10.48047/jocaaa.2025.34.12.34

Metrics and Measurement Frameworks for E-Commerce Search Systems

Prathyusha Bhaskar Karnam
Independent Researcher, USA

Abstract

The ability to search is a fundamental part of e-commerce technologies today as shoppers browse through billions of available products. An effective search tool will accurately decipher what the user wants to find, retrieve the relevant products quickly, and rank the results in an order that makes it easier for consumers to make a purchase. How well these systems do their job determines how satisfied customers will be, how many customers will convert from browsing to actual purchases, and ultimately influences the revenues generated by each of these companies. Organizations need to put in place a comprehensive quality measurement framework that will allow them to systematically track how well their search system performs through the various stages of the search pipeline. Measurements at a component level provide insight into areas of weakness within the individual components of the search tool, while measurements tied to the success of the business illustrate how well technical performance can connect back to business success or outcomes. Organizations need to use more than one strategy for measuring their search tool, which can include controlled experimentation, holdout testing, real-time monitoring, and a combination of offline and online validation techniques. System-level measurement includes tracking the retrieval of products covered, the accuracy of ranking the results, the latency between searching and retrieving products, and the number of errors made. Business-facing measurement includes the click-through rate, conversion rates, total dollar contribution to the business from every product retrieved, and the lifetime value of a customer. When a company has a robust quality measurement framework, it gives that business a competitive edge through systematic optimization of its search technology, speed in detecting regressions in its products, and data-driven prioritization of product development. By integrating the technical and business measurements, companies ensure that their engineering improvements translate into measurable user value and continue to support sustained growth for the business. Ultimately, the effective use of a quality measurement framework will lead to better experiences for the customer and support the long-term goals of a business in a highly competitive e-commerce environment.

Keywords: E-Commerce Search Systems, Information Retrieval Metrics, Ranking Evaluation, Business Impact Measurement, Search Quality Optimization

1. Introduction

Search functionality serves as a cornerstone of modern e-commerce platforms. Users interact with vast product catalogs that change rapidly and contain millions of items. A simple query can potentially return thousands of relevant products. The search system must accurately interpret user intent and retrieve appropriate items. It then ranks these items effectively and presents them in ways that facilitate quick decision-making [1].

The quality of search directly influences key business outcomes. User satisfaction depends heavily on finding relevant products quickly. Conversion rates rise when search results match user expectations. Revenue growth correlates with search effectiveness. In competitive markets, search quality becomes a primary differentiator. Users have abundant choices and high expectations for their shopping experience.

10.48047/jocaaa.2025.34.12.34

Search systems operate as complex, interconnected pipelines with multiple stages. Each component influences downstream performance. Query understanding affects retrieval quality. Retrieval coverage constrains ranking effectiveness. Ranking determines which products users ultimately see. This interdependence makes comprehensive measurement essential for system optimization [2].

Rigorous measurement enables teams to diagnose performance issues systematically. Engineers can compare algorithmic approaches quantitatively. Product teams can track improvements over time and prevent quality regressions. Most importantly, measurement links technical metrics to business outcomes. This connection allows organizations to optimize search systems reliably at scale. Traditional approaches focus primarily on end-result metrics like precision and recall. However, this narrow view misses critical upstream issues [1].

Effective measurement requires evaluating each pipeline component individually. System-level metrics reveal where problems originate and where improvements deliver maximum impact. Benchmark datasets like BEIR provide standardized evaluation frameworks across diverse retrieval tasks. These benchmarks enable zero-shot evaluation of information retrieval models without domain-specific training. This comprehensive approach creates true end-to-end visibility into search performance [2].

2. Core Components of E-Commerce Search Systems

The modern architecture of e-commerce search relies on various interlinked components, each with a specific role within the system and contributing to its overall performance. Understanding these components individually enables targeted measurement and optimization strategies. Major platforms have evolved sophisticated machine learning algorithms to handle search complexity at scale [3].

2.1 Query Understanding

Query understanding interprets user intent before retrieval begins. The system must map natural language queries to structured catalog attributes. This process determines which products the system should consider relevant. Intent classification identifies whether users seek specific products or browse categories. The system analyzes query structure, terms, and context to make these determinations.

