KUAISEARCH: AN E-COMMERCE SEARCH DATASET WITH AUTHENTIC QUERIES AND PRODUCT TEXTS FOR RECALL, RANKING, AND RELEVANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

E-commerce search connects user needs with massive product inventories, yet real-world systems face challenges from ambiguous queries, noisy product texts, and diverse user preferences. Recent advances in large language models (LLMs) offer new opportunities for semantic understanding and intent modeling, but existing e-commerce search datasets remain limited by heuristically constructed queries, popularity-based filtering, anonymized texts, and incomplete coverage of the search pipeline. These limitations hinder realistic evaluation of LLM-based e-commerce search. To address this gap, we introduce **KuaiSearch**, a large-scale dataset built from real user search interactions on the Kuaishou platform. KuaiSearch preserves authentic queries and natural-language product texts, retains cold-start users and long-tail products, and covers three key stages of the search pipeline: *recall*, *ranking*, and *relevance judgment*. We further provide comprehensive analyses of users, products, and queries, together with benchmark experiments on representative search tasks. Results demonstrate the value of KuaiSearch as a realistic benchmark for e-commerce search research. The code is available at `https://anonymous.4open.science/r/KuaiSearch-7230/`.

1 INTRODUCTION

E-commerce search plays a vital role in our daily lives by serving as the primary interface through which users express information needs, explore product content, and make purchasing decisions on online platforms (Tsagkias et al., 2021; Zhang et al., 2023). As the scale and diversity of online marketplaces continue to grow, e-commerce search has become a critical infrastructure component that directly influences user satisfaction, engagement, and conversion behavior, and largely determines the overall effectiveness of modern e-commerce systems. As a result, search quality is crucial for effectively connecting user demand with massive product inventories.

In e-commerce scenarios, user queries are often broad and ambiguous, typically consisting of brand names, product categories, or attribute-related keywords rather than well-specified product descriptions (Silverstein et al., 1999; Yano et al., 2016). Meanwhile, product-side texts are generally lengthy and noisy, as merchants may insert irrelevant or weakly related keywords to increase product exposure. Consequently, core information such as brands, attributes, and categories is often presented in an unstructured manner, weakening the discriminative capability of representation models (Wang et al., 2018; De Souza et al., 2018). In addition, e-commerce search exhibits substantial user heterogeneity, such that the same query may correspond to different budgets, preferences, and usage contexts. Effective search systems must therefore capture both query semantics and personalized user intent (Ai et al., 2019; Ai & Narayanan. R, 2021; Ai et al., 2017; Liu et al., 2022). Under these conditions, static relevance modeling between queries and product texts remains insufficient. Recent advances in large language models provide a promising direction for bridging this semantic gap through enhanced user profiling, preference modeling, and semantic understanding (Chen et al., 2025a; Zheng et al., 2025; Liu et al., 2025).

However, existing e-commerce search datasets remain limited in data scale, textual completeness, behavioral signals, and search-task coverage. Early studies have relied on Amazon datasets containing heuristically constructed queries, while filtering cold-start users and long-tail products (Van Gy-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Statistical comparison of e-commerce search datasets.

Dataset	# Users	# Items	# Queries	Text form
Amazon	192,403	63,001	3,221*	text
Amazon-C4	20,838	1,058,417	21,223*	text
JDsearch	173,831	12,872,736	171,728	anon'd
KuaiSearch-Lite	102,086	6,634,118	555,553	text
KuaiSearch	331,930	18,605,582	2,574,949	text

* Queries are constructed rather than collected from real user searches

sel et al., 2016; Ai et al., 2017; Hou et al., 2026). Such preprocessing produces data distributions that differ from real-world search scenarios. JDsearch provides real queries and multiple user–product interactions while preserving cold-start users and long-tail products (Liu et al., 2023); nevertheless, its anonymized textual content limits research on LLM-based search. More recently, Qilin provides natural-language text, but its setting is closer to content browsing, covers only one day of interactions, and lacks sufficient behavioral signals for modeling purchase intent (Chen et al., 2025b). Furthermore, most existing datasets focus on a single task, such as recall or ranking, making it difficult to evaluate multiple components within a shared real-world domain. Consequently, they remain insufficient for systematic research on real-world LLM-based e-commerce search.

To address the aforementioned challenges, we construct and release **KuaiSearch**, an e-commerce search dataset built upon one of the most popular short-video platforms in China, Kuaishou. KuaiSearch explicitly covers three key tasks in the e-commerce search pipeline—*recall*, *ranking*, and *relevance judgment*—and supports the evaluation of representative methods for each task within a shared e-commerce domain. The dataset consists of approximately 330,000 users, 18,000,000 products, and 2,500,000 real search queries, providing a large-scale benchmark for real-world e-commerce search research. During dataset construction, we fully retain real user search queries without applying any popularity-based filtering on products, and include users with diverse historical behavior lengths, thus preserving data distributions that closely reflect commercial environments. We further conduct multi-perspective analyses and benchmark representative models on KuaiSearch, demonstrating its research value and practical utility for e-commerce search.

In summary, the main contributions of this work are as follows:

- We release KuaiSearch, a large-scale dataset of real queries, full-category products, and long-term user behavior that preserves realistic e-commerce search distributions.
- KuaiSearch supports three key e-commerce search tasks: *recall*, *ranking*, and *relevance judgment*, providing a benchmark for representative methods across different tasks.
- We provide comprehensive dataset analysis and benchmark representative models on KuaiSearch, demonstrating its effectiveness and value for e-commerce search research.

2 Related Work

2.1 E-commerce Search

Early e-commerce search systems mainly rely on traditional information retrieval methods. Sparse retrieval approaches such as TF-IDF and BM25 (Salton & Buckley, 1988; Robertson et al., 2009) remain strong baselines due to their efficiency, while dense retrieval models such as DPR and ANCE (Karpukhin et al., 2020; Xiong et al., 2020) improve semantic matching by encoding queries and products into dense representations. Neural ranking models (Ai et al., 2019; Ai & Narayanan. R, 2021; Ai et al., 2017) further re-rank retrieved candidates with richer features for more fine-grained and personalized ranking. More recently, generative retrieval has emerged as a promising paradigm. Early works including DSI, DSI-QG, and LTRGR (Tay et al., 2022; Zhuang et al., 2022; Li et al., 2024) establish the foundation of generative retrieval, while OneSearch and OneSug (Chen et al., 2025a; Guo et al., 2025) explore unified generative modeling for e-commerce retrieval and ranking, and OneVision (Zheng et al., 2025) extends it to multimodal search. However, existing research is still constrained by the lack of large-scale e-commerce search datasets that jointly contain natural-language queries and authentic user behavior, limiting systematic evaluation in realistic scenarios.

