborative filtering based on user historical data. In addition to using generative LLMs to directly generate suggestions, we also utilize embeddings from pre-trained LLM models to capture both buyer query intent and synonymous queries in a latent space, thereby providing suggestions that are closer to the buyer's intent.

Our study reveals that while users are open to exploring diverse products, they tend to favor search results closely aligned with their intent when making purchase decisions. We propose an approach that enhances both the diversity and relevance of search suggestions, addressing the limitations of traditional collaborative filtering methods. Through real-world A/B testing on an e-commerce platform, we found that while the large language model (LLM)-based approach boosted user engagement, the embedding-based method led to a significant 0.77% improvement in Gross Merchandise Bought (GMB). This improvement translates into a notable revenue increase, highlighting the potential of LLM applications to optimize search query suggestion systems.

CCS Concepts

• **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Keywords

e-commerce, related search queries, LLM, embedding

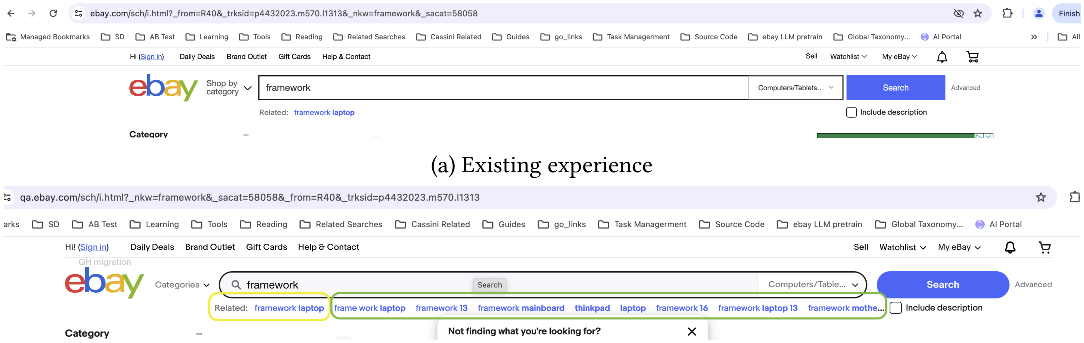
ACM Reference Format:
Hung Nguyen, Jayanth Yetukuri, Phuong Ha Nguyen, Lizzie Liang, Ishita Khan, Jiang yu, and Zhe Wu. 2025. Enhancing Related Searches Recommendation system by leveraging LLM Approaches. In *Proceedings of the second workshop on Generative AI for E-Commerce 2025, September 22, 2025.* ACM, New York, NY, USA, 6 pages.

1 Introduction

The journey of a buyer at a typical e-Commerce platform consists of issuing sequential search keywords a.k.a buyer *query* until he/she finds the desired product. The back-end Search Engine (SE) consists of a series of complex algorithms and models for improving buyer satisfaction [4, 9, 10]. The complexity of these steps drastically increases for a platform like eBay, which deals with various collector items and is driven by seller-provided details. To enhance buyer satisfaction and accelerate the decision making process, the notion of *Related Searches (RS)* [3] is introduced where alternate suggested queries for the buyer provided query are often displayed either beneath the *search bar* or at the *bottom of page* in an e-commerce website. These alternate queries are often of high quality, mined from user logs [11], presented directly to the buyers.

While the infrastructure of Related Searches provides a mechanism to surface query alternatives, the effectiveness of this module fundamentally depends on the relevance and intent-alignment of the suggested queries. Designing such a system requires a nuanced understanding of buyer behavior, particularly how users reformulate queries during their shopping journey. Presenting queries that were historically associated with prior searches does not guarantee

Genaiecom '25, Prague, CZ



(a) Existing experience

(b) Proposed experience

Figure 1: Enhancing related suggestions buyer experience for Non-head queries. The suggestion in yellow box is obtained using collaborative filtering, whereas the suggestions in the green box are obtained using the proposed approach.

improved user experience or coverage. Instead, there is a growing need to model the latent intent transitions embedded in user interactions and leverage them to craft more intelligent and adaptive query suggestions.