Accurate query understanding enables downstream components to function effectively. Misinterpreted queries lead to poor retrieval coverage regardless of ranking quality. The system must handle misspellings, synonyms, and implicit intent. It should recognize brand names, product categories, and attribute specifications [3]. Query expansion and reformulation techniques help bridge gaps between user language and catalog terminology. Search intention networks can personalize query auto-completion based on user behavior patterns [10].

2.2 Retrieval Systems

Retrieval systems surface candidate products from the full catalog. This stage prioritizes broad coverage of potentially relevant items. The retrieved set feeds into ranking algorithms that perform finer-grained ordering. Retrieval typically employs inverted indices, embedding-based search, or hybrid approaches. These methods balance efficiency with coverage to return candidates quickly [4].

Strong retrieval ensures that relevant products reach the ranking stage. Poor retrieval creates an upper bound on achievable search quality. Even a perfect ranking cannot surface items that the retrieval missed. The system must handle lexical matching for specific queries and semantic matching for broader intent. Real-time personalization using embeddings has proven effective for large-scale search applications. Embedding-based retrieval can capture semantic similarities that traditional keyword matching misses [4].

2.3 Ranking Systems

10.48047/jocaaa.2025.34.12.34

Ranking systems order retrieved candidates by relevance and business value. Machine learning models typically perform this ordering using features like click-through rates, conversion history, and product attributes. The ranking stage determines which products users actually see and in what order. This directly influences user engagement and purchase decisions. Discriminative semantic rankers can improve relevance by learning from user interaction patterns [5].

Effective ranking balances multiple objectives simultaneously. Relevance to query intent matters most for user satisfaction. Business considerations like inventory levels and profit margins also factor into ordering. Personalization signals help surface products matching individual preferences. The ranking model must calibrate these competing signals to optimize overall performance. Multi-modal user embeddings can enhance recommendation quality by incorporating diverse signal types [6].

2.4 Result Presentation

The result presentation determines how ranked products appear to users. Layout, imagery, and metadata influence click-through behavior. The system must decide what information to display for each product. Thumbnail quality, price visibility, and review ratings all affect user decisions. Presentation strategies should reduce cognitive load and facilitate quick comparisons.

Different query types may warrant different presentation formats. Navigational queries benefit from category-based layouts. Product-finding queries work well with grid displays. Comparison queries might show side-by-side attributes. The presentation layer adapts to intent and context to optimize user experience.

Component	Primary Function	Key Characteristics
Query Understanding	Interprets user intent and maps natural language to catalog attributes	Handles misspellings, synonyms, brand recognition, and intent classification
Retrieval Systems	Surfaces candidate products from full catalog for downstream processing	Employs inverted indices, embeddings, or hybrid approaches for broad coverage
Ranking Systems	Orders retrieved candidates by relevance and business value	Balances multiple objectives including relevance, inventory, and personalization
Result Presentation	Displays ranked products with appropriate layout and metadata	Adapts format based on query type to reduce cognitive load

Table 1: Core Components and Functions of E-Commerce Search Systems [3, 4]

10.48047/jocaaa.2025.34.12.34

3. Measurement Approaches for Search System Components

Evaluating search systems requires distinct metrics for each component. System-level measurements reveal specific weaknesses and optimization opportunities. Component metrics complement business metrics to provide comprehensive visibility into search performance. Classical information retrieval principles provide foundational frameworks for evaluation [7].

3.1 Query Understanding Metrics

Intent classification accuracy measures how correctly the system identifies user goals. Evaluators compare predicted intent labels against human annotations. High accuracy ensures downstream components receive appropriate signals. This metric requires carefully labeled evaluation sets that represent real query distributions. Annotation guidelines must define intent categories clearly and handle edge cases consistently [7].

Query reformulation rate tracks how often users modify their searches after viewing results. High reformulation rates indicate mismatches between interpreted intent and user expectations. The system measures reformulation by identifying successive queries within short time windows. Contextual reformulations suggest more severe understanding failures than simple refinements. Teams should analyze reformulation patterns to identify systematic interpretation errors. Machine learning techniques can predict when users will reformulate queries based on session context [8].

3.2 Retrieval System Metrics

Recall at K measures what fraction of relevant items appear in the top candidates. This metric compares retrieved sets against known relevant products for evaluation queries. Recall directly influences ranking performance by constraining the pool of orderable items. Low recall indicates insufficient catalog coverage or overly restrictive matching. Teams typically evaluate recall at multiple depth values to understand coverage at different positions [1].

