Adaptive E-Commerce Search Optimization Using Large Language Models and Dynamic User Interest Memory Networks

Nicolas Remos
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
nicolasramos@colostate.edu

Arjun C. Tripathi
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
contactarjun@ucf.edu

Rainer Fox
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
rainermail@ku.edu

Abstract

The rapid growth of e-commerce platforms has made product search a critical interface between consumers and vast digital catalogs, yet traditional retrieval and ranking pipelines struggle to capture the fluid and context-dependent nature of user intent. This paper presents a comprehensive system-level investigation of adaptive e-commerce search optimization that integrates large language models with dynamic user interest memory networks. We examine the architectural fusion of transformer-based language understanding with dynamically updating memory components that continuously encode short-term and long-term user preferences, session context, and cross-session behavioral signals. The discussion is organized around structural trade-offs, infrastructure design, deployment considerations, and governance challenges, including fairness, robustness, privacy, and sustainability. By situating the proposed approach within the broader evolution from feature-based learning to rank and neural retrieval models, we analyze how memory-augmented large language models can enable more coherent and personalized query matching and intent modeling. The paper does not present a single system implementation but instead develops a holistic research perspective that connects advances in dense retrieval, interest evolution networks, and zero-shot ranking capabilities with the systemic demands of production-scale e-commerce environments. Throughout, we emphasize the interplay between model expressiveness and operational constraints, the necessity of continuous model updating to combat concept drift, the policy implications of profiling user memory, and the environmental costs of serving large generative models at scale. The analysis is grounded in recent literature and provides forward-looking perspectives on building inclusive, privacy-preserving, and energy-efficient search systems that respect user autonomy while delivering high-quality discovery experiences.

Keywords

e-commerce search, large language models, user memory networks, personalization, information retrieval, fairness, system architecture.

1. Introduction

E-commerce search engines serve as the primary digital gateway through which millions of users daily navigate product catalogs, express commercial intent, and make purchasing decisions. The performance of these systems directly impacts customer satisfaction, retention, and platform revenue. Historically, the field of information retrieval has progressed through distinct paradigm shifts, beginning with lexical matching and moving toward learning to rank models that combine hundreds of handcrafted features to score query-document pairs [1]. While these approaches brought substantial improvements over purely statistical retrieval functions, they remained limited by their reliance on feature engineering and their limited capacity to model the deeper semantic and contextual dimensions of user queries. The subsequent rise of neural ranking models, which leverage distributed representations learned end-to-end from interaction data, marked a significant leap forward by enabling the capture of latent semantic similarities and query-document relevance patterns that were previously opaque [2]. Early neural approaches, however, often treated each query in isolation, disregarding the rich sequential and historical signals that characterize real user behavior on e-commerce platforms.

The evolution of neural architectures toward deeper and more capable models has been further accelerated by the introduction of pre-trained language models based on the transformer architecture, which have fundamentally reshaped natural language processing and information retrieval research [3]. These models, pre-trained on massive text corpora through self-supervised objectives, can be fine-tuned to excel at understanding complex query formulations, product descriptions, and the nuanced relationships between them [4]. Among pre-trained models, BERT and its variants demonstrated that bidirectional context modeling yields substantial gains on ranking benchmarks, while the scaling of generative pre-trained transformers to hundreds of billions of parameters gave rise to the era of large language models capable of few-shot and zero-shot generalization [5, 6]. The deployment of these models in e-commerce search opens new possibilities for intent understanding, query reformulation, and result summarization, yet simultaneously raises fundamental questions about latency, cost, consistency, and user trust.

Parallel to the developments in language modeling, research on recommendation and click-through rate prediction has produced a family of interest network architectures specifically designed to extract evolving user preferences from long and noisy behavioral sequences. The deep interest network introduced the concept of adaptively activating relevant historical interactions given a target item, thus moving beyond fixed user embedding vectors [7]. Subsequent advancements, such as the deep interest evolution network, incorporated an explicit evolution module that captures how interests shift over time, while search-based interest models pushed the frontier further by enabling the retrieval of long-term life-cycle behaviors through a two-stage attention mechanism [8, 9, 10]. These architectures are inherently dynamic because they do not compress a user into a static representation but instead compute on-the-fly interest vectors conditioned on the current query context and the entire behavioral trajectory. Despite their successes, they were originally designed to operate within recommendation pipelines and relied on fixed feature vocabularies, lacking the semantic generalization power that large language models provide.

