Enterprise Strategy Group by TechTarget

SHOWCASE

NOVEMBER 2024

# How Generative AI Is Changing E-commerce

Mark Beccue, Principal Analyst

**Abstract:** Artificial intelligence has been used to deliver superior e-commerce experiences for more than a decade. Generative AI has the potential to significantly level up customer experiences, specifically for search, recommendations, and personalizing user experiences, but the market is very much in the early exploratory stage. How can generative AI transform search, recommendations, and personalization? There are roadblocks, but with the right partner, e-commerce retailers can transform their customer engagement.

## Introduction: Generative AI's Potential Has E-commerce Buzzing

In the two short years since ChatGPT was launched, generative AI has inspired organizations across all types of industries to dream big. Unlike classical AI, which was designed to identify patterns in data and produce predictions and analysis, generative AI models can create new content. While classical AI requires organizations to curate and develop their own data sets, generative AI enables organizations to leverage massive, collective data sets as well as their own data. To tap classical AI, organizations typically need significant data science or data engineering expertise to construct and manage models, whereas with generative AI, data science experience is less of a requirement.

Importantly, generative AI is no longer just a concept; it's a technology that is quickly reaching the mass market adoption phase. Research from TechTarget's Enterprise Strategy Group found 30% of respondents have generative AI use cases in production, with 8% describing the production phase as "mature" and 22% describing it as "early."[1]

Retail organizations, including e-commerce retailers, have high hopes for generative AI. Take, for instance, how retailers expect to benefit from AI: 30% reported that they expect AI to improve customer conversations/interactions, 30% that they expect improved operational efficiency, 28% that they expect increased customer trust, and 22% that they expect improved customer experience.[2]

### Market Insight

41% of retail respondents are pursuing or planning to pursue AI-fueled product recommendations use cases.

Use cases for AI in retail are well underway. Supply chain management/optimization is the most prevalent AI use case retailers surveyed by Enterprise Strategy Group are currently pursuing or planning to pursue (cited by 50% of retail respondents). Other popular use cases being pursued included personalized marketing (48%), chatbots and virtual assistants (45%), and product recommendations (41%).[3]

[1] Source: Enterprise Strategy Group Complete Survey Results, *The State of the Generative AI Market: Widespread Transformation Continues*, September 2024.
[2] Source: Enterprise Strategy Group Complete Survey Results, *The State of Analytics and Business Intelligence Platforms: Infusing AI into Analytics*, April 2024.
[3] Ibid.

For e-commerce retailers, there is a lot of excitement in particular around what generative AI can potentially do to transform customer engagement services—specifically search, recommendations, and personalization:

- **Generative AI in search.** Retailers see the potential to improve accuracy and relevance of customer searches by understanding the intent of queries. Further, there is potential to generate answers or suggest relevant search results based on improved understanding of queries. For example, a search for "best laptops for college students" might not only show laptops but also include generated recommendations based on price range, user reviews, and specific needs (e.g., battery life, portability).
- **Generative AI in recommendations.** Instead of merely matching users with items based on predefined rules or past interactions, large language models (LLMs) have the potential to analyze a broader context from user behavior, preferences, and language inputs. For example, generative AI can analyze a user's shopping pattern and create tailored product suggestions that match, not just past purchases, but also inferred preferences based on similar users or contextual data.
- **Generative AI in personalization.** LLMs excel in personalizing a user's experience because of their ability to process and understand natural language inputs. Because of this, they can adapt content, messages, or offers for individual users by generating outputs that reflect the user's specific needs. For example, after analyzing a shopper's query, history, and preferences, generative AI might suggest uniquely bundled products or create custom promotions.

## The Problem: Generative AI Scalability and Accuracy/Performance Tradeoffs Thwart Efforts To Transform E-commerce Customer Engagement

While e-commerce retailers are very interested in leveraging generative AI to boost search, recommendations, and personalization, obstacles abound. The key challenges include:

- **Accuracy challenges for LLMs.** Because e-commerce search, recommendations, and personalization are so unique to each user, accuracy of such deliverables has to be spot-on. As much upside potential as they might represent, LLMs are notorious hallucinators, not only returning inaccurate responses, but also being highly confident responses are correct.
- **Inability to apply ML at production scale.** Many e-commerce retailers use architectures where simple retrieval is combined with a separate machine learning (ML)-based re-ranker or re-scorer service. While this architecture provides a clear separation of concerns, the downside is that it is unable to apply AI in a way that scales with data and request rates since it is bottlenecked on passing information to the separate re-ranker service. This problem has become more severe with the addition of vectors as these represent a one to two orders of magnitude increase in data volume for ranking.
- **Accuracy/performance tradeoffs.** Inability to apply ML to data at scale is typically sidestepped by only applying AI for relevance and personalized recommendations (scoring) to a small fraction of the result selected by simple methods. This limits the quality achieved by applying AI, ultimately leading to lost revenue.

