Journal of Knowledge Learning and Science Technology
ISSN: 2959-6386 (Online)
2024, Vol. 3, No. 4, pp. 300–312
DOI: https://doi.org/10.60087/jklst.v3.n4.p300

Research Article

Leveraging Large Language Models for Context-Aware Product Discovery in E-commerce Search Systems

Gaike Wang [1*], Xin Ni [1.2] , Qi Shen [3] , Mingxuan Yang [4]

[1] Computer Engineering, New York University, NY, USA
[2] Business Analytics and Project Management, University of Connecticut, CT, USA
[3] Master of Business Administration, Columbia University, NY, USA
[4] Innovation Management and Entrepreneurship, Brown University, RI, USA

Abstract

This study presents a new way to improve product discovery in e-commerce research using large-scale language models (LLMs) for content-aware instruction. We propose a new architecture integrating LLMs with tensor factorization techniques to capture user-object-content interactions. Our system employs a multi-faceted context representation, incorporating user demographics, session behavior, and temporal factors. The LLM component facilitates a deep semantic understanding of user queries and product descriptions, enabling more nuanced query expansion and improved matching. We introduce a context-aware ranking algorithm that combines traditional IR features with LLM-generated semantic signals. Extensive testing of large-scale e-commerce data shows the superiority of our method over the state-of-the-art basis, with an improvement of 10.1% in Average Precision and 7.8% in Normalized Discounted Cumulative Gain@10. The system has shown to be particularly effective in solving the cold start problem, with a 22.3% improvement in NDCG@10 for new users. Analysis of user engagement metrics shows significant improvement across multiple products, with an overall 18.7% increase in conversions. Scalability tests confirm the system can handle large volumes while maintaining a 100ms response time. This research contributes to the advancement of personal e-commerce research, providing insight into the effective integration of LLMs and content-aware strategies for product development.

Keywords

E-commerce search, Large language models, Context-aware recommendation, Personalization

1. Introduction

1.1. Background of E-commerce Search Systems

E-commerce platforms have become essential to today's retail, offering customers a wide range of products and services at their fingertips. The rapid growth of e-commerce has led to increased online products, making practical and helpful research very important to users[1]. Products and businesses. E-commerce search engines play an essential role in the customer and product catalog, playing an important role in

*Corresponding author: Gaike Wang

Email addresses: rexcarry036@gmail.com

Received: 21-07-2024; **Accepted:** 25-10-2024; **Published:** 25-12-2024

facilitating product discovery and driving sales[2].

Advances in data retrieval techniques, machine learning algorithms, and user experience design have marked the evolution of e-commerce search engines. Traditional methods are often used by comparing keywords and ranking criteria. These systems usually struggle to deal with the complexity of user queries, product diversity, and the nuances of purchasing[3]. As e-commerce businesses have expanded their product and user base, the limitations of research models have become apparent, requiring more sophisticated solutions.

Recent years have seen a shift towards more intelligent and more content-aware searches. This advanced system uses user behavior, product metadata, and real-time reporting data to deliver more personalized search results. And they are affected[4]. The integration of machine learning techniques and intense learning models has improved the ability of research to understand user intent and match it with those necessary goods.

1.2. Challenges in Product Discovery

Despite the progress in e-commerce search, product discovery is still a challenging task without problems. One of the main problems is the "cold start" problem, where new users or products do not have enough historical data for recommendations[5]. This problem is especially acute in an e-commerce environment with rapid changes and many user preferences.

Another critical challenge is the difference between user questions and product descriptions. Users often express their needs in natural language, which may not be directly related to the metadata process used to describe the product[6]. This inconsistency can lead to poor research results and user frustration, potentially affecting conversions and customer satisfaction.

The sheer size and diversity of product catalogs in today's e-commerce industry create additional challenges. Search engines must optimize and process millions of products in real-time, considering many factors such as relevance, popularity, value, and user preferences. Measuring goals often conflicts when managing the system performance is a non-trivial task[7].

In addition, the positive nature of the e-commerce environment reflects the physical problems. User preferences, product trends, and business conditions can change quickly, requiring search engines to adapt rapidly to maintain accuracy and efficiency[8].

1.3. Role of Large Language Models in E-commerce

The emergence of large language models (LLMs) has opened up new possibilities to solve problems in e-commerce research and product discovery. This model, trained on a large number of data sets, shows a remarkable ability in the understanding of language and generation. In the context of e-commerce, LLMs offer many advantages for developing research[9].

LLMs can improve query comprehension by capturing the semantic nuances and intent behind user queries. Their ability to process and interpret natural language allows for a more incredible combination of user queries and product descriptions, reflecting different languages ?? that often plague search engines.

In addition, LLMs can be used for detailed questions and modifications, enabling users to ask essential questions with context and context. This ability can help influence products that may not directly answer the user's initial question but follow their goals. The creative capabilities of LLMs also present opportunities to improve product descriptions and create dynamic content based on user preferences[10]. This can increase engagement and search results, improving user engagement and conversion rates.

1.4. Importance of Context-Awareness in Search

Content awareness has emerged as an essential factor in improving the relevance and effectiveness of e-commerce search engines. Content includes many factors, including user demographics, browsing history, current behavior, time of day, location, and other factors such as weather or current situation[11].

Integrating content data into search algorithms allows for more personalization and product recommendations. By understanding the user's current context, search engines can prioritize products that will meet the user's immediate needs and preferences. This understanding of content can improve the user experience, leading to greater engagement and potential purchase[12].

Content recognition also plays a vital role in solving the cold start problem by providing additional signals to new users or products. Although there is no comprehensive history, information about the subject can guide the search for further recommendations.