E-commerce search presents distinct challenges that differentiate it from general web search and even from recommendation. Queries are often extremely short, ambiguous, and imbued with implicit purchasing intent that may change rapidly across a single session. A user searching for “lightweight running shoes” in the morning might shift to “trail running shoes

waterproof" by evening, and without memory of the morning session, the search system loses critical signal about the user's evolving intent. Existing e-commerce search evaluation benchmarks, such as the large-scale shopping queries dataset, highlight the need for fine-grained relevance judgments that capture exact, substitute, complement, and irrelevant relationships, underscoring the multi-faceted nature of product search [11]. In this landscape, the integration of large language models with dynamic user interest memory networks promises a unified framework in which language understanding, personalization, and memory are jointly optimized. This paper explores such an integration from a systemic viewpoint, focusing on how memory-augmented large language models can be architected, deployed, and governed in real-world e-commerce platforms while balancing expressivity with efficiency, personalization with privacy, and innovation with societal accountability. The work by Yu [12] has recently outlined a conceptual framework for employing LLM-based user memory systems to enhance query matching and intent modeling, providing initial evidence that persistent, learnable user memory can significantly improve search relevance over stateless baselines. Building on this line of inquiry, the present analysis extends the discussion into the broader systems, infrastructure, and policy dimensions that determine whether such architectures can be responsibly operationalized at scale.

2. Background and Motivation

To appreciate the architectural innovations that underpin adaptive e-commerce search, it is important to understand the layered evolution of retrieval and ranking technologies and the constraints imposed by production environments. Classical search engines relied on inverted indices, term frequency-inverse document frequency scoring, and probabilistic models that operated on lexical match signals. The introduction of learning to rank enabled the incorporation of a wide array of features, including query-independent document priors, historical click-through rates, and various forms of user context, but the models were ultimately limited by the manual curation of feature sets and their inability to capture deep semantic relationships across vocabularies [1]. The advent of neural ranking models based on convolutional and recurrent architectures allowed for the learning of query-document similarity functions directly from raw text, bypassing some of the feature engineering bottleneck. Yet these early neural models were data-hungry and often required large amounts of human relevance labels, a scarce resource in many e-commerce settings. The use of weak supervision, where models are trained on user click logs used as noisy relevance signals, emerged as a practical workaround that connected neural ranking with the scale of behavioral data [2].

The transformer architecture fundamentally altered this landscape by introducing self-attention mechanisms that could model arbitrary dependencies across input sequences, eliminating the sequential bottlenecks of recurrent models and enabling large-scale parallelization [4]. Pre-trained transformer language models, when fine-tuned on in-domain ranking data, were shown to substantially outperform prior neural and feature-based rankers on a range of ad-hoc retrieval tasks, including product search [3, 5]. The generalization capabilities of these models meant that they could handle tail queries and previously unseen product descriptions more gracefully than models that relied on term-matching statistics. The scaling of model size to hundreds of billions of parameters further revealed emergent capabilities, such as the ability to perform passage re-ranking and query intent classification with only a handful of demonstrations in the prompt [6]. These generalist models can, in principle, be used directly as rankers through prompting strategies, although their

computational cost and latency remain prohibitive for real-time query-serving in high-traffic e-commerce environments unless heavily optimized through distillation, quantization, or hybrid retrieval pipelines.

At the same time, the recommendation community developed specialized architectures for user behavior modeling that are highly relevant to search personalization. The deep interest network addressed the limitation of fixed-dimensional user representation by adaptively weighting historical interactions based on their relevance to a candidate item, effectively implementing a form of soft memory retrieval [8]. The deep interest evolution network extended this by introducing an interest extraction layer that uses auxiliary losses to supervise the learning of evolving interest states, making the model more sensitive to the sequential nature of user actions [9]. The search-based interest model further recognized that for long user sequences spanning months or years, only a fraction of historical behaviors are relevant to any given session, and proposed using a sub-model to retrieve the top-k relevant long-term behaviors which are then fed into an attention mechanism [10]. These architectures, while designed for click-through rate prediction in recommendation, provide a conceptual template for the dynamic memory component of an adaptive search system. In e-commerce search, the analogous requirement is to maintain a memory of past queries, clicks, add-to-carts, and purchases, and to selectively retrieve and attend to those memories when processing a new query.