Zero result rate quantifies how often retrieval returns no candidates. This metric tracks the percentage of queries producing empty result sets. Rising zero-result rates signal problems with indexing, catalog coverage, or query processing. Teams should segment zero-result metrics by query characteristics to identify specific failure modes. Common causes include strict matching rules, limited inventory, or poor synonym handling. Heterogeneous benchmarks enable evaluation across diverse retrieval scenarios without requiring task-specific model training [1].

3.3 Ranking System Metrics

Normalized Discounted Cumulative Gain evaluates ranking quality with position-based discounting. NDCG uses graded relevance labels rather than binary judgments. Higher-ranked relevant items contribute more to the score than lower-ranked ones. The metric normalizes against the ideal ranking to enable cross-query comparison. NDCG particularly emphasizes top positions where users focus attention [2].

Mean Reciprocal Rank measures how quickly users find relevant results. MRR calculates the reciprocal of the first relevant item's position and averages across queries. Higher MRR indicates that relevant products appear early in result lists. This metric matters greatly for user satisfaction since most users examine only the top results. Mean Average Precision captures overall ranking quality across all relevant items. Learning to rank algorithms optimize these metrics through supervised training on labeled query-document pairs [2].

Model calibration through observed-to-expected ratios assesses prediction reliability. These ratios compare predicted engagement probabilities against actual user behavior. Well-calibrated models

10.48047/jocaaa.2025.34.12.34

maintain consistent ratios near target values across different segments. Poor calibration indicates that predicted scores misrepresent actual engagement likelihoods. Teams should monitor calibration across query types, product categories, and user segments.

3.4 Operational Performance Metrics

Latency measures response time from query submission to result display. This includes query processing, retrieval, ranking, and presentation time. When it comes to search results, customers in today's e-commerce world expect nearly immediate responses. Latency can create a poor experience for users and lead to decreased rates of conversion, so monitoring latency at multiple percentile levels for tail latencies is essential.

Error rate tracks the frequency of failed or incomplete search requests. Errors include timeouts, server failures, and partial result sets. This metric indicates system reliability and operational health. Rising error rates suggest infrastructure problems, dependency failures, or capacity constraints. Teams should alert on error rate spikes and investigate root causes immediately. Segmenting errors by type and component helps prioritize fixes.

Measurement Category	Purpose	Implementation Approach
Query Understanding Assessment	Evaluates accuracy of intent identification and user satisfaction	Compares predicted labels to annotations and tracks query reformulation patterns
Retrieval Coverage Evaluation	Determines fraction of relevant items surfaced and empty result frequency	Measures recall at various depths and monitors zero-result occurrences
Ranking Quality Assessment	Evaluates ordering effectiveness and prediction reliability	Uses position-based discounting and calibration monitoring across segments
Operational Performance Tracking	Monitors system responsiveness and reliability	Tracks response times at percentiles and frequencies of request failures

Table 2: Component-Level Measurement Approaches and Their Applications [1, 2]

4. Business Impact Measurement

Business metrics connect technical search quality to organizational outcomes. These measurements capture user behavior, engagement patterns, and revenue effects. Unlike system metrics, business metrics

10.48047/jocaaa.2025.34.12.34

reflect how search improvements translate into customer value and company growth. E-commerce platforms increasingly rely on machine learning to optimize these business outcomes [8].

4.1 Short-Term Engagement Metrics

Click-through rate measures what fraction of queries result in clicks. CTR indicates whether displayed results attract user interest. Low CTR suggests poor relevance, unappealing presentation, or mismatched intent. Teams should analyze CTR by query type and result position to identify patterns. First-position CTR particularly matters since users focus heavily on top results [3].

Add-to-cart rate tracks how often clicked items get added to shopping carts. ATCR indicates whether clicked products meet user expectations upon closer examination. This metric captures the transition from interest to purchase intent. Low ATCR relative to CTR suggests that results appear relevant but disappoint on detailed inspection. Product information quality, pricing, and availability all influence ATCR. Machine learning models can predict add-to-cart likelihood based on user behavior and product characteristics [8].

Conversion rate measures what percentage of searches lead to purchases. CVR directly connects search quality to revenue generation. This metric depends on the entire search-to-purchase funnel including checkout experience. Search teams should measure CVR with appropriate attribution windows to capture delayed purchases. Multi-touch attribution helps isolate search's specific contribution to conversions [3].

