JOURNAL OF INTERNATIONAL CRISIS AND RISK COMMUNICATION RESEARCH
ISSN: 2576-0017
2026, VOL 9, NO 1

How Large Language Models Are Transforming E-Commerce Search

Prathyusha Bhaskar Karnam

Independent Researcher, USA

Abstract

E-commerce search has evolved from simple keyword matching to sophisticated semantic understanding through the adoption of large language models. Traditional search systems, reliant on lexical matching and inverted indexes, struggle with vocabulary mismatches, implicit user intent, and fragmented product information across text and images. Large language models fundamentally transform every layer of the search stack by introducing genuine language comprehension capabilities. These models employ transformer-based architectures with attention mechanisms to generate dense vector representations that capture semantic relationships between queries and products, enabling the retrieval of relevant items even without exact keyword overlap. They consolidate previously separate functions like category classification, attribute extraction, and query rewriting into unified systems that process natural language queries holistically. In conversational contexts, neural models maintain dialogue state and track user preferences across multiple turns, while multimodal architectures extract comprehensive product attributes by synthesizing information from text descriptions, packaging images, and product demonstrations. Vision-language models enable entirely new search paradigms where users can query with images alongside textual constraints, reasoning about visual style and semantic meaning simultaneously. By combining semantic search with structured metadata filtering, these systems expand discovery beyond literal matches to suggest complementary products, substitutes, and themed bundles based on usage contexts and culinary traditions encoded in pre-training data. This transformation represents a fundamental architectural shift from rigid keyword matching to flexible semantic understanding, creating search experiences that feel intuitive, reduce query reformulation, and surface relevant products that traditional systems systematically overlook.

Keywords: Large Language Models, Semantic Search, E-Commerce Retrieval, Multimodal Understanding, Query Interpretation.

Introduction

The traditional pipeline in which e-commerce search has been realized is to match the keywords, rank, and then to make attempt to comprehend what the shoppers desire. This method, however practical, has not been able to cope with the complexity of natural language and subtle variations of how individuals communicate their needs. Its essence is the existence of a semantic divide between the search queries as defined by users and the description of the products in the catalog systems. The use of inverted indexes to do the matching of exact terms to keywords in traditional systems makes them prone to failure whenever the user uses a synonym or colloquialism, context-dependent languages that do not match the catalog names. The understanding of queries to search engines has widely reported such limitations, with how most of the traditional systems do not see the actual intent of user queries, especially where the queries were based on implicit attributes, the use of comparative language, or terminologies that are domain-specific and therefore need to be understood in the context of the query [1]. The inflexibility of the lexical matching implies that even the slightest change in wording can lead to radically different outcomes, frustrating users and depriving retailers of possible sales.

This is being redefined by large language models that are providing real language understanding to each layer of the search stack. These models do not simply correlate words, but they understand intent, reason under context, and come up with insights to make product discovery more accommodating and powerful. It has been shown that the use of transformer-based architecture, and especially BERT and its derivatives, is able to learn the semantic similarity between queries and product descriptions with impressive text-bidirectional processing and contextual relationships, which are absent in traditional models [2]. This change is felt in the ways systems access products with dense embedding representations, the ways they rank the results with contextual relevance scoring, the way they extract structured features out of unstructured catalog information, and even in the way they measure their own performance with automated relevance evaluation.

The replacement of the hard keyword matching by a softer semantic interpretation is a paradigm change in e-commerce search architecture, allowing systems to cope with the entire range of natural language richness that describes actual user behavior. Contemporary applications use pre-trained language models, which have learned extensive knowledge about language, enabling them to generalize between product categories as well as new query patterns that have never been seen before without needing to write extensive rules or category-related training data [1]. The models can also be useful in disambiguating complex queries, learning implicit constraints, and knowing when users are using indirect language to describe their preferences instead of using explicit filters. The encoding of queries and products into common semantic spaces allows these systems to find the relevant matches to the query when there is zero lexical overlap between the search terms and product metadata [2]. This functionality is especially useful in the case of long-tail and cross-category searches, as well as in situations where the user does not have the exact words needed, but can define what they desire in everyday speech. The outcome is a more intuitive search experience, requires less query reformulation, and brings forth useful products that would be systematically missed by traditional systems.