Table 2: Field Descriptions and Data Sizes of Different Components in KuaiSearch.

Table	Size	Fields
User	331,930	user id, gender, age, location
Item	18,605,582	item id, title, brand id/name, seller id/name, category L1 id/name, category L2 id/name, category L3 id/name
Recall	2,574,949	user id, session id, time index, query, impressed item ids, clicked item ids, purchased item ids
Ranking	81,401,477	user id, user statistical features, fan number, follow number, session id, time index, query, search entrance, recently clicked item ids, recently purchased item ids, target item id, target item statistical features, target item price, is clicked, is purchased
Relevance	46,422	query, title, brand name, seller name, attribute,score

2.2 Existing Datasets

Datasets are essential for training and evaluating retrieval systems. General-purpose benchmarks such as MS MARCO and NQ (Nguyen et al., 2016; Kwiatkowski et al., 2019) mainly contain explicit question-style queries and lack user behavioral context, making them less suitable for personalized e-commerce search, where queries are often short and ambiguous, product texts are noisy, and user behaviors exhibit strong personalization and long-tail characteristics. Existing e-commerce datasets also have notable limitations. Amazon-based benchmarks (Van Gysel et al., 2016; Ai et al., 2017; Hou et al., 2026) often construct queries heuristically from product metadata and filter cold-start users or long-tail products, leading to distribution shifts from real-world search. JDsearch (Liu et al., 2023) provides real queries and user–product interactions while preserving long-tail data, but anonymizes textual content, limiting LLM-based research. Qilin (Chen et al., 2025b) retains plaintext content, yet focuses more on content retrieval and contains only short-term interaction data, making it insufficient for modeling complex and long-term e-commerce search behaviors.

3 The KuaiSearch Dataset

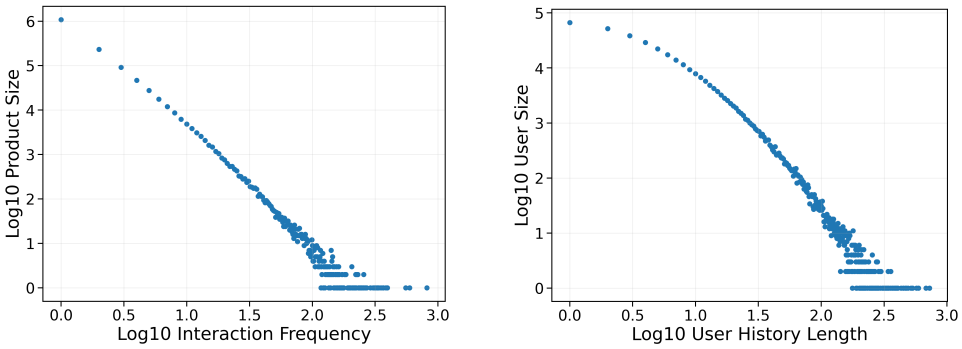
In this section, we introduce the KuaiSearch dataset, including its construction pipeline and data schema. KuaiSearch is built from real user search data collected from the e-commerce module of Kuaishou, covering diverse real-world search behaviors. The construction process consists of five components: user behavior collection, product metadata collection, user metadata collection, ranking data construction, and relevance data construction. Data from different components can be linked via shared session IDs for integrated analysis across components.

User Behavior Collection. We randomly sampled 331,930 users who exhibited core interaction behaviors during the two-month period starting on June 1, 2025. For each user, we collected all product search activities within the corresponding time window. For each search request, we retained the complete set of exposed items and recorded feedback signals, including clicks and purchases. Notably, we also retained search requests with no clicks or purchases to ensure the completeness of user behavior logs. During data construction, no constraints were imposed on the user population or the length of users' historical behaviors; as a result, the dataset naturally covers diverse user behavior patterns and more faithfully reflects real-world e-commerce search scenarios.

Product Metadata Collection. On the product side, we collect metadata, including titles, three-level category hierarchies, brand names, and seller names. All fields are retained in plaintext, enabling research on natural language understanding and semantic matching in e-commerce search.

User Metadata Collection. On the user side, we record basic demographic attributes, including gender, age group, and geographic location. These features provide auxiliary signals for modeling user heterogeneity and personalized search behavior.

Ranking Data Construction. Based on the complete user interaction logs, we further construct ranking data for each search request. Specifically, we augment the dataset with rich user-level and



(a) Product interaction frequency (b) User history search frequency

Figure 1: Log-log distributions of product interaction frequencies and searches per user.

product-level features. On the user side, these include the numbers of followees and followers, as well as clicks, exposures, purchases, and gross merchandise value (GMV) over the past 30 days. On the product side, we provide exposure count, click count, purchase count, and price. In addition, the ranking data explicitly record the search entrance to characterize differences in user behavior patterns and search intent across scenarios. Furthermore, we include the user's most recent 20 clicked products and most recent 20 purchased products as sequential behavior features, enabling more fine-grained modeling of preference evolution and intent dynamics.

Relevance Data Construction. To support research on relevance modeling, we construct a relevance dataset consisting of 46,422 query–product pairs collected from real user search scenarios. On the product side, we provide product titles, seller names, brand names, and attribute values. All query–product pairs are manually annotated by domain experts with e-commerce expertise using a four-level graded relevance scheme, thereby providing fine-grained supervision signals. Detailed annotation procedures and quality control measures are given in Appendix A.

KuaiSearch contains real-world e-commerce search interactions from 331,930 users. For privacy and compliance, identifiers are remapped and timestamps are represented using relative values (Appendix B). We also release KuaiSearch-Lite for efficient validation, ablation studies, and resource-constrained research (Appendix C). Table 1 compares both versions with existing datasets, while Table 2 and Section 4.2 summarize the data fields and relevance statistics, respectively.

4 Primary Data Analysis

This section presents a primary analysis of KuaiSearch, including (1) search interaction and content analysis over products, users, and queries, and (2) relevance data analysis.

4.1 Search Interaction and Content Analysis

In this section, we analyze e-commerce search characteristics in KuaiSearch from the perspectives of products, users, and queries, covering behavioral patterns, content properties, and distributions.

4.1.1 Product Analysis

First, we present the distribution of product interaction frequencies in the KuaiSearch dataset in Figure 1 (a). As shown, the interaction frequencies exhibit a pronounced long-tail distribution: a large number of products are interacted with only a few times by users (i.e., cold products), while only a small fraction of products receive frequent interactions. This pronounced long-tail distribution reflects the highly imbalanced popularity distribution of products in real-world e-commerce search scenarios. On the one hand, the vast majority of cold products constitute the main body of the product space, leading to severe data sparsity; on the other hand, a limited number of frequently interacted products concentrate most of the user behavior signals.