A critical question for engineers is determining the types of queries that should be suggested to users in order to enhance engagement or help them find relevant products. *Related Searches* is specifically designed to address this need. However, indiscriminately suggesting random queries in the hope of increasing engagement or facilitating a purchase is not an effective strategy. A more systematic solution involves employing collaborative filtering techniques by analyzing anonymized user search history logs. More specifically, one can construct pairs consisting of a user's initial search keyword and the corresponding suggestions based on historical search sessions, which can then be stored in a cache. When other users input the same initial keyword, the corresponding suggestions from the cache are offered. This approach is particularly effective for *head queries*, which are characterized by high user engagement and a wide range of related search queries. However, for a substantial portion of non-head queries—often numbering in the millions on a daily basis—it is challenging to generate an adequate set of suggestions for each query using collaborative filtering alone.

This study aims to enhance RS module with the intention of improving user engagement. We build on a success-aware production baseline—session-level collaborative filtering that only surfaces queries with demonstrated success (historical clicks). We provide two powerful approaches which leverage a variety of state of the art transformer-based models, generative Large Language Models (LLM) and Bidirectional Encoder Representations from Transformers (Bert) models. The first approach utilizes LLM to generate suggestions using prompt engineering and the latter approach **utilizes model embedding distance to retrieve nearest neighbors from this success-filtered suggestion table**. While the former showed success in performance metrics (CTR, Abandonment rate) **the latter—grounded in historically successful queries—** brings significant gain in both performance and business metrics (GMB, CTR, Abandonment rate). Our contributions with this study include:

(1) **User Intent Alignment:** Our analysis reveals that users tend to stay closely aligned with their original search intent while interacting with the RS module. This behavior highlights the importance of designing systems that respect and reinforce the user's initial search objectives, rather than introducing suggestions that might lead the user too far from their original query intent.
(2) **Proposed Algorithm:** We introduce a novel algorithm for query suggestions that goes beyond traditional collaborative filtering techniques. While collaborative filtering remains a common approach for generating recommendations based on historical user behavior, our method incorporates additional features such as query-context analysis and intent classification, resulting in more accurate and contextually relevant suggestions.
(3) **Business Impact:** To assess the effectiveness of our approach, we conducted A/B tests focused on key business metrics. The results of these tests show a statistically significant improvement, indicating that our approaches not only enhances user experience but also contributes positively to core business objectives such as conversion rates and user engagement.

2 Related Searches User Behavior Study

This section presents a focused analysis of user engagement behavior with the Related Searches (RS) module—a query suggestion interface positioned below the search bar [2], as illustrated in Figure 1. Recent findings [8] directly informed and motivated the expansion of our query suggestion system.

We collected two weeks of interaction data on the RS module and employed an in-house state-of-the-art Named Entity Recognition (NER) model [7] to segment user search queries into structured aspects. For this study, we selected the most frequent query aspects, including *brand*, *color*, *product type*, *size*, *material*, and *product model*. For example, the query "nike running shoes 11" is decomposed into: 'nike' as *brand*, 'running' as *occasionUsage*, 'shoes' as *product type*, and '11' as *size*.

We then analyzed the statistical relationship between the aspects of the user's initial query and those of the engaged suggestions.

Table 1: Buyer engagement (clicks) and conversion segmented by suggestion type. Each value shows the fractional lift in of engagement with respect to the lowest engagement. For example, queries having the suggestions with *Same Size* has *55.55%* more engagement with respect to *Added Size*. The top numbers at each category are highlighted in green.

Aspect Type	Click				Conversion			
	Same	Different	Removed	Added	Same	Different	Removed	Added
Brand	+0.00%	+16.28%	+9.74%	+1.60%	+16.18%	+0.00%	+4.41%	+2.94%
Product Type	+32.09%	+7.01%	+1.57%	+0.00%	+36.76%	+0.00%	+0.00%	+0.00%
Size	+55.55%	+43.89%	+6.16%	+0.00%	+71.79%	+0.00%	+15.38%	+43.59%
Material	+14.72%	+1.43%	+0.13%	+0.00%	+61.19%	+0.00%	+10.45%	+11.94%
Color	+86.86%	+51.96%	+47.65%	+0.00%	+85.94%	+20.31%	+54.69%	+0.00%
Model	+15.13%	+5.53%	+6.23%	+0.00%	+40.30%	+14.93%	+0.00%	+8.96%

We categorized suggestion relevance into four types: *Same Aspect* (shared aspect present in both initial and suggested queries), *Removed Aspect* (aspect present in the initial query but absent in the suggestion), *Added Aspect* (absent in the initial query but introduced in the suggestion), and *Different Aspect* (aspects differ entirely). We computed engagement metrics, specifically click-through and conversion rates, normalized as lifts relative to the lowest value in each category to protect absolute performance data.