Reimagining Search Retrieval

Traditional retrieval systems face a critical limitation: they can only find what's explicitly labeled. When someone searches for "healthy organic snacks," conventional systems look for those exact terms in product metadata. They'll surface items tagged with "organic" or "healthy" but completely miss unlabeled yet relevant options like hummus or fresh vegetables. This vocabulary mismatch represents a fundamental weakness in keyword-based approaches, where the retrieval effectiveness depends entirely on the overlap between query terms and indexed product attributes. Classical retrieval methods like BM25, which have dominated information retrieval for decades, operate purely on lexical matching principles and cannot bridge the semantic gap between different ways of expressing the same concept. The problem intensifies with sparse product catalogs where metadata quality varies significantly, and with diverse user populations who describe the same products using vastly different vocabulary based on their background, expertise level, and search context. Research on pretrained transformers for text ranking has demonstrated that traditional sparse retrieval approaches systematically underperform on queries requiring semantic understanding, as they lack the capacity to learn conceptual relationships between terms and rely entirely on surface-level word matching [3].

Language models address this through semantic understanding. By generating dense vector representations of both queries and products, these systems can match meaning rather than mere words. They capture that "healthy organic snacks" relates conceptually to fresh vegetables, even without shared terminology. Neural embedding approaches using transformer architectures encode semantic similarity into continuous vector spaces, where documents and queries that share conceptual meaning cluster regardless of lexical overlap. The application of models like BERT, which processes text bidirectionally through multi-headed self-attention mechanisms, enables the capture of contextual nuances that determine whether terms are truly relevant to a query's intent [3]. These architectures learn during pre-training on massive text corpora that terms like "wholesome," "nutritious," and "clean eating" all relate to health concepts, enabling them to retrieve relevant products even when exact query terms are absent from metadata. The transformer's attention mechanism weighs the importance of different tokens based on surrounding context, allowing the model to disambiguate polysemous terms and understand that "organic" in a food context differs from "organic" in chemistry or farming contexts.

Some implementations also create learned sparse representations that emphasize critical tokens like brand names or product codes, balancing semantic flexibility with precision. These hybrid systems combine the strengths of both approaches: dense vectors capture broad semantic relationships and handle vocabulary mismatch, while sparse representations ensure that specific, high-signal terms receive appropriate weight. Recent advances in learning effective representations for retrieval have focused on self-distillation techniques with adaptive relevance margins, where knowledge from larger teacher models is transferred to more efficient student models while maintaining retrieval quality [4]. This approach allows the student model to learn nuanced relevance signals by observing how the teacher model scores query-document pairs, with adaptive margins that account for varying difficulty levels across different query types. The self-distillation process with adaptive relevance margins helps models develop more discriminative representations that can better distinguish between highly relevant, marginally relevant, and irrelevant items [4]. This hybrid approach delivers both the broad understanding of semantic search and the accuracy of exact matching, filling gaps that have long plagued e-commerce retrieval while maintaining the computational efficiency necessary for real-time search applications serving millions of queries daily.

Table 1: Comparison of Traditional vs. Language Model-Based Retrieval Approaches [3, 4]

Aspect	Traditional Retrieval (BM25/Keyword)	Language Model-Based Retrieval	Improvement Achieved
Matching Approach	Lexical/keyword matching	Semantic understanding via embeddings	Conceptual vs. surface-level
Vocabulary Mismatch Handling	Poor - requires exact term overlap	Strong - matches meaning without shared terms	Handles synonym/conceptual variations
Query Understanding	Surface-level word matching	Contextual nuance through attention mechanisms	Disambiguates polysemous terms
Representation Type	Sparse inverted indexes	Dense vector spaces (hybrid approaches)	Combines precision with semantic breadth
Long-tail Query Performance	Weak - relies on term frequency	Strong - generalizes from pre-training	Better coverage for rare queries
Metadata Dependency	High - only finds explicitly labeled items	Low - infers relevance from semantic similarity	Retrieves unlabeled relevant products
Model Architecture	Rule-based/statistical	Transformer-based (BERT, self-attention)	Captures contextual relationships
Knowledge Transfer	Not applicable	Teacher-student distillation with adaptive margins	Efficient deployment at scale