4.1.2 User History Analysis

Next, we analyze user behavior patterns. Figure 1 (b) shows the frequency distribution of the number of searches per user in the KuaiSearch dataset. We observe that the dataset includes users with very limited search activity, as well as users with extremely rich search histories. This distribution indicates substantial heterogeneity in user shopping behavior: some users conduct only a few product searches, whereas others exhibit frequent search activity. Consequently, this characteristic creates opportunities for designing differentiated personalization strategies tailored to users with diverse behavioral patterns. For example, when user histories are sparse, models may rely more heavily on general preferences learned from the entire user population; in contrast, for users with extensive search histories, greater emphasis can be placed on modeling individual interests. Such diversity also provides a realistic setting for evaluating the robustness of personalized search models under varying levels of behavioral sparsity.

To assess whether behavioral histories in KuaiSearch encode user-specific preference signals, we compare each user's recent interests against two references: their own earlier behavior and the aggregated earlier behavior of all eligible users. For users with more than ten interactions, the latest ten clicked or purchased items represent recent interests, while the remaining interactions characterize long-term preferences. We model interest distributions at the brand and first-level category levels and quantify their differences using Jensen–Shannon (JS) divergence. As shown in Figure 2, recent interests are consistently closer to users' own histories than to the population distribution, with a larger gap at the category level than at the brand level. The broad divergence distributions further indicate temporal variation in user preferences. These results demonstrate that KuaiSearch captures both stable personalized interests and evolving behavioral patterns.

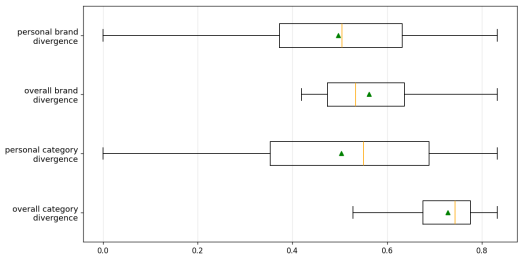


Figure 2: The boxplot of personal and overall interest divergence across users.

4.1.3 Query Analysis.

In this subsection, we provide an analysis of the search queries in KuaiSearch, focusing on query length distributions, user interaction patterns, and query intent ambiguity.

Figure 3 summarizes user engagement across different query length ranges. Medium-length queries dominate the distribution, with lengths of 5–6 and 7–8 together accounting for 44.68% of all requests. The session-level click-through rate (SessCTR) ranges from 0.496 to 0.531 across query lengths and does not exhibit a clear monotonic trend; nevertheless, very long queries (> 10) attain the highest SessCTR, suggesting a higher likelihood of triggering at least one click under more specific intent expressions. In contrast, the average number of clicked items per session (AvgClick) increases steadily with query length, rising from 1.06 to 1.33. This trend suggests that longer queries tend to express more specific user intents and lead to deeper interaction with the ranked results. Item-level click-through rate (ItemCTR) peaks at short-to-medium query lengths (3–4) and then gradually declines as query length increases, with a noticeable drop for very long queries (> 10). One possible explanation is that stronger intent encourages broader browsing and comparison behavior, dispersing clicks across a larger set of exposed items rather than concentrating them on a few.

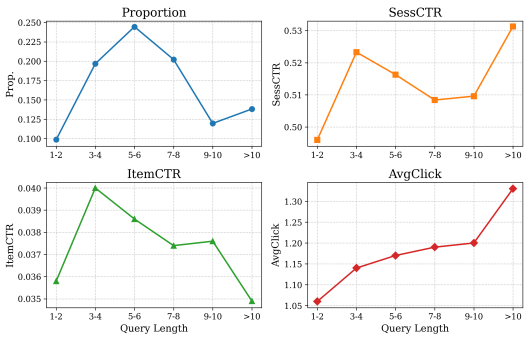
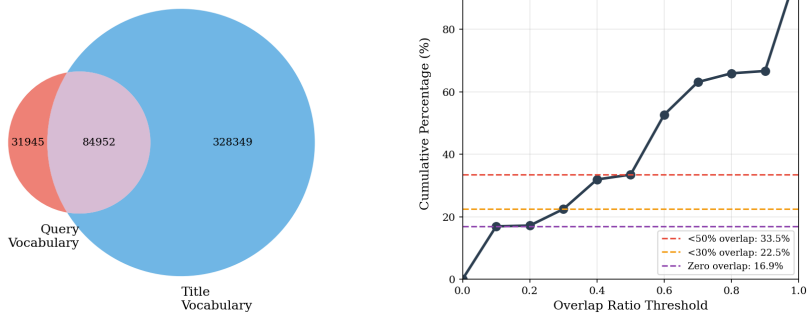


Figure 3: Engagement statistics across different query lengths.



(a) Vocabulary Overlap (b) Cumulative Distribution

Figure 5: Vocabulary asymmetry and token overlap between user queries and item titles.

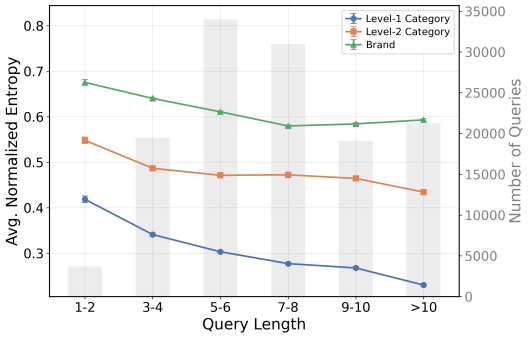


Figure 4: Query-level intent ambiguity across query lengths.

Beyond engagement patterns, we further examine query-level intent ambiguity through the category and brand distributions of positively interacted items, including both clicks and purchases. For each query, we aggregate all clicked and purchased products across sessions and deduplicate them to construct a unique interacted-item set. We then compute the normalized entropy of the first-level category, second-level category, and brand distributions within this set, where a higher entropy indicates that the query is associated with a more diverse range of potential product intents. As shown in Figure 4, category-level ambiguity generally decreases as query length increases. Specifically, the average first-level category entropy declines from 0.419 for queries of 1–2 characters to 0.230 for queries longer than 10 characters, while the corresponding second-level category entropy decreases from 0.548 to 0.434. These findings suggest that longer queries typically provide more explicit constraints, thereby narrowing the range of potentially relevant product categories. In contrast, brand-level entropy remains consistently high across all query-length groups, ranging from approximately 0.58 to 0.68. This pattern indicates that even relatively detailed queries often convey limited information about users' brand preferences. Overall, short queries exhibit particularly pronounced category-level ambiguity, whereas fine-grained brand preferences may require additional personalized signals derived from users' long-term historical interaction behaviors.