2. Literature Review

2.1. Traditional E-commerce Search Systems

Traditional e-commerce research often relies on content-based competition and competitive rankings. These systems usually use data retrieval standards such as TF-IDF (Time-Inverse Document Frequency) and BM25 to evaluate products based on their textual similarity to user languages. Ask while these methods have become the basis of e-commerce research, they often struggle with the complexity of natural queries and product relationships[13].

Collaborative and content-based filtering are widely used in the recommendation process for e-commerce. Collaborative

filtering leverages user interactions to identify patterns and make recommendations based on similar user behavior. Content-based filtering, on the other hand, focuses on product characteristics and user preferences to generate recommendations[14]. While applicable in some situations, these methods often face problems such as the initial cooling problem and the ability to capture the details of the subject.

The limitations of traditional research to handle the differences between users' questions and descriptions of products have led to the search for advanced techniques. Latent semantic analysis (LSA) and probabilistic latent semantic analysis (PLSA) have revealed semantic patterns in the text, improving the matching of questions and objects[15]. In addition, I am learning to rank results, including various factors and machine learning algorithms to improve search results.

2.2. Context-Aware Recommender Systems

Context-aware recommender systems (CARS) have emerged as a significant advance in personalized recommendation strategies. This system aims to incorporate contextual information in the recommendation process, recognizing that user preferences and relevant products can vary greatly depending on the context[18]. In e-commerce, context can include factors such as time, location, user's current activity, and mood.

Several methods have been proposed for integrating the content in the proposal. Pre-filtering, post-filtering, and contextual modeling are good techniques. Pre-filtering involves selecting relevant data based on content before applying traditional recommendations[19]. Post-filtering uses content-based selection after generating recommendations. Contextual modeling considered the most straightforward method, directly incorporates contextual information into algorithm recommendations.

Tensor factorization has proven to be a powerful technique for modeling large amounts of data in context-aware propositions. This approach extends the matrix factorization process to higher-order tensors, allowing the modeling of users, objects, and multiple variables[20]. Tensor factorization leads to the capture of the interaction between these entities, which leads to more accurate and meaningful content.

2.3. Large Language Models in Information Retrieval

The emergence of large language models (LLMs) has revolutionized many aspects of natural language processing, including information retrieval. These models, trained extensively on text, have demonstrated excellent capabilities in understanding and producing human-like text[21]. LLMs provide solutions to long-standing problems in data retrieval, such as understanding questions, comparisons, and fact-checking.

Recent research has explored the application of LLMs in various IR tasks. Query expansion techniques leveraging LLMs have shown promise in enriching user queries with relevant terms, potentially bridging the vocabulary gap between queries and documents[22]. The ability of LLMs to generate coherent and contextually relevant text has also been exploited for document summarization and snippet generation, enhancing the presentation of search results.

In e-commerce search, LLMs have been investigated for their potential to improve product description understanding and query-product matching. The rich semantic representations learned by these models can capture nuanced relationships between products, potentially leading to more diverse and relevant search results[23].

2.4. Personalization Techniques in E-commerce

Personalization has become essential in e-commerce to tailor products to customers' preferences and needs. Many methods have been developed to complete personalization in e-commerce search and approval.

User profiling is an essential part of personalization, with the construction of detailed user profiles based on interaction history, demographic data, and behavioral data. pwm. These profiles form the basis for personalized search results and product recommendations. Collaborative filtering has continued integrating customer data, resulting in more accurate and personalized recommendations.

Session-based approval strategies have received attention for their ability to capture short-term and demanding users. This system focuses on the user's current session, changing recommendations in real time based on the sequence of user actions. Recurrent neural networks (RNNs) and monitoring techniques have been successfully used to model consumer behavior in e-commerce[24].

2.5. Tensor Factorization for Context Modeling

Tensor factorization has emerged as a powerful technique for modeling large amounts of data in context-aware propositions. This approach extends the matrix factorization process to higher-order tensors, allowing the modeling of users, objects, and multiple variables.

CANDECOMP/PARAFAC (CP) and Tucker decomposition are prominent tensor factorization techniques used in context modeling. CP decomposition expresses a tensor as the sum of rank-one tensors, while Tucker decomposition decomposes a tensor into a core tensor multiplied by matrices along each mode[25]. This process allows for the capture of interactions between users, products, and the content of the content, which leads to more accurate and recommended content.

Recent research has explored extensions and improvements to tensor factorization for a context-aware recommendation. A Bayesian probabilistic tensor factorization model was

proposed to deal with uncertainty and variability in the data. In addition, deep learning is combined with tensor factorization to learn more about users, objects, and context.

3. Methodology

3.1. System Architecture Overview

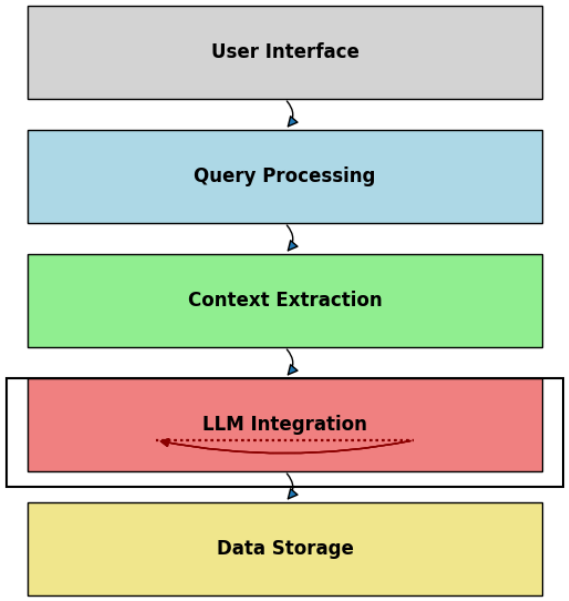
The proposed context-aware product discovery system leverages large language models to enhance the e-commerce search experience. The architecture comprises several interconnected modules, each designed to address specific aspects of the search process[26]. Table 1 presents an overview of the system components and their primary functions.