Average click position indicates where, in the result lists, users find relevant items. Lower values mean users discover appropriate products quickly without extensive scrolling. High click positions suggest that relevant items rank too low. This metric complements CTR by revealing result quality deeper in lists. Teams should track click position distributions rather than just averages.

Scroll depth tracks how far users navigate through search results before taking action. Shallow scrolling with high engagement suggests effective ranking and relevant top results. Deep scrolling may indicate that relevant items are buried or missing. This behavioral signal reveals whether users trust top-ranked items. Dashboard-based monitoring allows teams to have a real-time view of how their systems are functioning and the health of those systems; the teams utilize alerting and visualization tools to track KPIs and monitor metrics from week to week to enable them to quickly identify anomalies. This indicates that search surfaces relevant and compelling products. Short dwell times followed by return to results suggest mismatched expectations. Teams should analyze dwell time distributions across different product categories and query types.

4.2 Long-Term Business Metrics

Gross Merchandise Value captures total purchase value from search-driven transactions. GMV directly measures search's contribution to the company's revenue. This metric should attribute purchases to search appropriately using session tracking. GMV growth indicates that search improvements drive business outcomes. Teams should decompose GMV by product category, user segment, and query type [8].

Average Order Value tracks purchase sizes for search-initiated transactions. AOV indicates whether search surfaces high-value products effectively. Higher AOV suggests that ranking and recommendation logic successfully promote premium items. This metric matters particularly for businesses with broad price ranges. Teams should balance AOV optimization against conversion rate effects.

Customer retention tracks whether users continue engaging with the platform over extended periods. Strong search experiences encourage repeat usage and build platform loyalty. Retention metrics reveal whether search quality influences long-term user behavior. Teams should measure retention cohorts to understand how search improvements affect user stickiness. Higher retention translates to increased lifetime value and sustained revenue [8].

10.48047/jocaaa.2025.34.12.34

Customer Lifetime Value estimates the total value users generate throughout their relationship with the platform. CLV incorporates repeat purchases, retention, and engagement over time. Strong search experiences increase CLV by facilitating discovery and building user trust. This metric requires longitudinal tracking and predictive modeling. Multi-modal recommendation frameworks can enhance long-term user engagement by personalizing experiences across touchpoints [6].

Metric Category	Business Insight	Strategic Application
Engagement Funnel Tracking	Measures user progression from query to purchase	Identifies friction points and optimization opportunities across conversion path
Behavioral Pattern Analysis	Reveals user interaction depth and product discovery efficiency	Guides ranking improvements and interface design decisions
Revenue Attribution	Quantifies direct financial contribution from search functionality	Justifies infrastructure investment and algorithm development resources
Long-term Value Assessment	Estimates sustained user engagement and repeat purchase behavior	Supports retention strategies and loyalty program development

Table 3: Business Impact Indicators and Strategic Value [3, 8]

5. Search Measurement Strategies

Organizations employ multiple complementary approaches to evaluate search systems. Each strategy provides distinct insights into system performance and business impact. Combining these methods creates comprehensive visibility into search quality and effectiveness. Modern platforms leverage both offline and online evaluation to balance development speed with user safety [4].

5.1 Controlled Experimentation

Randomized controlled experiments enable teams to measure causal effects of search changes. Experiments randomly assign users to treatment and control conditions. This isolates the impact of specific modifications from broader traffic trends. Teams can test ranking algorithms, interface changes, or personalization features safely. Experiments typically run for sufficient duration to achieve statistical significance [4].

Experimental metrics span system and business levels simultaneously. Latency and recall track technical performance characteristics. CTR and conversion rate measure user engagement behaviors. Revenue metrics capture business impact directly. Teams should monitor multiple metrics simultaneously to detect unintended consequences. Positive effects on one metric might coincide with negative effects on others.

10.48047/jocaaa.2025.34.12.34

Embedding-based personalization approaches require careful experimentation to validate improvements across diverse user segments [6].

Statistical rigor ensures that experimental results reflect true effects rather than random variation. Sample size calculations determine how much traffic experiments require. Power calculations estimate the minimum detectable effect sizes. Significance testing quantifies confidence in observed differences. Teams should account for multiple comparison corrections when evaluating many metrics. Proper statistical methodology prevents false positive conclusions [6].