4.1.4 Vocabulary Asymmetry between Queries and Item Titles

This subsection analyzes the lexical gap between user queries and item titles in KuaiSearch. We use 3,035,445 clicked query–title pairs, tokenize texts with Jieba[1], and define token overlap as the proportion of query tokens appearing in the clicked item title. As shown in Figure 5, item titles contain 413,301 unique tokens, 3.54 times the query vocabulary of 116,897. Among them, 328,349 tokens appear only in titles, 10.3 times the 31,945 query-exclusive tokens. This asymmetry reflects different expression patterns: queries are concise and focus on categories or key attributes, while titles contain richer product information. At the instance level, 16.92% of clicked pairs have no token overlap and 33.48% have overlap below 50%. These results show that exact term matching alone is insufficient for many positive pairs. Thus, lexical methods such as BM25 (Robertson et al., 2009) may struggle with synonymy, implicit attributes, and cross-granularity matching, making KuaiSearch a challenging benchmark for semantic retrieval.

4.2 Relevance Data Analysis

[1] https://github.com/fxsjy/jieba

Table 4: KuaiSearch-Lite recall results; best and second-best values are bold and underlined.

	Lexical matching		*Embedding-based*				*Generative retrieval*		
Metric	**BM25**	**DocT5Query**	**DPR-SDE**	**DPR-ADE**	**BGE-base**	**BGE-base (fine-tuned)**	**DSI**	**LTRGR**	**LC-Rec**
R@10	0.071	0.078	<u>0.083</u>	0.082	0.074	**0.097**	0.062	0.069	0.071
HR@10	0.100	0.110	<u>0.121</u>	0.118	0.107	**0.138**	0.097	0.105	0.109
R@50	0.156	0.177	<u>0.208</u>	0.203	0.176	**0.229**	0.137	0.150	0.148
HR@50	0.209	0.238	<u>0.277</u>	0.271	0.235	**0.303**	0.202	0.218	0.218

Table 3: Distribution of relevance scores.

Score	**0**	**1**	**2**	**3**
Number	6,095	8,834	15,774	15,719
Proportion	13.13%	19.03%	33.98%	33.86%

The relevance data in KuaiSearch are annotated by e-commerce experts using a four-level graded relevance scheme. A score of 0 denotes clearly irrelevant items, typically due to category or attribute mismatches; 1 indicates weak relevance, where items belong to related categories but fail to satisfy the search intent; 2 represents partially relevant items that match the general category but violate important constraints such as brand or key attributes; and 3 indicates highly relevant items that fully satisfy the user's intent. This scheme provides fine-grained supervision for relevance modeling. Representative examples are shown in Table 10. As shown in Table 3, scores 2 and 3 account for most instances, indicating that KuaiSearch contains many non-trivial partially and highly relevant pairs. These cases are difficult to distinguish because they require reasoning over subtle category boundaries and attribute-level constraints. Overall, KuaiSearch provides a challenging benchmark for fine-grained relevance modeling in realistic e-commerce search.

5 Experiment

In this section, we report the performance of baseline approaches on recall, ranking, and relevance tasks. All experiments are conducted on our KuaiSearch-Lite benchmark, which improves experimental efficiency while preserving the representativeness of the evaluation. For trainable models, we report the average results over three runs using different seeds.

5.1 Recall Task Performance

In the recall task, we retain search requests with at least one click or purchase, using the last day of KuaiSearch-Lite recall logs for testing and the remaining logs for training. We evaluate Recall@K (R@K) and Hit Rate@K (HR@K), where R@K measures the proportion of interacted items retrieved in the top-K, and HR@K indicates whether any interacted item is retrieved. Baselines include lexical methods (BM25, DocT5Query (Robertson et al., 2009; Nogueira et al., 2019)), embedding-based methods (DPR-ADE, DPR-SDE, BGE-base-zh, and BGE-base-zh (fine-tuning) (Karpukhin et al., 2020; Dong et al., 2022; Xiao et al., 2024)), and generative methods (DSI, LTRGR, and LC-Rec (Tay et al., 2022; Li et al., 2024; Zheng et al., 2024)). DPR-ADE uses separate query and item encoders, while DPR-SDE and both BGE variants use shared encoders; BGE-base-zh (fine-tuning) is adapted on the training set. All methods use item titles. For DocT5Query, we train an mT5-base model (Xue et al., 2021) on the training set to generate 10 pseudo-queries per item, concatenated with item titles as indexed text. For generative retrieval, we use RQ-Kmeans (Luo et al., 2025; Deng et al., 2025) to assign semantic IDs with a three-layer codebook of size 512 per layer. For items sharing the same semantic ID, an additional deduplication position is appended to ensure uniqueness. Further implementation details are provided in Appendix D.

Table 4 reports the performance of different retrieval methods in terms of Recall and Hit Rate. Embedding-based retrieval methods outperform both lexical matching and generative retrieval methods across most evaluation metrics. Notably, BGE-base-zh after fine-tuning achieves the best performance across all metrics, demonstrating the effectiveness of adapting pretrained text representations to the e-commerce search domain. DPR-SDE also outperforms DPR-ADE, indicating its superior ability to model semantic relevance between queries and items. This observation is consistent with previous studies (Dong et al., 2022), which show that symmetric dual-encoder architectures with

Table 5: Ranking ROC-AUC; best and second-best values are bold and underlined.

Metric	DNN	Wide & Deep	DCN	DCN-v2	DIN	GDCN	WuKong	FinalMLP
ROC-AUC	<u>0.6258</u>	0.6217	0.6194	0.6239	**0.6262**	0.6224	0.6243	0.6220

shared parameters can alleviate misalignment in the embedding space, improving retrieval performance. In contrast, generative retrieval methods perform worse than embedding-based methods and achieve performance comparable to lexical matching methods. This gap mainly arises from the inherent complexity of e-commerce search: user queries are often ambiguous and multi-intent, product catalogs are vast with a long-tail distribution, and product texts are more noisy and unstructured compared with traditional retrieval settings. These factors make it particularly challenging for generative models to generate precise item identifiers during the recall phase.

5.2 Ranking Task Performance

In the ranking task, we use data from the last day of the KuaiSearch-Lite ranking dataset as the test set, while all preceding data are used for training. This chronological split reflects practical deployment scenarios, in which models are trained on historical interactions and evaluated on future user behavior. We evaluate model performance using ROC-AUC, which measures the ability to distinguish clicked items from non-clicked ones. We consider eight representative CTR prediction models: DNN (Covington et al., 2016), Wide & Deep (Cheng et al., 2016), DCN (Wang et al., 2017), DCN-v2 (Wang et al., 2021), DIN (Zhou et al., 2018), GDCN (Wang et al., 2023), WuKong (Zhang et al., 2024), and FinalMLP (Mao et al., 2023). These baselines cover diverse architectural paradigms, including multilayer perceptron-based models, explicit feature-crossing networks, attention-based user behavior modeling, and methods designed to capture high-order feature interactions. To ensure a fair comparison, all models use the same feature set, including user demographics such as gender and age, the 20 most recent clicked items and their categories, the target item and its category, and query and item-title embeddings extracted using a pretrained BGE-small-zh-v1.5 encoder (Xiao et al., 2024). The data split and preprocessing procedures are kept consistent across all baselines. Further implementation details are provided in Appendix D.