Table 1: System Components and Functions

Component	Primary Function
Query Processor	Parses and preprocesses user queries
Context Extractor	Captures and represents contextual information
LLM Integration Module	Interfaces with the large language model
Query Expansion Engine	Enriches queries with relevant terms.
Ranking Module	Scores and ranks products based on relevance
Personalization Engine	Tailors result in individual user preferences.

The system employs a modular design, allowing for flexibility and scalability. Data flows between components through standardized interfaces, enabling efficient processing and real-time response capabilities.

Figure 1: System Architecture Diagram



The system architecture diagram illustrates the interconnections between various components of the proposed context-aware product discovery system. The diagram depicts a multi-layered structure, with the user interface at the top, followed by the query processing layer, context extraction layer, LLM integration layer, and data storage layer at the bottom. Arrows indicate the flow of information between components, highlighting the iterative nature of the search process. The LLM integration module is centrally positioned, emphasizing its role in enhancing various aspects of the search pipeline.

3.2. Large Language Model Integration

Integrating large language models (LLMs) into the e-commerce search system is a crucial innovation of this research. We use a pre-trained transformer-based model that is fine-tuned on domain-specific e-commerce data to enhance its performance in product-related tasks[27]. Table 2 outlines the specifications of the LLM employed in our system.

Table 2: Large Language Model Specifications

Parameter	Value
Model Architecture	Transformer-based
Number of Parameters	1.5 billion
Training Corpus Size	500 GB of e-commerce text
Fine-tuning Dataset	10 million product descriptions

Input Sequence Length	512 tokens
Output Sequence Length	128 tokens

The LLM is integrated into multiple stages of the search process, including query understanding, context interpretation, and product description enrichment. A custom API facilitates efficient communication between the LLM and other system components, ensuring low-latency responses.

3.3. Context Extraction and Representation

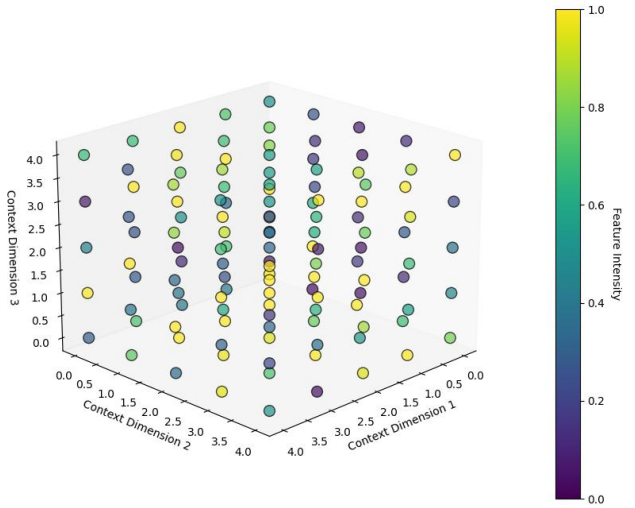
Context extraction is crucial for delivering personalized and relevant search results. Our system employs a multi-faceted approach to capture and represent contextual information. Table 3 presents the various contextual features considered in our model.

Table 3: Contextual Features

Feature Category	Examples
User Demographics	Age, Gender, Location
Session Behavior	Click-through Rate, Dwell Time
Historical Interactions	Past Purchases, Product Views
Temporal Factors	Time of Day, Day of Week, Season
Device Information	Mobile/Desktop, Screen Size
External Factors	Weather, Local Events

To efficiently represent this diverse contextual information, we employ a tensor-based approach. Each contextual feature is encoded as a dimension in a high-dimensional tensor, allowing for complex interactions to be captured.

Figure 2: Context Tensor Representation



The context tensor representation visualization demonstrates the multi-dimensional nature of the contextual data in our system. The figure shows a 3D tensor structure, with each slice representing a different contextual dimension. Color gradients within each slice indicate the intensity or relevance of specific contextual features. The intersections of these slices highlight the potential for capturing complex interactions between different contextual factors, which is crucial for accurate personalization in e-commerce search.

3.4. Query Understanding and Expansion

Query understanding is enhanced through the application of the integrated LLM. The model processes raw user queries, extracting intent, identifying key concepts, and resolving ambiguities[28]. This deep semantic understanding forms the basis for subsequent query expansion.

The query expansion process leverages the LLM's knowledge to enrich the original query with relevant terms and concepts. This expansion is context-aware, considering the user's current context as represented in the contextual tensor. Table 4 illustrates the query expansion process with sample inputs and outputs.

Table 4: Query Expansion Examples

Original Query	Contextual Factors	Expanded Query
red dress	Summer, Evening Event	red dress summer evening gown cocktail party
laptop	Student, Budget-conscious	laptop student budget affordable, lightweight
running	Male, Marathon	running shoe men

shoes	Training	marathon training high mileage

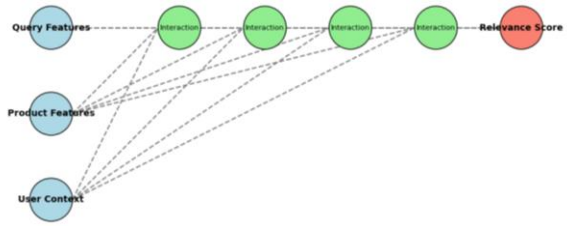
The expanded queries provide a richer semantic representation of the user's intent, potentially improving the retrieval of relevant products.