The availability of specialized datasets has catalyzed progress by providing common benchmarks that reflect the unique properties of e-commerce queries and product catalogs. The shopping queries dataset introduced by a major e-commerce platform offers a large collection of query-product pairs with graded relevance judgments including exact, substitute, complement, and irrelevant labels, capturing the subtleties of product substitutability and complementarity that are central to purchase decisions [11]. These distinctions are important because a user searching for a specific phone model may be satisfied with a close substitute if the exact match is out of stock, whereas showing a complementary accessory as the top result would be a failure. Such nuanced relevance relationships demand that search systems go beyond simple topical similarity and incorporate user context, inventory state, and price sensitivity. The recent work on LLM-based user memory systems explicitly tackles this challenge by maintaining persistent user profiles that can condition both query understanding and result ranking, thereby enabling more coherent intent modeling across sessions [12]. The preliminary results indicate that memory-augmented models can achieve higher precision on ambiguous queries and reduce the number of reformulations needed to find a satisfactory result, suggesting a promising direction for the next generation of e-commerce search engines.

3. Large Language Models in E-Commerce Search

The integration of large language models into e-commerce search is not a straightforward plug-and-play endeavor but rather a complex systems design problem that touches every component of the retrieval and ranking stack. At the core, LLMs bring a profound capacity for semantic understanding, enabling the system to accurately map free-text queries to product attributes and categories even when the vocabulary differs significantly. For instance, a query like “something breezy for summer hiking” contains no exact product terms, yet an LLM can infer the intent to find lightweight, breathable hiking apparel suitable for warm weather. This level of understanding is difficult to achieve with traditional entity extraction and category matching pipelines. Dense retrieval methods such as DPR leverage these semantic capabilities by encoding queries and documents into a shared dense vector space

where relevance is computed via inner product, allowing for efficient approximate nearest neighbor search [13]. When combined with a two-stage retrieve-and-rerank architecture, dense retrieval can serve as a scalable first stage that benefits from LLM-based query encoding while the computationally heavier LLM reranker is applied only to a small candidate set.

The architecture of LLM-based search components can be conceptualized along the axes of when and how the language model participates in the ranking process. In the encoding-only paradigm, the LLM is used as a text encoder to produce query and document representations, which are then scored using a lightweight similarity function such as cosine distance. This approach is amenable to pre-computation of document embeddings and incremental indexing, thereby reducing online serving cost. In a generative reranking paradigm, the LLM directly produces a relevance score or an explanation for each query-document pair, often through a tailored prompt template that includes the query, product title, description, and possibly user history. While this can yield state-of-the-art accuracy, it multiplies the serving cost by the number of candidate documents and is incompatible with strict latency budgets unless model distillation, speculation, or hardware acceleration is employed. A hybrid design that uses a smaller fine-tuned student model for candidate scoring and reserves the full LLM inference for the top few results or for challenging queries is a pragmatic compromise that many platforms are exploring.

The dynamic nature of e-commerce catalogs presents additional challenges that are often underappreciated in static retrieval benchmarks. New products are continuously added, inventory statuses change, and seasonal trends shift rapidly. Large language models that are fine-tuned on a historical snapshot or are frozen after pre-training can become stale, exhibiting degraded performance on out-of-distribution queries or failing to reflect up-to-date catalog knowledge. This creates a tension between leveraging the deep semantic priors encoded in pre-trained LLMs and the need for continuous alignment with the current catalog and user behavior. Potential solutions include retrieval-augmented generation where the LLM is supplied with up-to-date product knowledge at inference time, incremental fine-tuning on a stream of recent interactions, and the use of adapter modules that allow lightweight model updating without full retraining. None of these solutions is without cost, and the choice among them involves trade-offs between accuracy, freshness, engineering complexity, and the risk of catastrophic forgetting.

Beyond relevance ranking, LLMs enable new search experiences that blur the line between search and conversation. In a conversational search scenario, the user engages in a multi-turn dialog with the search assistant, refining their specifications, comparing products, and asking follow-up questions. Maintaining a coherent discourse state across turns requires that the LLM have access to the interaction history in an efficiently encoded form, which naturally points to the need for a dynamic memory module. The LLM can also generate product summaries, pros and cons lists, and personalized explanations based on the user's identified preferences, provided those preferences are reliably modeled and persistently stored. However, the generation of fluent and persuasive text raises risks of hallucination, where the model invents product features or makes unsupported claims, and of amplifying biases present in the training data, such as disproportionately recommending expensive products or stereotyping based on demographic proxies. The systems community must therefore develop robust evaluation frameworks that go beyond offline relevance metrics and include factuality, safety, and fairness dimensions measured through both automated and human evaluation protocols.