As shown in Table 1, user engagement—particularly clicks—is highest when the suggestion retains the same aspects as the original query. Interestingly, while users occasionally explore different *brands*, conversion rates consistently peak in the *Same Aspect* group across all other dimensions. This underscores a strong alignment with original intent during exploration and decision-making phases.

Based on this insight, we enhance our query suggestion system by generating additional intent-aligned suggestions. To supplement traditional collaborative filtering, we incorporate transformer-based embedding similarity to retrieve semantically coherent alternatives that maintain aspect consistency with the user's original search.

3 Methodology

This section outlines the core methodology behind our approach to improving buyer-facing search suggestion systems. We begin with a brief overview of the collaborative filtering strategy currently employed to generate the baseline recommendation system data [6]. We then detail two experimental approaches designed to enrich the set of related search suggestions, with a particular focus on expanding coverage for non-head (long-tail) queries. Due to the user interface constraints of e-commerce platforms—such as the limited real estate under the search bar—we restrict the number of displayed suggestions to 12 per query. Consequently, these enhancements yield more substantial gains for non-head queries, where existing coverage is often sparse, compared to head queries that typically already saturate the available suggestion slots.

3.1 Baseline: Collaborative Filtering

The current framework is based on a session-level collaborative filtering method that exploits co-occurrence patterns in user search behavior. As illustrated in Figure 2, we mine transitions between sequential queries issued within a user session to identify candidate reformulations. If a user issues query A followed by query B in the same session, we treat B as a potential suggestion for A. This method relies on the behavioral assumption that successive queries often reflect refinement (e.g., increased specificity or filter application) or semantic shifts toward complementary or alternative intents.

To ensure robustness and reduce noise, we apply frequency-based filtering: we retain only those query pairs (A, B) that appear in at least three unique user sessions. This frequency threshold helps eliminate one-off transitions and surface only statistically meaningful patterns. The resulting output is a suggestion table containing a set of primary queries and their associated reformulations or continuations. For example, if both (A, B) and (A, C) are frequent transitions, the system records A as the primary query and aggregates B and C as related suggestions.

The suggestion table is indexed and cached in the RS infrastructure, enabling real-time retrieval during user interactions. At runtime, when a user searches for query A, the system retrieves and displays the corresponding set of precomputed suggestions—e.g., B and C—within the allocated UI component. While effective for head queries, this method suffers from recall sparsity for less frequent (tail) queries, motivating the need for more dynamic and scalable suggestion generation strategies.

To further enhance the quality and utility of the mined query suggestions, a post-processing step is applied that specifically filters for *null* and *low recall* queries. This step is motivated by the observation that a substantial proportion of user dissatisfaction in e-commerce search stems from queries that fail to retrieve meaningful results—either due to inventory sparsity, overly specific query formulations, or vocabulary mismatch. By incorporating this recall-aware filtering step, we not only refine the quality of the suggestion dataset but also strategically focus on high-leverage queries where intelligent reformulation can deliver significant improvements in user experience, conversion rates, and overall search performance. This filtering process ensures that the system adapts to recall deficiencies in real time and prioritizes the enrichment of suggestions where they are most needed.