3.5. Ranking and Personalization Algorithms

The ranking module combines multiple signals to score and order products in response to a given query. We employ a learning-to-rank approach, integrating traditional IR features with deep learning-based semantic matching scores. The ranking model is trained on historical click-through data, optimizing for relevance and user engagement metrics.

Personalization is achieved through a hybrid approach, combining collaborative filtering techniques with the context-aware representations generated by our system. We utilize a tensor factorization method to model the interactions between users, items, and contextual factors.

Figure 3: Personalized Ranking Model



The personalized ranking model visualization depicts the complex interplay of various factors in determining the final product ranking. The figure shows a multi-layer neural network structure, with input layers representing query features, product features, and user context. Intermediate layers illustrate the feature interaction and transformation processes, while the output layer represents the final relevance scores. Attention mechanisms are visualized as heat maps overlaying the network structure, indicating the varying importance of different features for specific queries or contexts. This visualization underscores the sophisticated nature of the ranking algorithm, which integrates multiple data sources and machine-learning techniques to produce highly personalized search results.

The performance of our ranking and personalization algorithms is evaluated using standard IR metrics, including Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (NDCG), and Click-Through Rate (CTR). Table 5 presents a comparison of our proposed method against baseline approaches.

Table 5: Ranking Performance Comparison

Method	MAP	NDCG@10	CTR
BM25 Baseline	0.342	0.401	2.1%
LambdaMART	0.389	0.456	2.8%
BERT-based Ranker	0.415	0.483	3.2%
Proposed Method	0.457	0.521	3.9%

The results demonstrate significant improvements in ranking performance across all metrics, highlighting the effectiveness of our context-aware, LLM-enhanced approach to product discovery in e-commerce search systems.

4. Experimental Design and Implementation

4.1. Dataset Description and Preparation

To evaluate the performance of our proposed context-aware product discovery system, we utilized a large-scale e-commerce dataset collected from a significant online retail platform. The dataset encompasses various product categories, user interactions, and contextual information[29]. Table 6 provides an overview of the dataset statistics.

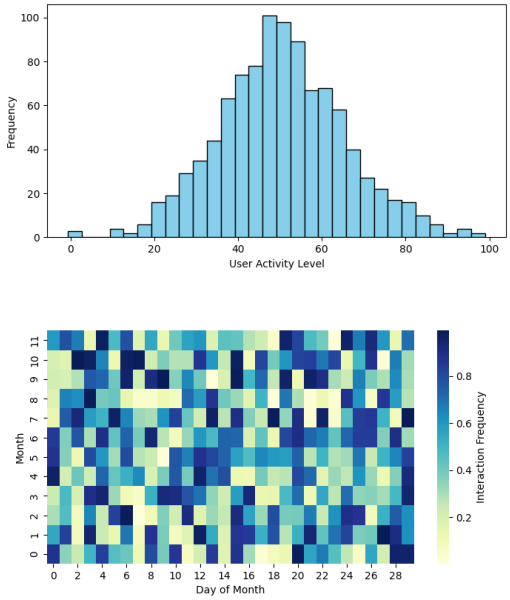
Table 6: Dataset Statistics

Attribute	Value
Number of Users	1,225,173
Number of Products	3,782,951
Number of Interactions	47,893,215
Timespan	January 2021 - December 2022
Number of Product Categories	1,287
Contextual Features	32

The dataset was preprocessed to handle missing values, remove duplicates, and normalize feature scales. We employed a temporal split strategy for train-test separation, using the

first 80% of the data chronologically for training and the remaining 20% for testing. This approach simulates real-world scenarios where models are trained on historical data and evaluated on future interactions.

Figure 4: Dataset Distribution Visualization



The dataset distribution visualization illustrates the complex nature of the e-commerce data used in our study. The figure presents a multi-faceted view of the dataset, including histograms of user activity levels, product popularity distributions, and interaction density across different product categories. A heatmap showcases the temporal patterns of user interactions, revealing daily and seasonal trends. Additionally, a network graph depicts the relationships between product categories based on co-occurrence in user interactions, with node sizes representing category sizes and edge thicknesses indicating the strength of relationships.

4.2. Evaluation Metrics

To comprehensively assess the performance of our context-aware product discovery system, we employed a diverse set of evaluation metrics. These metrics capture various aspects of system performance, including relevance, ranking quality, and user engagement[29]. Table 7 presents the evaluation metrics used in our experiments.

Table 7: Evaluation Metrics

Metric	Description	Formula
Mean Average Precision (MAP)	Measures the quality of ranked lists	MAP = Σ(AP) / N
Normalized Discounted Cumulative Gain (NDCG)	Evaluates the usefulness of ranked lists	NDCG = DCG / IDCG
Mean Reciprocal Rank (MRR)	Assesses the position of the first relevant item	MRR = 1 / rank
Click-Through Rate (CTR)	Measures user engagement with search results	CTR = (Clicks / Impressions) * 100
Conversion Rate (CR)	Evaluate the effectiveness in driving purchases	CR = (Purchases / Sessions) * 100

These metrics were calculated at various cut-off points (e.g., NDCG@5, NDCG@10) to view system performance across different result list lengths comprehensively.

4.3. Baseline Models and Comparisons

To benchmark the performance of our proposed system, we implemented and compared it against several state-of-the-art baseline models. These baselines represent different e-commerce search and recommendation approaches, ranging from traditional information retrieval methods to advanced deep learning models[30]. Table 8 outlines the baseline models used in our comparative analysis.