Enhancing Ranking and Query Interpretation

Beyond retrieval, language models improve how results are ordered and how queries are interpreted. In ranking, they provide fine-grained relevance scoring that captures nuances traditional features miss. Traditional learning-to-rank systems rely on hand-crafted features such as click-through rates, term frequency, and position bias signals, which often fail to capture the semantic alignment between query intent and document content. These models generate semantic signals that integrate into existing ranking systems without requiring complete infrastructure overhauls. The application of transformer-based architectures in ranking tasks has demonstrated significant improvements over conventional methods by leveraging contextualized representations that capture deeper semantic relationships between queries and candidate documents. Research on neural ranking approaches has shown that pre-trained language models can encode complex linguistic patterns and domain-specific knowledge that traditional feature engineering cannot easily replicate, enabling more accurate relevance assessments across diverse query types [5]. Their pre-training on massive text corpora helps them handle rare queries effectively, surfacing relevant results even for unusual or niche product searches. The models develop broad linguistic knowledge during pre-training that generalizes to domains and query patterns not explicitly seen during task-specific fine-tuning, enabling them to assess relevance for long-tail queries that appear infrequently in training data but represent a significant portion of real user traffic.

Query understanding benefits even more dramatically. Legacy systems typically employ separate models for category classification, attribute extraction, and query rewriting, leading to inconsistent predictions where the category classifier might suggest one domain while the attribute extractor identifies properties from another, creating conflicts that degrade search quality. Language models consolidate these functions into unified systems that generate structured outputs in a single pass, treating query understanding as a sequence-to-sequence generation task where the input is the raw user query and the output is a structured representation containing all relevant semantic elements. The utilization of BERT and transformer models in unified architectures has proven particularly effective for tasks requiring both extraction and generation capabilities, as these models can simultaneously identify relevant information spans and synthesize coherent, structured outputs [6]. A practical implementation uses a teacher-student architecture: a large model generates high-quality training examples from user queries, including rewritten versions and extracted attributes, then trains a smaller, efficient model for production use. The teacher model, typically containing billions of parameters, can perform few-shot or zero-shot query understanding by leveraging its extensive pre-training, while the student model, with fewer parameters, achieves comparable performance while maintaining acceptable latency for real-time applications. The hybrid approach combining extractive and abstractive capabilities allows these systems to both identify explicit attributes mentioned in queries and infer implicit requirements based on contextual understanding [6]. This approach handles both broad queries like "healthy food" and long-tail requests like "low-sugar chocolate oat milk" consistently, extracting categories and attributes simultaneously while rewriting queries for better matching. The unified architecture ensures that all components of query understanding share the same contextual representation, enabling the model to resolve ambiguities by considering multiple signals jointly rather than making isolated decisions that might conflict with each other. The integration of attention mechanisms allows these models to weigh different parts of the query differently based on their importance for various understanding tasks [5].

Table 2: Ranking Systems Comparison - Traditional vs. Language Model Approaches [5, 6]

Ranking Component	Traditional Learning-to-Rank	LLM-Enhanced Ranking	Key Advantage
Feature Engineering	Hand-crafted (CTR, term frequency, position bias)	Automatic semantic signal generation	Captures deeper semantic relationships
Semantic Alignment	Limited - misses query-document intent alignment	Strong - contextualized relevance scoring	Fine-grained nuance detection
Infrastructure Integration	Requires a complete system redesign	Integrates with existing systems	No infrastructure overhaul needed

Long-tail Query Handling	Poor - limited training data	Strong - generalizes from pre-training	Handles rare/niche searches effectively
Domain Knowledge Encoding	Manual feature creation per domain	Automatic from pre-training	Complex linguistic patterns encoded
Relevance Assessment	Surface-level feature matching	Deep contextual understanding	More accurate across diverse queries
Architecture	Gradient boosting with manual features	Transformer-based contextualized representations	Captures semantic relationships
Training Data Dependency	High - needs extensive labeled data	Lower - leverages pre-training knowledge	Generalizes to unseen patterns
Query Type Coverage	Struggles with unusual queries	Broad coverage, including edge cases	Pre-training enables generalization