Table 5 reports the ROC-AUC performance of the evaluated ranking models. The results lie within a narrow range of 0.6194–0.6262, indicating broadly comparable performance under identical feature and training settings. Nevertheless, in large-scale e-commerce systems, even modest offline improvements may translate into meaningful online gains. DIN achieves the best ROC-AUC of 0.6262, suggesting that attention-based interest modeling effectively identifies relevant historical behaviors. DNN follows closely with 0.6258, showing that a simple multilayer perceptron remains competitive when rich user, item, behavioral, and textual features are available. Among feature-interaction models, DCN-v2 improves over DCN from 0.6194 to 0.6239, while WuKong reaches 0.6243 through explicit high-order interaction modeling. However, GDCN and FinalMLP do not outperform DNN or DIN, indicating that under the current experimental setting, more complex interaction modules do not necessarily yield better performance. Overall, the results highlight the importance of both behavioral preference modeling and input features in e-commerce search ranking.

5.3 Relevance Task Performance

In the relevance task, we split the dataset, using 10% of samples for testing and the rest for training. Samples annotated with a relevance score of 3 are treated as positive instances, while all others are regarded as negative, formulating relevance modeling as a binary classification problem. Model performance is evaluated via ROC-AUC and PR-AUC: ROC-AUC measures ranking quality, and PR-AUC better reflects the ability to identify positive instances under class-imbalanced settings.

We evaluate three representative and widely-used relevance modeling approaches: bi-encoder-based methods (BGE (Xiao et al., 2024)), cross-encoder-based methods (BERT-Chinese, BERT-Multilingual (Devlin et al., 2019), and XLM-RoBERTa (Conneau et al., 2020)), and LLM-based generative classification methods (Qwen3 (Yang et al., 2025) and Llama3.2 (Dubey et al., 2024)). For the LLM-based generative classification approach, each query-item pair is represented as a natural language input along with the task prompt, and the candidate labels (“yes” and “no”) are treated

Table 6: Relevance results; best and second-best values are bold and underlined.

	BI-ENCODER		CROSS-ENCODER				GENERATIVE			
Metric	**BGE Base**	**BGE Large**	**BERT Chinese**	**BERT Multi.**	**XLM-R Base**	**XLM-R Large**	**Qwen3 0.6B**	**Qwen3 1.7B**	**Llama3.2 1B**	**Llama3.2 3B**
Size	0.10B	0.32B	0.11B	0.11B	0.27B	0.55B	0.60B	1.70B	1.00B	3.00B
ROC-AUC	0.7475	0.7531	0.7606	0.7737	0.7941	0.8005	0.7994	**0.8215**	0.7602	<u>0.8093</u>
PR-AUC	0.5791	0.6052	0.6041	0.6383	0.6658	<u>0.6756</u>	0.6524	**0.6966**	0.5927	0.6696

as target tokens for LoRA fine-tuning (Hu et al., 2022). During inference, the model performs a single forward pass without autoregressive decoding, and the conditional probability distribution over the candidate label tokens is extracted at the final token position. The probabilities corresponding to "yes" and "no" are normalized via a softmax operation, and the probability of "yes" is used as the final relevance score for subsequent ranking and binary classification evaluation. Further implementation details are provided in Appendix D.

As shown in Table 6, bi-encoder-based methods exhibit weaker performance. This is primarily due to the fact that queries and items are encoded independently and matched via vector similarity, limiting the model's ability to capture fine-grained semantic interactions required for accurate relevance judgment. In contrast, cross-encoder-based methods achieve better performance, validating the effectiveness of explicitly modeling token-level interactions between queries and items. The performance of generative classification models varies across model families and scales. Among all evaluated models, Qwen3-1.7B achieves the best performance. In particular, Qwen3-1.7B achieves the highest scores on both ROC-AUC and PR-AUC, substantially outperforming the bi-encoder and cross-encoder baselines. Notably, Qwen3-0.6B surpasses Llama3.2-1B and achieves performance comparable to Llama3.2-3B, demonstrating the strong modeling capability and favorable parameter efficiency of the Qwen3 family for relevance modeling. Overall, these results suggest that models capable of capturing fine-grained semantic relationships between queries and items perform better in relevance modeling, and generative classification methods exhibit significant potential. A complementary ablation analysis on product metadata is provided in Appendix E.2.

6 CONCLUSION

We introduced KuaiSearch, a large-scale e-commerce search dataset constructed from authentic Kuaishou interactions. It preserves real queries, natural-language product texts, structured metadata, and diverse user behaviors without popularity-based filtering, while retaining long-tail products and cold-start users. KuaiSearch supports three key e-commerce search tasks: *recall*, *ranking*, and *relevance judgment*. Our analyses highlight query ambiguity, lexical gaps between queries and product titles, and heterogeneous user preferences, revealing practical challenges in real-world e-commerce search. Benchmark experiments further demonstrate its value for studying personalized retrieval, ranking, and fine-grained semantic relevance, while providing reference results for representative methods. We hope KuaiSearch serves as a useful resource for developing and evaluating more robust LLM-based e-commerce search systems.

7 LIMITATIONS

KuaiSearch reflects a Chinese e-commerce platform and contains linguistic, cultural, and domain-specific biases associated with its language, user population, product catalog, and shopping practices. Chinese expressions and tokenization may affect benchmark results, which should not be generalized to other languages or markets without validation. Translation may alter query ambiguity and does not reproduce authentic shopping behavior in the target market. Even after translation, differences in product availability, brand familiarity, and shopping conventions may limit the transferability of the observed findings. Future work should evaluate models using datasets from diverse languages, markets, and user populations to assess cross-domain generalizability and identify findings that remain consistent across settings.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and literature retrieval and discovery. Specifically, they were used to improve grammar, clarity, and phrasing, and to help identify potentially relevant prior work. All AI-assisted edits and literature suggestions were reviewed and verified by the authors. Generative AI tools were not used for research ideation or execution, drafting substantive sections of the paper, generating synthetic datasets, or proving mathematical claims. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

The construction and release of KuaiSearch follow Kuaishou Technology's data privacy and security governance procedures. All released data are de-identified, with identifiers remapped, demographic attributes aggregated, geographic information coarsened, and absolute timestamps converted to relative indices. The data release underwent internal legal and data-security compliance review. Further details are provided in Appendix B.