3.2 Approach 1: Embedding based suggestions

In this approach, we implement a post-processing technique over the existing suggestion data table. Specifically, we utilize a domain-specific BERT model [1], which is specifically pre-trained on one billion of queries and items' titles at an e-commerce website, to generate high dimensional embeddings for the primary queries in the existing suggestion table. This table, containing primary queries

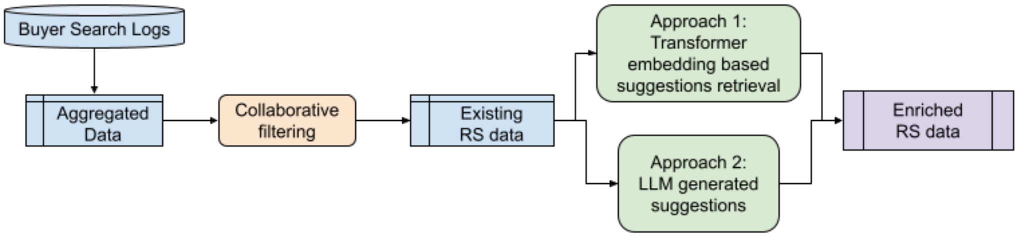

Buyer Search Logs
Aggregated Data
Collaborative filtering
Existing RS data
Approach 1: Transformer embedding based suggestions retrieval
Approach 2: LLM generated suggestions
Enriched RS data

Figure 2: An overview of the framework for generating Related Suggestions in an e-commerce domain.

Table 2: Samples from the related suggestions dictionary using both traditional approach and proposed approach.

Primary Query	Collaborative filtering	Embedding based suggestions	LLM-based suggestions
global gladiators	global gladiators sega genesis	gladiator premier cabinet samurai pegasus	global gladiator game global gladiator movie poster
golf iron	golf iron set	golf wedge, used golf driver ping bag, iron set, driving iron	used golf irons, best golf irons golf iron covers, golf iron shafts

and their embeddings, is indexed in a database. Each primary query embedding is then used to perform a nearest neighbor search using cosine similarity against the pre-indexed primary query embeddings in the rest of the table. Once the nearest neighbor records in the table are found, their suggestions are appended after removing duplicates, if any exist. This method significantly enriches the suggestion table while maintaining suggestion quality, as all suggestions are derived from a high-quality existing database.

This approach is straightforward yet significantly improves the quantity of query recommendations. More importantly, it ensures their quality by considering semantic similarities rather than relying solely on random suggestions. This leads to a better user experience, higher search relevance, and increased conversion rates. Overall, leveraging a pre-trained BERT model for query reformulation not only enhances the number of suggestions but also ensures that they are not merely based on surface-level keyword matching. Instead, they are grounded in deeper semantic relevance, improving both the quality and diversity of suggestions.

3.3 Approach 2: LLM generated suggestion

This approach explores the use of a Large Language Model (LLM) to generate query suggestions. Unlike traditional retrieval-based methods that rely on historical data, LLMs can generate novel, contextually relevant query suggestions by understanding natural language patterns. The key objective here is to determine whether an LLM-based query suggestion system can enhance search relevance and improve user engagement on an e-commerce platform. For this experiment, we utilized open sourced Llama-3 [5], a large-scale language model available via HuggingFace framework, which topped the leader board at the time of running this experiment. Llama-3 is designed to generate coherent and contextually appropriate text, making it a viable candidate for generating RS suggestions. The following prompt was used to instruct the LLM to generate related search queries:

> "**`You are a search expert at an e-commerce website like eBay and your job is to generate related queries that will be shown under the search bar. Given a user search query, please recommend 6 related search queries that user might be interested in exploring and making purchases. Example 1: ‘user query’: yerse, ‘related search queries’: [yerse dresses; yerse tops and blouses; yerse knitwear; yerse outwear; yerse accesories; yerse coats]. Example 2: ‘user query’: 08 Tribeca front sway bar bushes, ‘related search queries’: [08 Tribeca suspension parts; 08 Tribeca front sway bar; 08 Tribeca bushings replacement; 08 Tribeca automotive parts; Subaru SUV maintenance parts; Tribeca bushings replacement].`**"

This prompt was designed based on human judgment, by examining result samples, ensuring clarity and alignment with typical user expectations. The simplicity of the prompt was intended to test the baseline performance of the model without additional context or fine-tuning. The generated suggestions, especially for non-head queries (i.e., long-tail or less frequent queries), were stored in a cache. This caching mechanism aimed to reduce latency and enhance scalability, ensuring that once an LLM-generated suggestion was created, it could be reused efficiently.