5.2 Holdout Experiments

Holdout experiments extend testing over longer timeframes to capture delayed effects. A portion of traffic remains excluded from new features for weeks or months. This approach reveals impacts on retention, lifetime value, and other long-term outcomes. Holdout groups provide ongoing baselines for measuring cumulative improvements. Teams use holdouts when immediate engagement metrics don't fully capture value.

Long-term holdouts require careful maintenance and monitoring throughout their duration. Teams must ensure holdout groups remain representative as user populations evolve. Technical infrastructure must reliably assign and maintain user treatment status. Periodic tracking shows how effects develop over time. Teams should balance holdout duration against the opportunity costs of withholding improvements.

5.3 Continuous Monitoring

Although holdout testing demonstrates how customer retention and lifetime value change over time as effects of marketing become apparent, these immediate metrics will not capture all long-term effects on customer retention and lifetime value because there is often a delay between the two. This proactive approach catches regressions before they significantly impact users. Monitoring should cover both technical and business metrics comprehensively [7].

Service-level agreements define acceptable performance thresholds formally. SLAs codify expectations for latency, error rates, and availability. Breaching SLAs triggers escalations and incident response procedures. Teams use SLA compliance rates as operational health indicators. SLA-driven monitoring ensures consistent system reliability at scale. Information retrieval systems require continuous quality assessment to maintain performance standards [7].

Anomaly detection algorithms flag unusual metric patterns automatically. Machine learning models learn normal behavior and identify deviations. This catches issues that simple threshold alerts miss easily. Anomaly detection works particularly well for metrics with temporal patterns. Teams investigate detected anomalies to determine if intervention is needed [9].

Alerting systems notify teams when metrics exceed predefined thresholds. Alert design balances sensitivity against false positive rates. Critical alerts page on-call engineers immediately. Warning-level alerts aggregate for periodic review. Teams should regularly tune alert thresholds based on operational experience. Effective alerting enables rapid response to quality degradations. Deep learning techniques can improve anomaly detection accuracy by learning complex patterns in system behavior [9].

5.4 Offline and Online Evaluation

Offline evaluation uses historical data and labeled datasets to assess system changes. Teams compute metrics like NDCG and recall on logged interactions or curated test sets. This enables rapid iteration without exposing users to experimental changes. Offline evaluation catches obvious regressions before deployment. However, offline metrics imperfectly predict online outcomes [1].

Online evaluation observes real user interactions with production systems. Live traffic provides ground truth for engagement and business metrics. Online testing validates that offline improvements translate to

10.48047/jocaaa.2025.34.12.34

actual user value. Teams should correlate offline and online metrics to build predictive relationships. Strong offline-online correlation enables confident deployment decisions. Zero-shot evaluation frameworks enable testing across diverse domains without requiring extensive labeled data for each scenario [1].

Hybrid approaches combine offline and online evaluation strategically. Teams use offline metrics for initial screening and rapid iteration. Promising changes graduate to small-scale online experiments. Successful experiments scale to full deployment with continued monitoring. This funnel balances development velocity with deployment safety. Multiple validation stages reduce risk while maintaining iteration speed [10].

Replay-based evaluation simulates online serving using logged traffic. Counterfactual reasoning estimates how historical users would have responded to new algorithms. This technique provides intermediate validation between offline and live experiments. Replay evaluation helps predict online performance more accurately than static offline metrics. Search intention networks can be evaluated through replay to assess personalization quality [10].

Strategy Type	Primary Application	Key Advantage
Controlled Experimentation	Tests algorithmic and interface modifications with isolated user groups	Establishes causal relationships between changes and measurable outcomes
Extended Holdout Testing	Captures delayed effects on retention and lifetime value	Reveals long-term impacts beyond immediate engagement signals
Real-time Monitoring Systems	Provides continuous visibility into performance and health indicators	Enables rapid detection and response to quality degradations
Combined Evaluation Approaches	Integrates historical analysis with live traffic observation	Balances development speed with validation of actual user value

Table 4: Evaluation Strategies and Their Organizational Benefits [4, 7]

10.48047/jocaaa.2025.34.12.34

Conclusion

Comprehensive measurement frameworks prove essential for maintaining and improving e-commerce search systems at scale. The complexity of modern search pipelines demands evaluation approaches that span multiple dimensions and system components. Technical metrics provide visibility into retrieval coverage, ranking accuracy, query understanding, and operational reliability. Business metrics connect these technical characteristics to user behavior, engagement patterns, and revenue outcomes. Organizations must implement both system-level and business-facing measurements to optimize search effectively.