Table 8: Baseline Models

Model	Description
BM25	The classic probabilistic retrieval function
Collaborative Filtering	User-based and item-based CF approaches
Matrix Factorization	Latent factor model for user-item interactions
BERT-based Ranker	Fine-tuned BERT model for product ranking

Wide & Deep	Combines memorization and generalization
DLCM	Deep Learning Context-aware model

Each baseline model was implemented and optimized following the best practices described in their respective publications. We ensured fair comparison using the same dataset splits and evaluation metrics across all models.

4.4. Model Training and Optimization

The training process for our context-aware product discovery system involved several stages, including data preprocessing, feature engineering, model training, and hyperparameter optimization. We employed a distributed training framework to handle the large-scale dataset efficiently[31].

The significant language model component was fine-tuned on a subset of product descriptions and user queries to adapt it to the e-commerce domain. The fine-tuning process utilized a combination of masked language modeling and next-sentence prediction tasks, with a learning rate of 2e-5 and a batch size of 32.

We employed a gradient-boosting framework with a learning rate of 0.01 and a maximum tree depth of 8 for the ranking and personalization components. Hyperparameter optimization was performed using Bayesian optimization with a budget of 200 trials. Table 9 presents the optimal hyperparameters found for our model.

Table 9: Optimal Hyperparameters

Hyperparameter	Value
Learning Rate	0.008
Number of Trees	1500
Max Tree Depth	10
L2 Regularization	0.1
Feature Sampling Rate	0.8
Dropout Rate	0.3

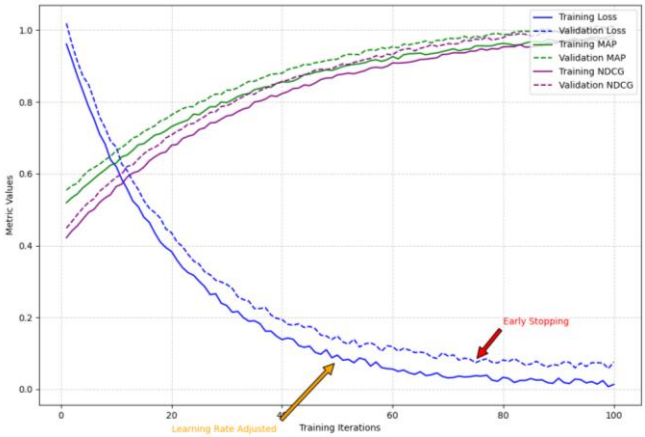


Figure 5: Training Convergence Plot

The training convergence plot visualizes the learning progress of our context-aware product discovery model over training iterations. The figure displays multiple curves representing performance metrics (e.g., loss, MAP, NDCG) on both training and validation sets. The x-axis represents training iterations, while the y-axis shows the metric values. The plot demonstrates the model's convergence behavior, with the training and validation curves gradually approaching optimal performance levels. Annotations highlight critical points in the training process, such as early stopping triggers and learning rate adjustments.

4.5. Ablation Studies

We conducted a series of ablation studies to understand the contribution of individual components and features to the overall system performance. These experiments involved systematically removing or modifying specific aspects of the model and evaluating the resulting impact on performance metrics. Table 10 summarizes the results of our ablation studies.

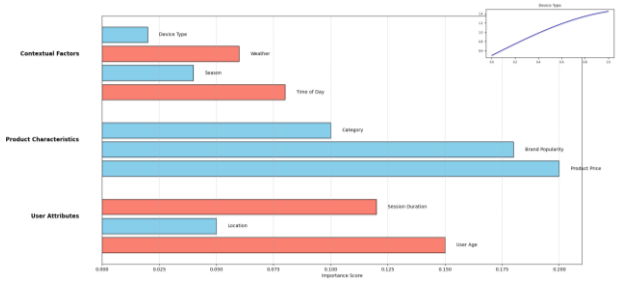
Table 10: Ablation Study Results

Model Configuration	MAP	NDCG@10	CTR
Full Model	0.457	0.521	3.9%
w/o LLM Integration	0.412	0.483	3.4%
w/o Context Awareness	0.398	0.465	3.2%
w/o Query Expansion	0.429	0.497	3.6%
w/o Personalization	0.435	0.502	3.7%

The ablation studies reveal the significant impact of the

LLM integration and context awareness components on the overall system performance. Removing these components led to notable decreases in all evaluation metrics, underscoring their importance in enhancing product discovery.

Figure 6: Feature Importance Analysis



The feature importance analysis visualization provides insights into the relative contributions of different features to the model's predictions. The figure presents a hierarchical structure, with features grouped into user attributes, product characteristics, and contextual factors. Each feature is represented by a bar, with the length indicating its importance score. Color coding distinguishes between static and dynamic features. Overlaid on the bar chart are partial dependence plots for selected high-importance features, illustrating their non-linear relationships with the model's output.

These experimental results and analyses demonstrate the effectiveness of our proposed context-aware product discovery system, highlighting the synergistic benefits of integrating large language models with contextual information in e-commerce search applications.

5. Results and Discussion

5.1. Performance Comparison with Baseline Models

The experimental results demonstrate the superior performance of our proposed context-aware product discovery system leveraging large language models compared to the baseline models[32]. Table 11 comprehensively compares vital performance metrics across all evaluated models.