4. Dynamic User Interest Memory Networks

The concept of a dynamic user interest memory network is central to the vision of an adaptive e-commerce search system that treats each user as a continuously evolving entity rather than a collection of static attributes. At a high level, the memory network is a learnable component that reads from and writes to a user-specific state, encoding the user's historical interactions, declared preferences, inferred intents, and contextual signals into a structured representation that can be efficiently accessed during query processing. The design of such a memory can draw inspiration from end-to-end memory networks, which were among the first architectures to demonstrate that recurrent reading and writing to an external memory array can be learned end-to-end via gradient-based optimization [14]. In the context of e-commerce, the memory must handle heterogeneous interaction types including clicks, purchases, dwell times, price filters, brand preferences, and even returns, distilling these into a compact set of interest vectors that are actionable for retrieval and ranking.

A salient structural distinction is between fixed-size user embeddings and content-addressable memory stores that grow with the user's history. The former approach compresses the entire user history into a single dense vector, which is computationally efficient but loses the ability to recall specific past interactions that might be acutely relevant to a new query. Content-addressable memory, in contrast, maintains a collection of historical event representations that are indexed by similarity to the current query. When the user issues a new query, the memory network retrieves the most relevant past interactions and their associated product embeddings, aggregates them using attention, and produces a query-specific user context vector that is concatenated with the query encoding before scoring. This architecture mirrors the design principles established by deep interest networks but extends them with the semantic generalization of large language model encodings and the ability to store multimodal signals such as images and structured attributes.

The temporal dynamics of user interests demand that the memory network support both short-term session memory and long-term cross-session memory. Short-term memory captures the immediate sequence of actions within a single browsing session, such as a series of search refinements and product views that progressively clarify the user's intent. This memory must be rapidly updateable and have a relatively short decay window, as the user's focus may shift dramatically between sessions. Long-term memory retains information about the user's stable preferences, such as brand loyalties, size constraints, price sensitivity, and lifestyle categories, which persist over months or years. The search-based interest model demonstrated in recommendation that distinguishing between short-term and long-term behavioral sequences and retrieving relevant long-term memories using a specialized search mechanism can significantly improve predictive accuracy [10]. Adapting such a two-tier design for e-commerce search requires additional considerations, such as how to handle ephemeral purchase intents versus persistent shopping patterns, and how to manage the trade-off between recommendation-oriented browsing and directed purchase search.

The training of dynamic memory networks in a search setting raises unique challenges related to data sparsity, label noise, and the cold-start problem. Unlike recommendation, where implicit feedback is abundant, search interactions are often sparser and more targeted. A user may issue only a handful of queries in a session, and relevance labels are typically collected only for a small fraction of query-document pairs through click modeling or human annotation. Memory networks must therefore be trained with self-supervised objectives that pre-train the memory encoders on large volumes of unlabeled behavioral sequences before

fine-tuning on scarce relevance signals. Techniques such as next-query prediction, session continuity modeling, and contrastive learning between consecutive sessions can provide useful auxiliary tasks that help the memory extract meaningful interest trajectories without requiring explicit labels. Additionally, the cold-start problem, where a new user has no interaction history, must be addressed by initializing the memory from aggregated population priors and rapidly adapting it as the first few interactions arrive. This online adaptation capability is crucial for maintaining a personalized experience from the very first session, and it necessitates efficient incremental learning algorithms that can update user memory states without full recomputation from scratch.

The memory network also serves as a natural locus for implementing privacy controls and data minimization. Instead of storing raw interaction logs, the memory can be designed to store only abstracted, aggregated interest representations that are insufficient to reconstruct specific items or sessions. Differential privacy mechanisms can be applied during the memory update step to ensure that the inclusion or exclusion of any single interaction has a bounded influence on the stored representation. The tension between personalization accuracy and privacy preservation is well-documented, and prior work has shown that even supposedly anonymized behavioral datasets can be vulnerable to deanonymization attacks when correlated with auxiliary information [16]. Thus, the architectural design of the memory must incorporate defense-in-depth strategies, including user-controllable memory erasure, time-limited retention policies, and on-device memory computation where the user interest profile is stored and processed locally, with only aggregated intent signals shared with the search backend. These considerations elevate the memory network from a purely performance-enhancing component to a critical element of a trustworthy and regulatory-compliant system architecture, aligning with the principles of privacy by design that are increasingly mandated by data protection regulations globally.

5. System Architecture and Integration

Bringing together large language models and dynamic user interest memory networks into a coherent end-to-end search system requires careful orchestration across retrieval, ranking, and serving layers, with each layer introducing distinct requirements for latency, throughput, consistency, and fault tolerance. The canonical architecture begins at the query understanding stage, where the raw user query is enriched by the memory network to produce an intent-augmented representation. This augmented query embedding is then dispatched to a multi-stage retrieval pipeline. The first stage, typically an approximate nearest neighbor search over a dense vector index, retrieves a set of several hundred candidates based on query-product embedding similarity [13]. The second stage, a lighter-weight scoring model that may be a distilled version of the ranking LLM, reduces the candidate set to a few dozen high-relevance items while incorporating real-time business rules such as inventory filtering and price constraints. The final stage applies the full LLM-based reranker, now armed with the detailed user memory context, to score and order the remaining candidates, and optionally generate personalized explanations or promotional snippets. This staged funnel is essential because a full LLM inference on every product in a catalog of millions is economically and temporally infeasible even with massive hardware acceleration.