Despite the promising potential of LLM-generated suggestions, several challenges were identified:

(1) The generated suggestions often lacked semantic variety, producing queries that were too similar to one another. The model struggled to incorporate e-commerce-specific language and product taxonomy, leading to suggestions that were too general.

(2) This was evident when analyzing user engagement metrics. i) Engagement increased, meaning users interacted with suggestions more frequently. ii) However, gross merchandise bought (GMB) decreased, indicating that users were unable to find relevant products to purchase. This suggests that while the model generated syntactically valid queries, they did not necessarily align with buyer intent or industry-specific terminology.

In addition to recall-based filtering, the generated query candidates are subjected to a *domain-specific relevance filtering* step. This filtering is designed to ensure that only semantically coherent and contextually appropriate suggestions are surfaced to end users, particularly in e-commerce environments where lexical ambiguity or overgeneralization can result in irrelevant or misleading recommendations. This domain-specific relevance filtering step is critical not only for maintaining the quality and precision of query suggestions but also for ensuring trust and utility in the overall user experience. It minimizes the risk of suggestion-based query drift and enhances the coherence of user navigation paths within the search funnel. After this filtering step, there are 58% (approach 1) and 62% (approach 2) of the primary queries have been added new suggestions.

4 Evaluation

In this section, we present the outcomes of a controlled A/B test conducted to compare different approaches for query suggestion within the Related Searches (RS) module. The production-grade collaborative filtering system serves as the **Control** baseline, while two experimental variants—an embedding-based approach and a generative LLM-based approach—are treated as **Treatment** arms.

Our evaluation framework spans both business-centric and behavioral performance metrics to assess each method's influence on user engagement and marketplace efficiency. Table 3 summarizes the results. Business outcomes were quantified using Gross Merchandise Bought (GMB) and Bought Item Count (BIC), which serve as proxies for revenue generation and item-level transaction volume, respectively. The embedding-based approach yielded a statistically significant uplift across both GMB and BIC, indicating superior alignment with user purchase intent. Buyers exposed to these suggestions not only completed more purchases but also tended to explore and buy a larger variety of items—evidence of improved product discovery and intent fulfillment.

The LLM-based approach, while effective in increasing exploratory engagement, did not translate into proportional gains in revenue. This is likely due to the lack of domain grounding in its generative output, which occasionally led to off-target or irrelevant suggestions. As a result, the method showed diminished conversion effectiveness compared to the embedding-based variant.

From a behavioral standpoint, we examined Click-Through Rate (CTR), Abandonment Rate, and RS Driven Engagement. The embedding based approach showed marked improvement in CTR, reflecting better top-k suggestion relevance. It also achieved a lower abandonment rate, suggesting that users were more often able to find satisfactory results without exiting or reissuing queries—thereby reducing friction in the search process. Although the LLM-based method showed a moderate increase in session-level engagement, this engagement did not yield a commensurate increase in downstream conversions. This imbalance indicates that while users found the generative suggestions intriguing, they were less actionable in a commerce context.

Table 3: Online A/B test results with a 95% confidence interval. The top two metrics focus on business outcomes, while the bottom three metrics assess performance.

Metric	Improvement (%)	
	Approach 1	Approach 2
Gross Merchandise Bought	+0.77	-0.74
Bought Item Count	+0.5	neutral
Click Through Rate	+3.75	+1.34
Abandonment Rate	-0.31	neutral
RS driven engagement	+5.00	+11.10

Overall, the embedding-based approach offers a more effective balance between exploration and intent satisfaction for high-stakes e-commerce search. By retrieving from a success-filtered corpus and using task-specific embeddings trained on large-scale in-session reformulations, it produces semantically grounded, intent-preserving suggestions that align with user needs and available inventory, translating into measurable lifts in purchases and gross merchandise bought (GMB) alongside improvements in CTR and abandonment.

In contrast, LLM-generated suggestions excel at linguistic diversity and generalization but lack explicit alignment to historical success signals (clicks/purchases) and catalog constraints, which can depress precision in domain-sensitive settings. This explains why LLMs may underperform despite stronger surface fluency.

Looking forward, hybrid architectures—e.g., using LLMs to propose diverse candidates for tail queries and then applying success-aware, embedding-based re-ranking (or purchase-weighted scoring) with inventory awareness—appear promising. Such designs can preserve behavioral fidelity while retaining the contextual flexibility of LLMs, improving relevance, diversity, and scalability in production related-search systems.