REFERENCES

Qingyao Ai and Lakshmi Narayanan. R. Model-agnostic vs. model-intrinsic interpretability for explainable product search. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 5–15, 2021.

Qingyao Ai, Yongfeng Zhang, Keping Bi, Xu Chen, and W Bruce Croft. Learning a hierarchical embedding model for personalized product search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 645–654, 2017.

Qingyao Ai, Daniel N Hill, SVN Vishwanathan, and W Bruce Croft. A zero attention model for personalized product search. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 379–388, 2019.

Ben Chen, Xian Guo, Siyuan Wang, Zihan Liang, Yue Lv, Yufei Ma, Xinlong Xiao, Bowen Xue, Xuxin Zhang, Ying Yang, et al. Onesearch: A preliminary exploration of the unified end-to-end generative framework for e-commerce search. *arXiv preprint arXiv:2509.03236*, 2025a.

Jia Chen, Qian Dong, Haitao Li, Xiaohui He, Yan Gao, Shaosheng Cao, Yi Wu, Ping Yang, Chen Xu, Yao Hu, et al. Qilin: A multimodal information retrieval dataset with app-level user sessions. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3670–3680, 2025b.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pp. 7–10, 2016.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8440–8451, 2020.

Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pp. 191–198, 2016.

Jose GC De Souza, Michael Kozielski, Prashant Mathur, Ernie Chang, Marco Guerini, Matteo Negri, Marco Turchi, and Evgeny Matusov. Generating e-commerce product titles and predicting their quality. In *Proceedings of the 11th international conference on natural language generation*, pp. 233–243, 2018.

Jiaxin Deng, Shiyao Wang, Kuo Cai, Lejian Ren, Qigen Hu, Weifeng Ding, Qiang Luo, and Guorui Zhou. Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment. *arXiv preprint arXiv:2502.18965*, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Zhe Dong, Jianmo Ni, Daniel M Bikel, Enrique Alfonseca, Yuan Wang, Chen Qu, and Imed Zitouni. Exploring dual encoder architectures for question answering. *arXiv preprint arXiv:2204.07120*, 2022.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Xian Guo, Ben Chen, Siyuan Wang, Ying Yang, Chenyi Lei, Yuqing Ding, and Han Li. Onesug: The unified end-to-end generative framework for e-commerce query suggestion. *arXiv preprint arXiv:2506.06913*, 2025.

Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation: Benchmarking llms as semantic encoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3251–3265, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pp. 6769–6781, 2020.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. Learning to rank in generative retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8716–8723, 2024.

Jiongnan Liu, Zhicheng Dou, Qiannan Zhu, and Ji-Rong Wen. A category-aware multi-interest model for personalized product search. In *Proceedings of the acm web conference 2022*, pp. 360–368, 2022.

Jiongnan Liu, Zhicheng Dou, Guoyu Tang, and Sulong Xu. Jdsearch: A personalized product search dataset with real queries and full interactions. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2945–2952, 2023.

Zhiding Liu, Ben Chen, Mingyue Cheng, Enhong Chen, Li Li, Chenyi Lei, Wenwu Ou, Han Li, and Kun Gai. Towards context-aware reasoning-enhanced generative searching in e-commerce. *arXiv preprint arXiv:2510.16925*, 2025.

Xinchen Luo, Jiangxia Cao, Tianyu Sun, Jinkai Yu, Rui Huang, Wei Yuan, Hezheng Lin, Yichen Zheng, Shiyao Wang, Qigen Hu, et al. Qarm: Quantitative alignment multi-modal recommendation at kuaishou. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 5915–5922, 2025.

Kelong Mao, Jieming Zhu, Liangcai Su, Guohao Cai, Yuru Li, and Zhenhua Dong. Finalmlp: an enhanced two-stream mlp model for ctr prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 4552–4560, 2023.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human-generated machine reading comprehension dataset. 2016.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*, 2019.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 3(4):333–389, 2009.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. Analysis of a very large web search engine query log. In *Acm sigir forum*, volume 33, pp. 6–12. ACM New York, NY, USA, 1999.

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

Manos Tsagkias, Tracy Holloway King, Surya Kallumadi, Vanessa Murdock, and Maarten De Rijke. Challenges and research opportunities in ecommerce search and recommendations. In *ACM Sigir Forum*, volume 54, pp. 1–23. ACM New York, NY, USA, 2021.

Christophe Van Gysel, Maarten de Rijke, and Evangelos Kanoulas. Learning latent vector spaces for product search. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pp. 165–174, 2016.

Fangye Wang, Hansu Gu, Dongsheng Li, Tun Lu, Peng Zhang, and Ning Gu. Towards deeper, lighter and interpretable cross network for ctr prediction. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pp. 2523–2533, 2023.

Jingang Wang, Junfeng Tian, Long Qiu, Sheng Li, Jun Lang, Luo Si, and Man Lan. A multi-task learning approach for improving product title compression with user search log data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD'17*, pp. 1–7. 2017.

Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the web conference 2021*, pp. 1785–1797, 2021.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 641–649, 2024.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*, 2020.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pp. 483–498, 2021.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yuki Yano, Yukihiro Tagami, and Akira Tajima. Quantifying query ambiguity with topic distributions. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 1877–1880, 2016.

Buyun Zhang, Liang Luo, Yuxin Chen, Jade Nie, Xi Liu, Daifeng Guo, Yanli Zhao, Shen Li, Yuchen Hao, Yantao Yao, et al. Wukong: Towards a scaling law for large-scale recommendation. *arXiv preprint arXiv:2403.02545*, 2024.

Xue Zhang, Fusen Guo, Tao Chen, Lei Pan, Gleb Beliakov, and Jianzhang Wu. A brief survey of machine learning and deep learning techniques for e-commerce research. *Journal of Theoretical and Applied Electronic Commerce Research*, 18(4):2188–2216, 2023.

Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. Adapting large language models by integrating collaborative semantics for recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 1435–1448. IEEE, 2024.

Zexin Zheng, Huangyu Dai, Lingtao Mao, Xinyu Sun, Zihan Liang, Ben Chen, Yuqing Ding, Chenyi Lei, Wenwu Ou, Han Li, et al. Onevision: An end-to-end generative framework for multi-view e-commerce vision search. *arXiv preprint arXiv:2510.05759*, 2025.

Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1059–1068, 2018.