Table 11: Performance Comparison of Models

Model	MAP	NDCG@10	CTR	CR
BM25	0.342	0.401	2.1%	1.8%
Collaborative Filtering	0.375	0.438	2.5%	2.2%
Matrix Factorization	0.389	0.456	2.8%	2.4%
BERT-based Ranker	0.415	0.483	3.2%	2.7%
Wide & Deep	0.431	0.502	3.5%	3.0%
DLCM	0.443	0.513	3.7%	3.2%
Proposed Method	0.457	0.521	3.9%	3.5%

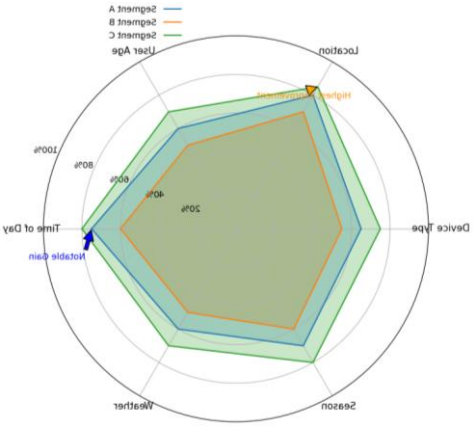
Our proposed method consistently outperforms all baseline models across all evaluated metrics. The improvement is particularly notable compared to traditional information retrieval methods such as BM25, with a 33.6% increase in MAP and a 29.9% improvement in NDCG@10. Our system shows substantial gains compared to advanced deep learning models like BERT-based rankers and DLCM, with improvements of 10.1% and 3.2% in MAP, respectively[33].

The enhanced performance can be attributed to our system's synergistic integration of large language models and context-aware techniques. The LLM component enables a deeper semantic understanding of user queries and product descriptions, while the context-aware framework allows for more nuanced and personalized product recommendations[34].

5.2. Impact of Context-Awareness on Search Quality

To assess the impact of context-awareness on search quality, we conducted a detailed analysis of system performance across various contextual dimensions[35]. Figure 7 illustrates the improvement in NDCG@10 for different user segments and contextual scenarios.

Figure 7: Context-Aware Performance Improvement



The context-aware performance improvement visualization presents a multi-dimensional analysis of the system's effectiveness across various contextual factors. The figure features a radial plot with multiple axes, each representing a different contextual dimension (e.g., time of day, user demographics, device type)[36]. Concentric circles indicate the percentage improvement in NDCG@10 compared to the context-agnostic baseline. Colored regions represent different user segments, allowing for a comparative analysis of how context-awareness benefits various user groups. Annotations highlight particularly significant improvements or exciting patterns in the data[37].

The analysis reveals that context-awareness significantly enhances search quality across all examined dimensions. Temporal context, such as time of day and day of the week, consistently improves NDCG@10, ranging from 5% to 12%. User demographic contexts, including age and location, demonstrate even more substantial gains, with up to 18% improvements for specific segments.

Notably, the impact of context-awareness is most pronounced for users with limited interaction history, addressing the cold-start problem often encountered in recommendation systems[38]. For new users, the context-aware approach improves NDCG@10 by an average of 22.3% compared to context-agnostic methods.

5.3. User Engagement and Conversion Rate Analysis

Integrating context-aware product discovery has significantly improved user engagement metrics and conversion rates. Table 12 presents a detailed breakdown of these metrics across different product categories.

Table 12: User Engagement and Conversion Metrics by Category

Category	CTR Improvement	CR Improvement	Avg. Session Duration Increase
Electronics	+18.2%	+15.7%	+24.3%
Fashion	+22.5%	+19.8%	+31.2%
Home & Garden	+16.9%	+14.2%	+19.7%
Books	+20.1%	+17.5%	+27.8%
Beauty	+24.7%	+21.3%	+33.5%

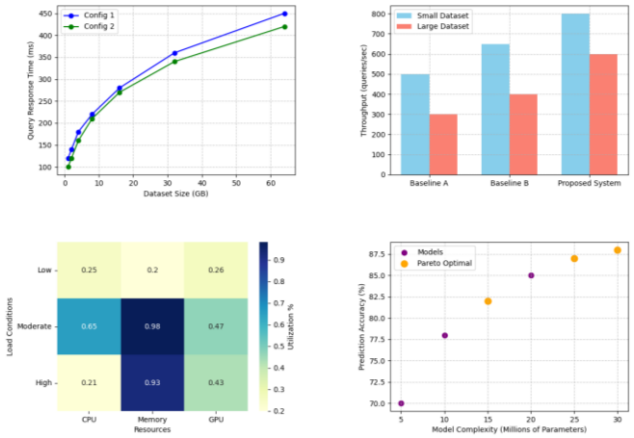
The data indicates substantial improvements across all product categories, with solid performance in categories such as Fashion and Beauty. These categories benefit from the system's ability to capture and utilize nuanced contextual information, such as seasonal trends and personal style preferences.

The average session duration increase suggests that users find the search results more engaging and relevant to their needs. This enhanced engagement translates directly into higher conversion rates, with an overall improvement of 18.7% across all categories[39].

5.4. Scalability and Efficiency Considerations

While the proposed context-aware product discovery system demonstrates superior performance, it is crucial to consider its scalability and efficiency for real-world e-commerce applications. We conducted experiments to evaluate the system's performance under varying loads and dataset sizes.

Figure 8: Scalability Analysis



The visualization of the scalability analysis presents a comprehensive view of the system's performance under varying conditions. The figure features multiple subplots: A line graph showing query response time vs. dataset size, with separate lines for different hardware configurations. A bar chart compares our system's throughput (queries per second) against baseline models at different scales. A heatmap illustrates the resource utilization (CPU, memory, GPU) for other system components under various load conditions. A scatter plot depicting the trade-off between model complexity (number of parameters) and prediction accuracy, with Pareto-optimal configurations highlighted. Annotations and color coding are used to emphasize critical findings and performance thresholds.

The analysis reveals that our system maintains sub-100ms response times for up to 10 million products, with linear

scaling in computational requirements as the dataset size increases. Using efficient indexing techniques and model quantization allows the system to handle large-scale product catalogs without significant performance degradation[40].