## The Right Approach: Distributed Architecture, Hybrid Search To Deliver Scale, Accuracy

To address the key challenges to generative AI-fueled customer engagement services, e-commerce retailers should consider the following recommendations:

- **LLM accuracy.** Challenges should be addressed through a growing and evolving corpus of model management best practices. These include customizing models (fine-tuning, retrieval-augmented

generation [RAG], and prompt engineering), leveraging LLM built-in safety systems, deploying human-in-the-loop backstops, and developing and implementing a comprehensive AI risk management/governance framework.

- **Scale.** Distributed architectures are a better approach than centralized and non-distributed architectures for achieving scale. They enable both vertical and horizontal scaling, providing the flexibility to increase capacity by distributing data, queries, and AI models across multiple nodes.
- **Reducing accuracy/performance tradeoffs.** Search accuracy improves through hybrid search, which combines multiple data types, including vectors, text, and structured as well as unstructured data. Hybrid search balances the precision of keyword searches with the broader relevance captured by embeddings or ML-based rankings. For example, a search result can be ranked based on both exact keyword matches and semantic closeness determined by a vector search model. Distributed architecture also plays a role in reducing accuracy/performance tradeoffs because it ensures fault tolerance.

## Driving Success With Vespa.ai

Vespa.ai is a platform for applications that leverage AI and data to deliver experiences to end users online in real time, such as search, recommendation, personalization, and RAG. It has a sterling track record in delivering these customer engagement services to well-known retail brands, including some of the largest retailers in the US. Vespa.ai's high-performance architecture, capable of scaling to over 100,000 queries per second, empowers retailers to deliver more engaging shopping experiences, ultimately driving increased revenue.

Vespa.ai takes a best-in-class, enterprise-grade approach to delivering e-commerce customer engagement services:

- **High performance at scale.** Vespa's optimized low-latency query execution, real-time data updates, and advanced ranking algorithms enable sophisticated and efficient searches.
- **Search accuracy.** Vespa supports precise hybrid search combining multiple data types, including vectors, text, and structured as well as unstructured data. ML ranks and scores results, ensuring relevance to the user.
- **Elasticity for fluctuating demands.** Vespa supports horizontal and vertical scaling, dynamically adding and removing hardware capacity to optimize cost while ensuring low latency.
- **Adaptability.** Vespa is highly customizable and offers APIs as well as SDKs for seamless integration with various applications and data sources, enabling users to build applications specific to their requirements.
- **Resiliency.** By distributing data, queries, and ML models across multiple nodes, Vespa ensures scalability and fault tolerance, which are crucial for large-scale deployments.
- **RAG.** Vespa supports RAG, which enhances LLMs in generative AI by integrating external knowledge retrieval with text generation. This enables more accurate, up-to-date, and contextually relevant responses.

Vespa has been leveraging AI in e-commerce for years. Here are a few examples of how Vespa's AI expertise matters for e-commerce:

- **ML model integration for ranking.** Vespa runs trained ML models, such as those created with TensorFlow, PyTorch, or XGBoost, directly within its search and query pipelines. This helps improve the relevance of results by applying advanced ranking models that use learned embeddings or features.
- **Vector nearest neighbor search.** Vespa is a pioneer in supporting advanced vector search, an ML-driven technique that enables the searching and ranking of high-dimensional vectors (embeddings). These

vectors are typically derived from ML models trained to capture the semantic meaning of data (e.g., text, images).

- **Feature engineering and feature stores.** Vespa allows real-time feature computation at query time, critical for ML models that need to access dynamic features, such as user behavior or contextual data. These features can be fed into the models to adjust search rankings or recommendations.
- **Online learning and continuous updates.** Vespa facilitates real-time updates to ML models, data pipelines, ranking, and application logic. It can support models that need continuous updates based on user interactions (e.g., recommendation systems that adapt to user preferences over time). Vespa can serve models that are updated regularly without needing to pause or restart the entire system.
- **Hybrid search (combining text and ML).** Vespa combines traditional keyword-based and ML-driven vector searches. This hybrid search approach balances the precision of keyword searches with the broader relevance captured by embeddings or ML-based rankings. For example, a search result can be ranked based on both exact keyword matches and semantic closeness determined by a vector search model.
- **Recommendations and personalization.** Vespa uses ML models to encode items and users as vectors. Based on user preferences, Vespa can recommend content or products. The platform supports real-time querying and ranking, which enables fast and relevant personalized recommendations.
- **Structured navigation.** Vespa can group all the matches to a query or all recommended content hierarchically and aggregate values over the groups to provide user aids such as navigation in a product catalog, seamlessly integrated with search and recommendation.

## Conclusion

The retail industry is under continued, significant transformation. Customer engagement services are one of the cornerstones of retail success. Generative AI is poised to amplify competitive advantages for e-commerce players savvy enough to leverage it in search, recommendations, and personalization. However, scaling solutions and accuracy/performance tradeoffs can significantly hamper generative AI-infused solutions.

It makes sense to find and team up with AI-savvy, experienced search, recommendation, and personalization experts that have the technical knowledge and proven track record to help retailers move from ideas to results. Vespa has developed solutions designed to deliver on the enormous potential generative AI has in order to power retailers to greater engagement and, ultimately, more revenue.



**About Enterprise Strategy Group**
TechTarget's Enterprise Strategy Group provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

contact@esg-global.com
www.esg-global.com