The integration of memory into this pipeline raises several systems-level design decisions. One fundamental question is whether the memory network is part of the online serving path or operates as an offline feature generation system that pre-computes user interest vectors which are then joined at query time. The online approach allows the memory to reflect the

very latest user actions, including those within the current session, but demands that memory reads and writes execute within tight latency budgets—typically single-digit milliseconds for end-to-end query processing. The offline approach simplifies online serving but can introduce staleness and prevents the model from reacting to session-in-progress intent shifts. A promising compromise is to maintain a short-term session memory online, as a lightweight key-value store with rapid access, while relegating the heavier long-term memory retrieval and aggregation to an offline or nearline pipeline that updates user interest vectors asynchronously. This design mirrors the online-offline consistency trade-offs widely studied in large-scale recommender systems, where feature freshness must be balanced against computational cost.

Infrastructure considerations extend to the management of model serving for LLMs. Deploying a large transformer model in a real-time, high-traffic setting demands sophisticated serving frameworks that support model parallelism, dynamic batching, and adaptive precision. Techniques such as quantization to int8 or int4, token-level speculative decoding, and kernel fusion can reduce per-token latency substantially, but each introduces potential accuracy regressions that must be validated against the e-commerce relevance metrics. Moreover, the memory network itself can be implemented as a separate microservice or co-located with the ranking model, depending on the coupling of query augmentation and scoring. Tightly coupled architectures where the memory read produces an attention bias that directly influences the LLM's self-attention computation may yield the highest accuracy but require end-to-end training and joint deployment, whereas a loosely coupled solution that passes memory summaries as additional input tokens is more modular and easier to iterate on but may underutilize the memory signals given the LLM's limited context window. The choice between these integration patterns is driven by the team's velocity, the maturity of the model serving infrastructure, and the acceptable level of cross-team dependency.

Given the scale of modern e-commerce platforms, which can serve billions of queries per day across multiple regions, the system design must also address global distribution and data locality. Privacy regulations may prevent user behavioral data from leaving its country of origin, requiring that the memory network be trained and served on regional clusters, with only model gradients or encrypted aggregates propagated to a global model coordinator. Federated learning and secure aggregation protocols offer pathways to achieve global model improvements while respecting data sovereignty, though they introduce additional engineering complexity and can slow down model convergence. Similarly, the retrieval index must be sharded and replicated across regions to ensure low latency for all users, and the inventory-level signals needed for filtering must be propagated with low delay. These distributed systems concerns are not merely implementation details but fundamentally shape the architectural choices: a design that appears compelling in a single-datacenter research setup can become untenable when faced with the full scale and regulatory patchwork of a global deployment.

6. Structural Trade-Offs and Deployment Considerations

Every decision in the design of an adaptive LLM-powered search system with user memory is situated within a multidimensional trade-off space defined by accuracy, latency, cost, fairness, and maintainability. One of the most salient trade-offs involves the granularity and longevity of the memory stored per user. Storing detailed, high-resolution event sequences maximizes the potential for accurate personalization but also increases storage cost, retrieval latency, and privacy exposure. Conversely, compressing user histories into fixed-size interest vectors

reduces these burdens but risks losing rare but decisive preference signals, such as a one-time purchase of a specialized tool that indicates a niche hobby. The memory architecture must therefore implement a lossy compression scheme that is informed by the downstream ranking objective, ensuring that information that is discriminative for future intent is preserved while redundant or outdated signals are discarded. This is formally a challenging representation learning problem that sits at the intersection of information theory and continual learning.

Another critical structural trade-off arises from the tension between model specialization and generalization. E-commerce encompasses diverse verticals—electronics, fashion, groceries, home goods—each with its own vocabulary, user behavior patterns, and relevance criteria. Training a single monolithic LLM and memory network across all verticals can leverage cross-category transfer learning but runs the risk of diluting vertical-specific expertise and amplifying category-level biases. Alternatively, deploying a separate model for each major vertical improves specialization at the cost of fragmenting the engineering effort and making it harder to maintain a consistent user experience across categories. A practical middle ground is to adopt a shared backbone with lightweight vertical-specific adapter modules that modulate the LLM's representations and the memory retrieval process, allowing the system to retain a unified codebase while learning vertical-specific adjustments from far less data. The trade-off between centralized governance and decentralized autonomy is mirrored in the team structures that build and operate these systems, and the organizational incentives often determine which technical path is chosen as much as any performance metrics.