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

A Relevance Annotation Quality and Reliability

The quality and reliability of the relevance annotations are fundamental to the validity of the model evaluation results reported in this work. To ensure the accuracy and consistency of the annotated data, we designed and implemented a rigorous annotation procedure and quality control mechanism. All query–item pairs were independently annotated by three annotators, who had no access to the decisions made by the other annotators. The annotators had frontline industrial experience in e-commerce search algorithms and received comprehensive and systematic training based on predefined annotation guidelines before the formal annotation process. In addition, annotators were required to provide an explicit written justification for each assigned label to support subsequent quality control. This mandatory requirement encouraged annotators to make decisions according to predefined criteria rather than arbitrary subjective judgments. Before applying majority voting, we first measured inter-annotator agreement on the raw annotations. The results show that the three annotators reached agreement on 93% of the relevance labels, with a Fleiss' κ score of 0.89. For samples with annotation disagreements, we adopted a tiered resolution strategy. General disagreements were resolved through majority voting, whereas samples with substantial disagreements first underwent structured group discussions, during which the annotators re-examined the annotation guidelines and specific cases, followed by a second round of independent re-annotation. These annotation procedures and quality control measures effectively reduced inconsistency and arbitrariness in the annotations, providing a reliable foundation for evaluating relevance models.

B Data Privacy and Ethical Compliance

The development and release of the KuaiSearch dataset follow Kuaishou Technology's data privacy governance framework. All data collection, processing, and release procedures underwent internal legal and data security compliance reviews and were conducted in accordance with the relevant requirements of the Personal Information Protection Law of the People's Republic of China. The compliance review further confirmed that the relevant user authorization is covered by the applicable platform privacy policy. Before public release, we applied multiple de-identification measures to protect user privacy. Specifically, all user, item, seller, and other related identifiers were replaced with irreversible random identifiers. Age information was aggregated into seven predefined age groups, geographic information was retained only at the provincial level for users, and all absolute timestamps were converted into relative time indices. The released dataset contains no direct personal identifiers and cannot be linked back to original platform accounts. These de-identification procedures, together with the internal compliance review process, substantially reduce potential re-identification risks.

C CONSTRUCTION OF KUAISEARCH-LITE

To facilitate efficient experimentation while preserving realistic user behavior patterns, we construct KuaiSearch-Lite by randomly selecting 102,086 users from the full KuaiSearch dataset and retaining their complete search and interaction records within a continuous one-month period. The subset is constructed at the user level rather than by independently sampling individual search requests or interaction instances. This design preserves the temporal continuity of user behavior histories and maintains the associations among queries, exposed items, clicks, and purchases. For each selected user, we retain the corresponding user information, search requests, exposed items, clicked and purchased items, and associated product metadata to construct the recall and ranking data. Consequently, KuaiSearch-Lite preserves the main behavioral and structural characteristics of the full dataset while substantially reducing the computational cost of model training and evaluation. The resulting subset contains 555,553 queries and 6,634,118 products, providing an efficient and representative benchmark for model development, ablation studies, and reproducible evaluation.

D EXPERIMENTAL DETAILS

This section reports the key implementation details of the recall, ranking, and relevance experiments. More comprehensive model-specific configurations, preprocessing procedures, and training and evaluation scripts are available in our released codebase.

D.1 RECALL EXPERIMENTAL DETAILS

For lexical retrieval, BM25 is implemented with Jieba-based Chinese word segmentation, and the hyperparameters are set to $k_1 = 1.2$ and $b = 0.75$. For dense retrieval methods, the maximum sequence lengths of both queries and item titles are set to 64 tokens. Query–item relevance is measured by inner-product similarity, and all item representations are indexed using `Faiss IndexFlatIP`, enabling exact maximum-inner-product search over the candidate corpus. For generative retrieval methods, we adopt prefix-tree-constrained decoding to restrict the generation space to valid semantic item identifiers. DSI and LTRGR are implemented with mT5-base as the backbone, whereas LC-Rec adopts Qwen3-0.6B-Instruct.

D.2 RANKING EXPERIMENTAL DETAILS

All ranking baselines use the same feature and training configurations. The embedding dimensions of user IDs and item IDs are both set to 32, while those of gender, age group, and category levels 1–3 are set to 8, 16, 16, 32, and 64, respectively. User behavior sequences retain the 20 most recent clicked items. All models are optimized using Adam with an initial learning rate of 1×10^{-3}, a weight decay of 1×10^{-5}, and binary cross-entropy loss. The batch size is 20,000, and the maximum number of training epochs is 20.

D.3 RELEVANCE EXPERIMENTAL DETAILS

For all relevance models, the maximum input sequence length is set to 512 tokens; for bi-encoders, this limit is applied separately to the query and item inputs. Bi-encoder models encode queries and items independently and compute relevance scores using cosine similarity between their representations. Cross-encoder models concatenate and jointly encode query and item texts to capture fine-grained token-level interactions. Qwen3-0.6B, Qwen3-1.7B, Llama3.2-1B, and Llama3.2-3B are fine-tuned using LoRA with rank $r = 8$, scaling factor $\alpha = 32$, and dropout rate 0.1. LoRA adapters are applied to the major linear projection modules, including `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`.

E Ablation Studies

E.1 Ranking

We conduct ablation experiments on feature groups for the ranking task using the best-performing DIN model to quantify the contribution of different signal types. Table 7 reports the ROC-AUC performance when removing each feature group individually while keeping all others intact.

Table 7: Ablation study on feature groups for ranking (DIN).

Setting	**ROC-AUC**
DIN	0.6262
w/o text	0.6078
w/o history	0.6206
w/o demographics	0.6197
w/o category	0.6249

The results show that textual features provide the most important signal for ranking, as their removal causes the largest ROC-AUC drop from 0.6262 to 0.6078. This finding highlights the central role of query–item semantic matching in e-commerce ranking, as textual representations directly capture the alignment between user queries and product descriptions. Demographic and historical behavior features also contribute substantially, with their removal reducing ROC-AUC to 0.6197 and 0.6206, respectively. These results suggest that both group-level user characteristics and sequential behavioral preferences provide complementary signals for personalized ranking. Category features yield a smaller but still positive contribution, with ROC-AUC decreasing to 0.6249 when they are removed. Overall, these findings demonstrate the value of the heterogeneous multi-field features provided by KuaiSearch for developing and evaluating ranking models.

E.2 Relevance

We conduct ablation experiments on the product metadata inputs for the relevance task using the parameter-efficient Qwen3-0.6B model. Table 8 reports the PR-AUC performance when progressively adding different metadata fields, starting from the basic product title (T) and incrementally including brand (B), seller name (S), and attribute (A) information.

Table 8: Ablation study on product metadata for relevance (Qwen3-0.6B).

Setting	T	T+B	T+B+S	T+B+S+A
PR-AUC	0.6383	0.6414	0.6432	0.6524

The results show that relevance performance consistently improves as richer product metadata is introduced, with attribute information bringing the largest incremental gain of 0.0092 PR-AUC. This is because attribute fields contain fine-grained product specifications such as size, material, style, and functional features, which are critical for distinguishing between products that share similar titles but differ in key characteristics that directly match user intent.