Natural Language Understanding and Catalog Intelligence

Language models excel at interpreting complex, conversational queries that traditional systems cannot parse. When someone asks for "a quick, healthy breakfast for two," the model identifies multiple elements: meal type, preparation time, dietary preference, and serving size. Traditional rule-based natural language understanding systems struggle with such multi-faceted queries because they lack the contextual reasoning needed to simultaneously extract and relate multiple constraints. Neural approaches to conversational information retrieval have fundamentally transformed how systems understand and respond to natural language queries by leveraging deep learning architectures that can capture semantic relationships and contextual dependencies across multiple utterances [7]. In session-based search, it can incorporate conversation history to infer implicit preferences, enabling more natural, context-aware interactions. The ability to maintain conversational context across multiple turns allows these models to understand references to previously mentioned items, refine searches based on feedback, and learn user preferences without requiring explicit restatement. Research on conversational information retrieval demonstrates that neural models can effectively track dialogue state, resolve coreferences, and interpret elliptical queries where users omit information they assume the system already knows from previous interactions [7]. The models employ attention mechanisms that weigh the relevance of previous utterances when processing current queries, enabling them to resolve ambiguities that would be impossible to address without conversational memory. These architectures process entire conversation histories jointly rather than treating each query in isolation, allowing them to build increasingly refined representations of user intent as dialogues progress and to distinguish between new information requirements and clarifications of existing requests.

Catalog intelligence sees similar improvements through multimodal understanding. Traditional systems struggle to extract product attributes when information appears across text and images. A tissue product might mention "3 packs" in its description but display "80 sheets" only on packaging. Language models that process both text and images can extract comprehensive structured attributes—size, flavor, dietary claims, computed values—even when details are fragmented. The integration of vision and language processing through unified multimodal architectures enables these systems to cross-reference information sources and compute derived attributes that require synthesis across modalities. Recent advances in dense video captioning and multimodal understanding have shown that models capable of processing temporal visual information alongside textual descriptions can extract rich, structured representations from complex multimedia content [8]. This dramatically improves metadata coverage and consistency, enhancing search quality, filtering accuracy, and personalization capabilities across large product catalogs. The models employ cross-modal attention mechanisms that allow visual features to inform text interpretation and vice versa, enabling them to resolve contradictions, fill gaps where one modality lacks information, and validate extracted attributes against multiple evidence sources. Research on dense captioning techniques demonstrates that multimodal transformers can generate detailed, temporally grounded descriptions by attending to both visual features and linguistic context, a capability that translates directly to product catalog enrichment, where attributes must be extracted from packaging images, product demonstrations, and textual specifications simultaneously [8]. By processing product images alongside textual metadata, these systems can identify attributes such as color variations, size comparisons visible in packaging, ingredient lists shown in nutritional labels, and usage scenarios depicted in product photography, creating richer and more accurate product representations that directly improve downstream search and recommendation performance.

Table 3: Natural Language Understanding - Traditional vs. Neural Conversational Systems [7, 8]

NLU Component	Traditional Rule-Based Systems	Neural LLM-Based Systems	Key Capability
Query Complexity Handling	Struggles with multi-faceted queries	Extracts multiple elements simultaneously	Contextual reasoning for constraints
Multi-element Extraction	Sequential, isolated processing	Parallel identification (meal type, time, dietary, size)	Semantic relationship capture
Session-Based Context	Limited or no history tracking	Full conversation history integration	Context-aware interactions
Dialogue State Tracking	Manual state management	Automatic dialogue state tracking	Neural architectures for dependencies
Coreference Resolution	Rule-based, often fails	Effective cross-utterance reference resolution	Attention-weighted previous utterances
Implicit Preference Learning	Cannot infer from history	Learns preferences across turns	Conversational memory mechanisms
Elliptical Query Handling	Fails without explicit information	Interprets omitted information from context	Assumes shared knowledge from dialogue
Query Processing Approach	Treats each query independently	Processes entire conversation jointly	Builds refined intent representations
Ambiguity Resolution	Limited without full context	Resolves using conversation history	Attention mechanisms on utterances
Intent Refinement	Static per query	Progressive refinement across dialogue	Distinguishes new vs. clarification requests