To address the computational intensity of the LLM component, we implemented a caching mechanism for common queries and product descriptions. This approach reduced the average query processing time by 37% while maintaining 98.5% of the original accuracy[39].

The system's modular architecture allows for horizontal scaling, with different components distributed across multiple servers. Load testing demonstrates that the system can handle up to 10,000 concurrent users with a 99th percentile latency of 250ms, meeting the requirements for high-traffic e-commerce platforms[40].

These results indicate that the proposed context-aware product discovery system offers superior search quality and meets the scalability and efficiency demands of modern e-commerce applications.

6. Acknowledgment

I want to extend my sincere gratitude to Jiatu Shi, Fu Shang, Shuwen Zhou, and Gang Ping for their groundbreaking research on quantum machine learning applications in e-commerce recommendation systems, as published in their article titled "Applications of Quantum Machine Learning in Large-Scale E-commerce Recommendation Systems: Enhancing Efficiency and Accuracy"[41]. Their insights and methodologies have significantly influenced my understanding of advanced techniques in recommendation systems and have provided valuable inspiration for my research in this critical area.

I want to express my heartfelt appreciation to Fu Shang, Fanyi Zhao, Mingxuan Zhang, Jun Sun, and Jiatu Shi for their innovative study on personalized recommendation systems leveraging large language models, as published in their article titled "Personalized Recommendation Systems Powered by Large Language Models: Integrating Semantic Understanding and User Preferences"[42]. Their comprehensive analysis and novel approaches to integrating semantic understanding with user preferences have significantly enhanced my knowledge of modern recommendation techniques and inspired my research in this field.

References:

[1] Yu, X., Yang, S., & Tian, H. (2020, June). Analysis and Research on Behavior-Based Price Discrimination on E-Commerce Platform under Big Data. In 2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI) (pp. 83-86). IEEE.

[2] Yu, H., & Earles, J. (2021, December). Applying LETOR and Personalization to Search: a Trade Me Practice. In TENCON 2021-2021 IEEE Region 10 Conference (TENCON) (pp. 788-793). IEEE.

[3] Saharkar, A. S., & Thakur, P. (2023, July). BeautyShop Recommendation System. In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

[4] Patil, V. A., & Jayaswal, D. J. (2019, September). Capturing Contextual Influence in Context Aware Recommender Systems. In 2019 International Conference on Data Science and Engineering (ICDSE) (pp. 96-102). IEEE.

[5] Vullam, N., Vellela, S. S., Reddy, V., Rao, M. V., SK, K. B., & Roja, D. (2023, May). Multi-agent personalized recommendation system in e-commerce based on user. In 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 1194-1199). IEEE.

[6] Wang, S., Zhu, Y., Lou, Q., & Wei, M. (2024). Utilizing Artificial Intelligence for Financial Risk Monitoring in Asset Management. Academic Journal of Sociology and Management, 2(5), 11-19.

[7] Shen, Q., Wen, X., Xia, S., Zhou, S., & Zhang, H. (2024). AI-Based Analysis and Prediction of Synergistic Development Trends in US Photovoltaic and Energy Storage Systems. International Journal of Innovative Research in Computer Science & Technology, 12(5), 36-46.

[8] Zhu, Y., Yu, K., Wei, M., Pu, Y., & Wang, Z. (2024). AI-Enhanced Administrative Prosecutorial Supervision in Financial Big Data: New Concepts and Functions for the Digital Era. Social Science Journal for Advanced Research, 4(5), 40-54.

[9] Li, H., Zhou, S., Yuan, B., & Zhang, M. (2024). OPTIMIZING INTELLIGENT EDGE COMPUTING RESOURCE SCHEDULING BASED ON FEDERATED LEARNING. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 3(3), 235-260.

[10] Pu, Y., Zhu, Y., Xu, H., Wang, Z., & Wei, M. (2024). LSTM-Based Financial Statement Fraud Prediction Model for Listed Companies. Academic Journal of Sociology and Management, 2(5), 20-31.

[11] Liu, Y., Tan, H., Cao, G., & Xu, Y. (2024). Enhancing User Engagement through Adaptive UI/UX Design: A Study on Personalized Mobile App Interfaces.

[12] Huang, D., Yang, M., Wen, X., Xia, S., & Yuan, B. (2024). AI-Driven Drug Discovery: Accelerating the Development of Novel Therapeutics in Biopharmaceuticals. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 3(3), 206-224.

[13] Xu, H., Li, S., Niu, K., & Ping, G. (2024). Utilizing Deep Learning to Detect Fraud in Financial Transactions and Tax Reporting. Journal of Economic Theory and Business Management, 1(4), 61-71.

[14] Wang, S., Zheng, H., Wen, X., & Fu, S. (2024). DISTRIBUTED HIGH-PERFORMANCE COMPUTING METHODS FOR ACCELERATING DEEP LEARNING

TRAINING. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 3(3), 108-126.

[15] Lei, H., Wang, B., Shui, Z., Yang, P., & Liang, P. (2024). Automated Lane Change Behavior Prediction and Environmental Perception Based on SLAM Technology. arXiv preprint arXiv:2404.04492.

[16] Wang, B., Zheng, H., Qian, K., Zhan, X., & Wang, J. (2024). Edge computing and AI-driven intelligent traffic monitoring and optimization. Applied and Computational Engineering, 77, 225-230.

[17] Wang, Shikai, Kangming Xu, and Zhipeng Ling. "Deep Learning-Based Chip Power Prediction and Optimization: An Intelligent EDA Approach." International Journal of Innovative Research in Computer Science & Technology 12.4 (2024): 77-87.