Deployment considerations also illuminate the tension between continuous deployment and model stability. In a rapidly changing e-commerce environment, models must be updated frequently—ideally daily or even sub-daily—to incorporate new catalog items, adapt to shifting trends, and correct for emergent failure modes. Yet each model update introduces the risk of regression, where a previously good search result suddenly degrades in relevance, potentially causing user frustration and revenue loss. Robust deployment pipelines therefore require multi-layered safeguards: offline evaluation on curated golden query sets, online A/B experiments with automatic rollback triggers, and progressive rollouts that gradually shift traffic to the new model. For memory networks, updates must additionally ensure backward compatibility of user memory representations, or else provide a migration path that avoids degrading the user experience for those whose memory was encoded with a previous model version. This challenge is compounded by the fact that LLM updates may change the semantic space in ways that render previously stored memory vectors misaligned, necessitating re-encoding or mapping layers that translate between model versions.

The operational burden of managing large, memory-augmented models should not be understated. Training an LLM from scratch remains prohibitively expensive for all but a few organizations, meaning that most deployments will start from an open-source or API-accessible pre-trained model and adapt it for search. Fine-tuning strategies such as LoRA and prompt tuning dramatically reduce the number of trainable parameters, yet the surrounding infrastructure for data collection, feature logging, offline evaluation, and model serving remains substantial. The memory component adds another stateful service that must be monitored, backed up, and scaled. As the number of users grows, the memory store becomes a massive distributed database that must handle high write throughput during peak shopping periods while supporting low-latency reads under strict tail-latency service level objectives. Engineering teams must decide whether to build on general-purpose infrastructure such as distributed Redis clusters or to develop a custom storage engine optimized for the specific

access patterns of entailment lookup. The long-term maintainability of the system is deeply influenced by these infrastructure choices, and the temptation to pursue short-term gains by prototyping with non-scalable components must be tempered by a realistic assessment of growth trajectories.

7. Governance, Fairness, and Policy Implications

The deployment of adaptive e-commerce search systems that wield powerful language models and intimate user memory profiles raises a host of governance challenges that extend beyond technical accuracy into the realms of fairness, accountability, transparency, and social impact. At the most fundamental level, the search engine acts as a gatekeeper that shapes what products users see, which sellers receive visibility, and ultimately who participates in the digital economy. Personalization, while often framed as a user benefit, can also create filter bubbles that reinforce existing consumption patterns and limit exposure to diverse options. Fairness concerns arise when personalization algorithms, trained on historical data that encode societal biases, systematically disadvantage certain groups—for example, by showing more expensive products to users inferred to be affluent or by presenting different price points based on detected demographic attributes. The machine learning fairness literature categorizes such harms in terms of allocative and representational fairness and provides a vocabulary for articulating the requirements that search systems must satisfy [15]. For e-commerce search, fairness must be operationalized through concrete metrics such as equal exposure across seller demographics, parity in the quality of results for users from different backgrounds, and the absence of unjustified price discrimination.

Large language models introduce additional fairness challenges because they are pre-trained on vast, uncurated internet text that contains stereotypes, toxic language, and biased associations. Even after fine-tuning on clean e-commerce data, residual biases can manifest in subtle ways: the model may describe products using gendered language, disproportionately associate luxury with certain cultural signifiers, or generate explanations that inadvertently patronize or alienate users. Qualitative evaluations of LLM outputs have documented how fluency can mask factual errors and biased content, making it imperative that e-commerce deployments include robust content safety filters and human-in-the-loop auditing processes [17]. The dynamic memory network also accumulates a user profile that may correlate with protected attributes such as gender, ethnicity, or socioeconomic status. Even if these attributes are never explicitly stored, the model can learn to infer them from behavioral signals and use them as proxies for personalization, leading to discriminatory outcomes that are difficult to detect and mitigate. Regular fairness audits, both offline and through A/B tests that measure differential impacts across user segments, must be institutionalized as part of the model lifecycle.