Expanding Discovery and Enabling Visual Search

Language models transform search from a purely matching exercise into a discovery engine. When someone searches for "sushi," traditional systems return matching products but miss complementary items like soy sauce, rice vinegar, or wasabi that complete the experience. Models can generate contextually relevant suggestions for complementary products, substitutes, or themed bundles, significantly expanding discovery opportunities. The application of semantic search combined with structured metadata filtering enables systems to understand the broader context of user queries and retrieve products that satisfy implicit intent rather than just explicit keyword matches. Research on query attribute modeling has demonstrated that combining semantic understanding with attribute-based filtering can substantially improve search relevance by capturing both the conceptual meaning of queries and the specific constraints users intend to apply [9]. The models accomplish this by reasoning about product usage contexts, culinary traditions, and typical consumption patterns encoded in their pre-training data, enabling them to suggest complementary items even for products with sparse co-purchase history. By modeling query attributes explicitly, these systems can decompose complex searches into semantic components and structured filters, allowing them to retrieve products that match the user's conceptual intent while respecting specific requirements like dietary restrictions, price ranges, or brand preferences [9]. This capability proves particularly valuable for expanding discovery in categories where users may not know what complementary items exist or where traditional "frequently bought together" signals are unavailable due to catalog sparsity or novelty.

Multimodal capabilities enable entirely new search paradigms. Users can search with images, screenshots, or combinations of visual and textual input. A photo of a sneaker paired with "find similar but more affordable" allows the system to reason about style, color, and design while applying price filters. Vision encoders work alongside language understanding to match queries against catalog embeddings, surfacing relevant items even when visual attributes aren't captured in text descriptions. Learning transferable visual models from natural language supervision has proven to be a breakthrough approach that enables systems to understand images through the lens of textual descriptions, creating models that can generalize across visual concepts without requiring explicit training examples for every category [10]. This proves particularly valuable for fashion, furniture, and home décor, where visual similarity matters as much as textual specifications. The integration of contrastive learning approaches enables these systems to learn joint embeddings where images and text descriptions of the same product cluster together in a shared semantic space, while visually or conceptually similar items from different products are positioned nearby. Research has shown that pre-training vision models on large-scale image-text pairs using contrastive objectives allows the models to develop robust visual representations that align with natural language concepts, enabling zero-shot transfer to new visual recognition tasks without fine-tuning [10]. The models employ dual-encoder architectures that process images and text independently before computing similarity in a shared embedding space, allowing efficient retrieval at scale while maintaining the flexibility to handle arbitrary combinations of visual and textual query components. The natural language supervision approach means these models can understand complex visual concepts described in text, such as "art deco furniture with geometric patterns" or "minimalist sneakers with chunky soles," and match them against product images even when such specific attributes aren't explicitly tagged in the catalog metadata [10].

Table 4: Search Discovery Capabilities - Traditional Matching vs. LLM-Based Discovery Engines [9, 10]

Discovery Feature	Traditional Search Systems	LLM-Based Discovery Systems	Key Enhancement
Search Paradigm	Pure keyword matching	Context-aware discovery engine	Understands broader context
Complementary Products	Misses related items (e.g., soy sauce for sushi)	Suggests contextually relevant complements	Usage context reasoning
Intent Understanding	Explicit keywords only	Implicit intent recognition	Semantic + metadata filtering
Product Relationship Reasoning	Limited to co-purchase history	Reasons about usage contexts and traditions	Pre-training knowledge encoding
Substitute Suggestions	Not available or basic	Intelligent substitute recommendations	Conceptual understanding
Themed Bundle Creation	Manual curation only	Automatic contextual bundling	Culinary/usage pattern knowledge
Query Decomposition	Keyword-based only	Semantic components + structured filters	Attribute modeling

Sparse Catalog Handling	Fails without co-purchase data	Works with sparse history	Semantic reasoning
Constraint Application	Basic filtering	Conceptual intent + specific constraints	Dietary, price, brand integration
Discovery in New Categories	Limited without historical data	Strong generalization capability	Pre-trained knowledge transfer
Complex Search Handling	Struggles with multi-faceted queries	Decomposes and satisfies all components	Query attribute modeling