5 Conclusion

This paper presents a hybrid framework for enhancing query suggestion systems in e-commerce by combining collaborative filtering with semantic retrieval models, including transformer-based embeddings (e.g., BERT) and generative Large Language Models (LLMs). Through rigorous A/B testing in a production environment, we demonstrate that while LLM-based approaches can increase user engagement by offering linguistically diverse and novel query suggestions, embedding-based methods grounded in buyer reformulation data consistently deliver superior outcomes in business-critical metrics such as Gross Merchandise Bought (GMB) and Click-Through Rate (CTR). These results underscore the importance of domain-aligned semantic representation in driving search effectiveness and conversion.

Our methodology reveals that embeddings trained on historical buyer behavior yield recommendations that are both intent-aware and transactionally effective—outperforming LLMs in high-precision commercial use cases. Future work will explore the integration of prompt engineering to further align generative outputs with marketplace dynamics, as well as hybrid architectures that leverage the semantic grounding of embeddings and the contextual flexibility of LLMs. This direction aims to bridge personalization and relevance in query suggestions, supporting scalable improvements in user experience and revenue generation.

References

[1] Dan Schonfeld Chen Xue, Jesse Lute and Guoping Han. [n. d.]. How eBay Created a Language Model With Three Billion Item Titles - innovation.ebayinc.com. https://innovation.ebayinc.com/tech/engineering/how-ebay-created-a-language-model-with-three-billion-item-titles Accessed 21-02-2025.

[2] Keri Engel. 2025. *Related Searches: Boost Your SEO with User Search Data.* https://explodingtopics.com/blog/related-searches

[3] Mohammad Al Hasan, Nish Parikh, Gyanit Singh, and Neel Sundaresan. 2011. Query suggestion for E-commerce sites. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (Hong Kong, China) *(WSDM '11)*. Association for Computing Machinery, New York, NY, USA, 765–774. doi:10.1145/1935826.1935927

[4] Sharon Hirsch, Ido Guy, Alexander Nus, Arnon Dagan, and Oren Kurland. 2020. Query Reformulation in E-Commerce Search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1319–1328.

[5] Meta AI. [n. d.]. Meta Llama 3. https://ai.meta.com/blog/meta-llama-3/. Accessed: 2025-08-22.

[6] Jacob Murel. [n. d.]. What is collaborative filtering? | IBM — ibm.com. https://www.ibm.com/think/topics/collaborative-filtering#:~:text=Collaborative%20filtering%20is%20an%20information,have%20interacted%20with%20that%20item. [Accessed 27-02-2025].

[7] Musen Wen, Deepak Kumar Vasthimal, Alan Lu, Tian Wang, and Aimin Guo. [n. d.]. Building Large-Scale Deep Learning System for Entity Recognition in E-Commerce Search - dl.acm.org. https://dl.acm.org/doi/abs/10.1145/3365109.3368765 Accessed 21-02-2025.

[8] Jayanth Yetukuri, Mehran Elyasi, Samarth Agrawal, Aritra Mandal, Rui Kong, Harish Vempati, and Ishita Khan. 2025. AI Guided Accelerator For Search Experience. *arXiv e-prints* (2025), arXiv–2508.

[9] Jayanth Yetukuri and Ishita Khan. 2025. Intent-Aware Neural Query Reformulation for Behavior-Aligned Product Search. *arXiv e-prints* (2025), arXiv–2507.

[10] Jayanth Yetukuri, Yuyan Wang, Ishita Khan, Liyang Hao, Zhe Wu, and Yang Liu. 2024. Multifaceted Reformulations for Null Low queries and its parallelism with Counterfactuals. In *"2024 IEEE 40th International Conference on Data Engineering (ICDE)"*. "5327–5333". doi:"10.1109/ICDE60146.2024.00401"

[11] Yang Zhang and Olfa Nasraoui. 2016. Mining search and browse behavior for query recommendation in e-commerce. *ACM Transactions on the Web (TWEB)* 10, 4 (2016), 1–25.

Received 10 August 2025