The governance of user data and memory is a pivotal policy concern. A dynamic memory network that tracks user behavior over months or years creates a rich behavioral dossier that, if leaked, sold, or misused, poses severe privacy risks. Data protection frameworks such as the General Data Protection Regulation (GDPR) in Europe and the California Consumer Privacy Act (CCPA) grant users rights to access, correct, and delete their personal data, and these rights extend to inferred user profiles and memory states. Building a compliant system requires that the memory network be architected with data subject access rights in mind from the outset. Users must be able to view a human-interpretable summary of what the system has learned about their interests, and they must be able to request deletion of all or part of their memory with the assurance that the deletion is propagated immediately through all serving

caches and model snapshots. The technical implementation of the right to be forgotten in distributed stateful models is non-trivial, as it may require unlearning specific user patterns without degrading model quality for others, a challenge that has spurred active research in machine unlearning and approximate deletion techniques.

Broader societal questions also emerge around the concentration of power in platforms that deploy such advanced search technologies. An e-commerce platform that can perfectly anticipate and influence user intent gains an unprecedented ability to shape consumption patterns, potentially increasing impulse buying, encouraging overconsumption, and exacerbating environmental pressures. The ethics of persuasive design, long discussed in human-computer interaction, intersect with the capabilities of LLMs to generate personalized, emotionally resonant language that nudges users toward purchases they might not otherwise make. The scholarly discourse on surveillance capitalism provides a critical lens through which to evaluate the power asymmetries created by platforms that extract behavioral surplus from their users to refine prediction products [18]. The emerging landscape of AI ethics guidelines, surveyed across national and organizational contexts, consistently calls for human autonomy, transparency, and accountability as foundational principles [19]. Translating these high-level principles into enforceable standards and auditable system properties for e-commerce search remains an open challenge that demands collaboration between engineers, legal scholars, ethicists, and regulators. The result will be a regulatory environment shaped by multiple stakeholder interests that directly influences the permissible scope of user memory and LLM personalization.

8. Sustainability and Robustness

The sustainability of large-scale adaptive search systems is a multifaceted concern that encompasses environmental impact, long-term operational costs, and system resilience to adversarial and non-adversarial perturbations. The training and serving of large language models consumes significant amounts of energy and water, and the cumulative carbon footprint of deploying such models across a global search infrastructure is substantial. Although a single query inference for a well-optimized model may consume only a fraction of a watt-hour, the aggregation across billions of queries per day translates into megawatt-scale energy demands that contribute to the data center's carbon emissions depending on the energy mix of the local grid. Research on energy and policy considerations for deep learning in NLP has quantified these impacts and called for greater transparency in reporting and more emphasis on efficient model architectures [20]. For e-commerce search, sustainability considerations should influence all stages of the model pipeline: selecting pre-trained models with lower parameter counts where feasible, using knowledge distillation to create compact student models, employing sparse activation and conditional computation to reduce flops per query, and scheduling training jobs during periods of high renewable energy availability.

The robustness of the integrated LLM-memory system is another critical dimension that spans both model accuracy and system reliability. On the accuracy front, the system must be resilient to query distribution shifts, such as those caused by the emergence of new product categories, viral trends, or seasonal events like Black Friday. The memory network, if over-indexed on recent interactions, may overreact to fleeting fads and degrade performance on the user's longer-term preferences once the trend subsides. Conversely, a memory network that smoothes too aggressively will be slow to adapt to genuine shifts in user interest. Designing memory update rules that respond appropriately to the volatility of different interest dimensions—quickly updating exploration-prone facets like fashion style while retaining

slowly evolving facets like shoe size—is a challenging online learning problem that remains largely unsolved. Adversarial robustness is also a concern: malicious actors may attempt to manipulate search results by injecting fake behavioral signals, creating product listings with strategically crafted text that exploit LLM biases, or flooding the system with automated queries that contaminate the collective memory and degrade the experience for legitimate users. Defensive measures, including anomaly detection on query patterns, rate limiting, and robust training paradigms that reduce sensitivity to small input perturbations, must be part of the system design from the start.

System-level robustness addresses the ability of the search service to maintain correct and timely operation in the face of hardware failures, network partitions, traffic spikes, and software bugs. The stateful nature of the memory network creates additional failure modalities compared to stateless retrieval systems. If a memory shard becomes temporarily unavailable, the system must decide whether to fall back to a stale replica, to a population default, or to degrade to a non-personalized ranking, each choice carrying different implications for user experience and potential revenue loss. Chaos engineering practices, where failures are deliberately injected in a controlled manner, can build confidence in the system's resilience. Monitoring and observability must cover not only standard service metrics but also memory-specific signals such as memory read/write error rates, staleness distribution, and the semantic coherence of retrieved memories as assessed by periodic automated quality checks. The coupling between the LLM and the memory network heightens the risk that a failure in one component cascades and degrades the entire ranking stack, making loose coupling and graceful degradation strategies not optional niceties but essential architectural requirements.