F Long-tail Scenario Analysis

We further conduct a long-tail analysis on the ranking task using the best-performing DIN to evaluate performance across different popularity segments. We define head items as the top 20% most frequently interacted products, while the remaining 80% of products are considered tail items. Similarly, head users are defined as the top 20% most active users with the most search sessions, while the remaining users are regarded as tail users.

These results reveal a substantial performance gap between head and tail segments. The model achieves substantially higher ROC-AUC on head items and head users, which can be attributed to

Table 9: DIN model performance on head and tail segments

Group	Head	Tail
Item ROC-AUC	0.6606	0.6180
User ROC-AUC	0.6473	0.6189

Table 10: Representative query–item pairs across relevance scores in KuaiSearch. Scores range from 0 (irrelevant) to 3 (highly relevant).

Score	Query	Item Title	Brand	Seller	Key Attributes
3	眼影盘橘朵 (Judydoll eyeshadow)	Judydoll橘朵单色眼影哑光日常新手淡妆闪片牛郎 (Judydoll single eyeshadow matte daily nude glitter)	橘朵 (Judydoll)	俏俏美妆优选 (Qiaoqiao Beauty Selection)	易上色, 眼影盘, 珠光, 单色 (easy color, eyeshadow palette, pearl, single color)
2	帽子棉女童 (Cotton girl's hat)	爱贝迪拉婴儿帽子棉帽子四季款新生婴幼保暖胎帽护囟门帽宝宝可爱 (Aibedila baby cotton hat, four seasons, keep warm, protect fontanel, cute)	爱贝迪拉 (Aibedila)	爱贝迪拉母婴旗舰店 (Aibedila Maternity & Baby Flagship Store)	胎帽, 棉, 翻边 (fetal hat, cotton, turn-up edge)
1	日奈绘画 (Rina painting)	节奏盒子描摹本画画本儿童线稿手绘描摹控笔练习绘画本图画本涂色 (Rhythm box tracing book, children's line drawing, hand tracing, pen control practice, coloring book)	艺丛 (Yicong)	彦和文具专营店 (Yanhe Stationery Store)	通用, 8岁–14岁, 艺丛 (universal, ages 8–14, Yicong)
0	防熊三件套 (Bear-proof three-piece set)	细细条冬天毛绒可爱小熊帽子围巾手套一体三件套骑车防寒护耳帽 (Xixitiao winter plush cute bear hat scarf gloves set, cycling cold-proof earflap hat)	细细条 (Xixitiao)	细细条旗舰店 (Xixitiao Flagship Store)	毛绒, 儿童, 细细条 (plush, children, Xixitiao)

the abundance of interaction data available for these segments, enabling more reliable statistical feature estimation and more accurate preference modeling. In contrast, tail items suffer from severe data sparsity and noisy statistical features, while tail users have limited historical behavior signals, making personalized ranking particularly challenging. These findings highlight that long-tail ranking remains a critical challenge in e-commerce search.

G Data License and Intended Use

KuaiSearch is released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. It is intended to support research and development in e-commerce retrieval, ranking, relevance modeling, personalization, and related areas. Users should provide appropriate attribution and comply with applicable laws and regulations.

H METRIC DEFINITIONS

To ensure scientific reproducibility, we provide formal mathematical definitions for the engagement, retrieval and ranking, personalization, and query analysis metrics used in this paper.

H.1 ENGAGEMENT METRICS

- **Session CTR (SessCTR)**: Proportion of search sessions with at least one click:

$$\mathrm{SessCTR} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathbb{I}\left(|\mathcal{C}_s| > 0\right),$$

  where $\mathcal{S}$ is the set of search sessions, $\mathcal{C}_s$ is the clicked-item set for session s, and $\mathbb{I}(\cdot)$ is the indicator function.

- **Item CTR**: Proportion of exposed items that are clicked:

$$\mathrm{ItemCTR} = \frac{\sum_{s \in \mathcal{S}} |\mathcal{C}_s|}{\sum_{s \in \mathcal{S}} |\mathcal{E}_s|},$$

  where $\mathcal{E}_s$ is the exposed-item set for session s.

- **AvgClick**: Average number of clicked items per search session:

$$\mathrm{AvgClick} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} |\mathcal{C}_s|.$$

H.2 RETRIEVAL & RANKING METRICS

- **Recall@K (R@K)**: Measures the average proportion of positively interacted items retrieved in the top-K results:

$$\mathrm{R@K} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \frac{|\mathrm{TopK}(s) \cap \mathcal{I}_s|}{|\mathcal{I}_s|},$$

  where $\mathcal{S}$ is the set of evaluated search requests, $\mathrm{TopK}(s)$ is the retrieved top-K item set, and $\mathcal{I}_s$ is the set of clicked or purchased items for request s.

- **Hit Rate@K (HR@K)**: Measures the proportion of search requests with at least one positively interacted item in the top-K results:

$$\mathrm{HR@K} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathbb{I}\left(|\mathrm{TopK}(s) \cap \mathcal{I}_s| > 0\right),$$

  where $\mathbb{I}(\cdot)$ is the indicator function.

- **ROC-AUC**: Measures the model's ability to distinguish positive from negative samples across classification thresholds.

- **PR-AUC**: Measures performance across precision–recall trade-offs and is particularly informative under class imbalance.

H.3 PERSONALIZATION AND QUERY ANALYSIS METRICS

- **Jensen–Shannon (JS) Divergence**: Measures the difference between two interest distributions P and Q:

$$\mathrm{JS}(P\|Q) = \frac{1}{2} D_{\mathrm{KL}}(P\|M) + \frac{1}{2} D_{\mathrm{KL}}(Q\|M),$$

  where $M = \frac{1}{2}(P + Q)$ and $D_{\mathrm{KL}}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence. A larger value indicates greater distributional discrepancy.

- **Normalized Entropy**: Measures query-level intent ambiguity based on the attribute distribution of interacted items:

$$\widetilde{H}(q) = -\frac{\sum_{i=1}^{N} p_i(q) \log p_i(q)}{\log N},$$

where $p_i(q)$ is the proportion of interacted items associated with the i-th attribute value, and N is the number of values in the attribute space. A higher value indicates greater category or brand diversity and stronger intent ambiguity.

- **Token Overlap Ratio**: Measures lexical overlap between a query and its corresponding item title:

$$\mathrm{Overlap}(q, t) = \frac{|T_q \cap T_t|}{|T_q|},$$

where T_q and T_t denote the query and item-title token sets, respectively. A lower value indicates a larger lexical gap.