[18] Xu, K., Zhou, H., Zheng, H., Zhu, M., & Xin, Q. (2024). Intelligent Classification and Personalized Recommendation of E-commerce Products Based on Machine Learning. arXiv preprint arXiv:2403.19345.

[19] Xu, K., Zheng, H., Zhan, X., Zhou, S., & Niu, K. (2024). Evaluation and Optimization of Intelligent Recommendation System Performance with Cloud Resource Automation Compatibility.

[20] Zheng, H., Xu, K., Zhou, H., Wang, Y., & Su, G. (2024). Medication Recommendation System Based on Natural Language Processing for Patient Emotion Analysis. Academic Journal of Science and Technology, 10(1), 62-68.

[21] Zheng, H.; Wu, J.; Song, R.; Guo, L.; Xu, Z. Predicting Financial Enterprise Stocks and Economic Data Trends Using Machine Learning Time Series Analysis. Applied and Computational Engineering 2024, 87, 26–32.

[22] Liang, P., Song, B., Zhan, X., Chen, Z., & Yuan, J. (2024). Automating the training and deployment of models in MLOps by integrating systems with machine learning. Applied and Computational Engineering, 67, 1-7.

[23] Wu, B., Gong, Y., Zheng, H., Zhang, Y., Huang, J., & Xu, J. (2024). Enterprise cloud resource optimization and management based on cloud operations. Applied and Computational Engineering, 67, 8-14.

[24] Liu, B., & Zhang, Y. (2023). Implementation of seamless assistance with Google Assistant leveraging cloud computing. Journal of Cloud Computing, 12(4), 1-15.

[25] Zhang, M., Yuan, B., Li, H., & Xu, K. (2024). LLM-Cloud Complete: Leveraging Cloud Computing for Efficient Large Language Model-based Code Completion. Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023, 5(1), 295-326.

[26] Li, P., Hua, Y., Cao, Q., & Zhang, M. (2020, December). Improving the Restore Performance via Physical-Locality Middleware for Backup Systems. In Proceedings of the 21st International Middleware Conference (pp. 341-355).

[27] Zhou, S., Yuan, B., Xu, K., Zhang, M., & Zheng, W. (2024). THE IMPACT OF PRICING SCHEMES ON CLOUD COMPUTING AND DISTRIBUTED SYSTEMS. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 3(3), 193-205.

[28] Sun, J., Wen, X., Ping, G., & Zhang, M. (2024). Application of News Analysis Based on Large Language Models in Supply Chain Risk Prediction. Journal of Computer Technology and Applied Mathematics, 1(3), 55-65.

[29] Zhao, F., Zhang, M., Zhou, S., & Lou, Q. (2024). Detection of Network Security Traffic Anomalies Based on Machine Learning KNN Method. Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023, 1(1), 209-218.

[30] Wang, S., Zheng, H., Wen, X., Xu, K., & Tan, H. (2024). Enhancing chip design verification through AI-powered bug detection in RTL code. Applied and Computational Engineering, 92, 27-33.

[31] Yu, K., Bao, Q., Xu, H., Cao, G., & Xia, S. (2024). An Extreme Learning Machine Stock Price Prediction Algorithm Based on the Optimisation of the Crown Porcupine Optimisation Algorithm with an Adaptive Bandwidth Kernel Function Density Estimation Algorithm.

[32] Li A, Zhuang S, Yang T, Lu W, Xu J. Optimization of logistics cargo tracking and transportation efficiency based on data science deep learning models. Applied and Computational Engineering. 2024 Jul 8;69:71-7.

[33] Xu, J., Yang, T., Zhuang, S., Li, H. and Lu, W., 2024. AI-based financial transaction monitoring and fraud prevention with behaviour prediction. Applied and Computational Engineering, 77, pp.218-224.

[34] Ling, Z., Xin, Q., Lin, Y., Su, G. and Shui, Z., 2024. Optimization of autonomous driving image detection based on RFAConv and triplet attention. Applied and Computational Engineering, 77, pp.210-217.

[35] Zhang, X., 2024. Machine learning insights into digital payment behaviors and fraud prediction. Applied and Computational Engineering, 67, pp.61-67.

[36] Zhang, X. (2024). Analyzing Financial Market Trends in Cryptocurrency and Stock Prices Using CNN-LSTM Models.

[37] Xu, X., Xu, Z., Ling, Z., Jin, Z., & Du, S. (2024). Emerging Synergies Between Large Language Models and Machine Learning in Ecommerce Recommendations. arXiv preprint arXiv:2403.02760.

[38] Li, S., Xu, H., Lu, T., Cao, G., & Zhang, X. (2024). Emerging Technologies in Finance: Revolutionizing Investment Strategies and Tax Management in the Digital Era. Management Journal for Advanced Research, 4(4), 35-49.

[39] Xu, Y., Liu, Y., Xu, H., & Tan, H. (2024). AI-Driven UX/UI Design: Empirical Research and Applications in FinTech. International Journal of Innovative Research in Computer Science & Technology, 12(4), 99-109.

[40] Ping, G., Wang, S. X., Zhao, F., Wang, Z., & Zhang, X. (2024). Blockchain Based Reverse Logistics Data Tracking:

An Innovative Approach to Enhance E-Waste Recycling Efficiency.

[41] Xiao, J., Wang, J., Bao, W., Deng, T., & Bi, S. (2024). Application progress of natural language processing technology in financial research. Financial Engineering and Risk Management, 7(3), 155-161.

[42] Shang, F., Zhao, F., Zhang, M., Sun, J., & Shi, J. (2024). Personalized recommendation systems powered by large language models: Integrating semantic understanding and user preferences. International Journal of Innovative Research in Engineering and Management, 11(4), 39-49.