Multiple evaluation strategies complement each other to provide a thorough system assessment. Through controlled experimentation, teams can evaluate the individual impact of various changes based on randomisation and statistical testing. Frequent regression monitoring will ensure that products are held to a minimum baseline quality and can be delivered without customers experiencing negative product interactions that would result in them leaving. Teams can use both offline and online measurements to deliver fast-paced development and maintain the ability to deliver actual user value. Each of these methods provides a solution to a piece of the measurement issue.

Effective measurement delivers tangible organizational benefits beyond simple performance tracking. Early regression detection prevents quality degradations from reaching users at scale. Component-level metrics enable precise diagnosis of issues and targeted optimization efforts. Quantified business impact justifies continued investment in search infrastructure and algorithm development. Benchmarking capabilities support realistic goal setting and strategic planning. Teams with a strong experimental culture can feel confident in the rigor of their measurements when testing new ideas for future developments. By integrating technical metrics with business metrics, the improvements made in engineering can be converted into measurable user satisfaction and ultimately lead to billion-dollar businesses through the increase in business growth. Search optimization becomes data-driven rather than intuition-based when comprehensive measurement exists. Resource allocation flows toward opportunities with the highest expected impact on user experience and revenue. Cross-functional teams align around shared metrics and common quality goals. Measurement transparency builds organizational confidence in search as a strategic capability.

E-commerce environments continue evolving with changing catalogs, rising user expectations, and advancing competitor capabilities. Organizations that master search measurement gain sustainable competitive advantages through systematic quality improvement. Robust measurement frameworks enable adaptation to these dynamics while maintaining reliable performance. The systematic optimization approach that measurement enables supports long-term success. By investing in measurement setup and processes, organisations make better decisions, iterate quickly, and deliver better user experiences. Comprehensive search measurement should be viewed by organisations as an essential infrastructure that supports both operational and financial success, not as a burden.

References

1. Nandan Thakur, et al., "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models," arXiv, 2021. Available: https://arxiv.org/abs/2104.08663
2. Hang Li, "Learning to Rank for Information Retrieval and Natural Language Processing, Second Edition," Springer, 2015. Available: https://link.springer.com/book/10.1007/978-3-031-02155-8

10.48047/jocaaa.2025.34.12.34

3. Vernon Elliot, "Optimizing E-Commerce Searches: The Evolution of Amazon's Machine Learning Algorithms for Keyword Relevance," Signalytics, 2023. Available: https://signalytics.ai/optimizing-ecommerce-searches-evolution-of-amazons-a9/
4. Mihajlo Grbovic, et al., "Real-time Personalization using Embeddings for Search Ranking at Airbnb," ACM Digital Library, 2018. Available: https://dl.acm.org/doi/10.1145/3219819.3219885
5. Yinqiong Cai, et al., "A Discriminative Semantic Ranker for Question Retrieval," ACM Digital Library, 2021. Available: https://dl.acm.org/doi/pdf/10.1145/3471158.3472227
6. Aditya Pal, et al., "PinnerSage: Multi-Modal User Embedding Framework for Recommendations at Pinterest," arXiv, 2020. Available: https://arxiv.org/abs/2007.03634
7. Baeza-Yates Ricardo and Ribeiro-Neto Berthier, "Modern Information Retrieval," 3rd Edition, ACM Press, 1999. Available: https://web.cs.ucla.edu/~miodrag/cs259-security/baeza-yates99modern.pdf
8. Elias Dritsas, et al., "Machine Learning in E-Commerce: Trends, Applications, and Future Challenges," IEEE Xplore, 2025. Available: https://ieeexplore.ieee.org/document/11009009
9. Yuanguo Lin, et al., "A Comprehensive Survey on Deep Learning Techniques in Educational Data Mining," arXiv, 2024. Available: https://arxiv.org/html/2309.04761v3
10. Wei Bao et al., "Search Intention Network for Personalized Query Auto-Completion in E-Commerce," arXiv, 2024. Available: https://arxiv.org/html/2403.02609v1