Conclusion

Large language models are a radical rethink of e-commerce search, not a simple enhancement of the existing systems. These models overcome longstanding restrictions that have bedeviled traditional keyword-based search methods by introducing veritable language comprehension, multimodal reasoning, and generation capabilities to all the members of the search stack. They bridge vocabulary divides by semantic embedding representations, integrate fragmented query comprehension systems into consistent architectures, achieve more information-rich catalog intelligence by synthesizing information between text and visual modalities, and support visual search paradigms that integrate image and text cognition. This change of superficial word recognition to deeper contextual understanding enables systems to support the overall complexity of natural language, conversational queries with implicit constraints, long-tail searches with sparse training data, and multi-faceted queries that need simultaneous extraction of multiple semantic items. With the continuous development of transformer architectures and the elaboration of pre-training methods, the demarcation between search and conversation is becoming increasingly blurred and provides shopping experiences capable of understanding not only what users are typing or displaying in images, but what they actually want to discover. This change allows retailers to appear with the appropriate products in the case where conventional systems will break down entirely, minimizes consumer aggravation by interpreting intentions more precisely, and opens up entirely new horizons of product discovery that can generate engagement and convert in ways never achievable by lexical matching.

References

[1] Yi Chang et al., “Query Understanding for Search Engines,” ResearchGate, January 2020. Available: https://www.researchgate.net/publication/347446452_Query_Understanding_for_Search_Engines

[2] Kalinka Jasinska et al., "BERT-based similarity learning for product matching," December 2020, ResearchGate. Available: https://www.researchgate.net/publication/350871039_BERT-based_similarity_learning_for_product_matching

[3] Jimmy Lin et al., "Pretrained Transformers for Text Ranking: BERT and Beyond," October 2020, ResearchGate. Available: https://www.researchgate.net/publication/344639090_Pretrained_Transformers_for_Text_Ranking_BERT_and_Beyond

[4] Lukas Gienapp et al., "Learning Effective Representations for Retrieval Using Self-Distillation with Adaptive Relevance Margins," July 2024, ResearchGate. Available: https://www.researchgate.net/publication/382739302_Learning_Effective_Representations_for_Retrieval_Using_Self-Distillation_with_Adaptive_Relevance_Margins

[5] Jin Wang et al., "EDCEW-LLM: Error detection and correction in English writing: A large language model-based approach," A. Hassan et al., October 2025, ScienceDirect. Available: https://www.sciencedirect.com/science/article/pii/S1110016825008750

[6] Divya S et al., "Unified extractive-abstractive summarization: a hybrid approach utilizing BERT and transformer models for enhanced document summarization," November 2024, ResearchGate. Available: https://www.researchgate.net/publication/385934152_Unified_extractive-

abstractive_summarization_a_hybrid_approach_utilizing_BERT_and_transformer_models_for_enhanced_document_summarization
[7] Jianfeng Gao et al., "Neural Approaches to Conversational Information Retrieval," January 2023, ResearchGate. Available:
https://www.researchgate.net/publication/369311610_Neural_Approaches_to_Conversational_Information_Retrieval
[8] Iqra Quasim et al., "Dense Video Captioning: A Survey of Techniques, Datasets, and Evaluation Protocols," January 2025, ResearchGate. Available:
https://www.researchgate.net/publication/388015109_Dense_Video_Captioning_A_Survey_of_Techniques_Datasets_and_Evaluation_Protocols
[9] Karthik Menon et al., "Query Attribute Modeling: Improving search relevance with Semantic Search and Meta Data Filtering," August 2025, ResearchGate. Available:
https://www.researchgate.net/publication/394362651_Query_Attribute_Modeling_Improving_search_relevance_with_Semantic_Search_and_Meta_Data_Filtering
[10] Alec Ladford et al., "Learning Transferable Visual Models From Natural Language Supervision," ResearchGate, February 2021. Available:
https://www.researchgate.net/publication/349704314_Learning_Transferable_Visual_Models_From_Natural_Language_Supervision