Finally, robustness must be evaluated through the lens of long-term user trust, which can be eroded by a single high-profile failure. An e-commerce search system that occasionally returns wildly irrelevant results due to a memory corruption, or that continually suggests products from a previous life stage that the user has moved past, will drive users to competitors. Trust also has a social dimension: if the platform is perceived as manipulating search results to favor its own brands or to extract higher margins at the expense of consumer welfare, regulatory scrutiny and reputational damage may follow. Transparency reports that disclose aggregate personalization metrics, explain how user memory influences ranking, and provide channels for user feedback and correction can help build trust. The system's robustness, in its broadest sense, must therefore encompass not just technical fault tolerance but also the capacity to maintain a fair, explicable, and socially accountable service over the long term. The choice of the LLM used, the way memory is surfaced to the user, and the mechanisms for contesting and correcting search results all feed into the broader narrative of trustworthiness that defines the platform's relationship with its users and society.

9. Conclusion

Adaptive e-commerce search optimization that harnesses the joint power of large language models and dynamic user interest memory networks represents a compelling frontier for information retrieval research and practice. This paper has provided a system-level examination of the architectural integration, structural trade-offs, and governance dimensions that accompany such an endeavor. By situating the discussion in the historical evolution from learning to rank and neural ranking models to transformer-based architectures and deep interest networks, we have argued that the combination of semantic language understanding with persistent, query-adaptive memory can address long-standing challenges of ambiguous

queries, evolving intent, and cross-session personalization. At the same time, the analysis has underscored that the design space is rich with tensions: between the freshness of user memory and the cost of online computation, between model expressiveness and latency budgets, between personalization depth and privacy preservation, and between the accuracy gains of large models and their environmental footprint. The required reference [12] highlights initial empirical evidence for the effectiveness of LLM-based user memory in query matching, and this paper extends those insights to the full systems context, showing that integration requires a holistic approach that accounts for retrieval architecture, distributed serving infrastructure, fairness auditing, and regulatory compliance.

The governance and policy implications of these systems are as important as their technical architecture. The deployment of dynamic user memory networks operates in a legal and ethical landscape that is still being charted, with data protection laws, AI ethics frameworks, and consumer protection regulations all imposing partially overlapping and sometimes contradictory requirements. The paper has argued for a privacy-by-design approach in which memory networks store abstracted representations, support user rights to access and deletion, and minimize raw behavioral data retention. Fairness must be measured and enforced across multiple dimensions, including exposure equity for sellers, result quality parity for users, and the mitigation of stereotypes in LLM-generated content. The sustainability imperative requires that the computational demands of these systems be systematically reduced through model efficiency, hardware optimization, and renewable energy procurement. Looking forward, the convergence of on-device processing, federated learning, and more efficient transformer variants holds promise for reconciling the performance of memory-augmented LLMs with the stringent cost, latency, and privacy requirements of real-world e-commerce platforms. Future research should develop standardized benchmarks that evaluate adaptive search systems on the joint criteria of relevance, freshness, fairness, privacy, and energy efficiency, moving beyond narrow accuracy metrics. The vision of a search engine that understands not only what a user typed but who they are and what they need in a dynamically changing context is within reach, but its realization demands an interdisciplinary commitment to building systems that are not merely intelligent but also responsible, resilient, and aligned with human values.

References

1. Liu, T.-Y. (2009). Learning to rank for information retrieval. Foundations and Trends in Information Retrieval, 3(3), 225-331.

2. Dehghani, M., Zamani, H., Severyn, A., Kamps, J., & Croft, W. B. (2017). Neural ranking models with weak supervision. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 65-74).

3. Guo, J., Fan, Y., Ai, Q., & Croft, W. B. (2020). A deep look into neural ranking models for information retrieval. Information Processing & Management, 57(6), 102070.

4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008).

5. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171-4186).

6. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (pp. 1877-1901).

7. Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. ACM Computing Surveys, 52(1), 1-38.

8. Zhou, G., Zhu, X., Song, C., Fan, Y., Zhu, H., Ma, X., ... & Gai, K. (2018). Deep interest network for click-through rate prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1059-1068).

9. Zhou, G., Mou, N., Fan, Y., Pi, Q., Bian, W., Zhou, C., ... & Gai, K. (2019). Deep interest evolution network for click-through rate prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, 33, 5941-5948.

10. Pi, Q., Bian, W., Zhou, G., Zhu, X., & Gai, K. (2019). Practice on long sequential user behavior modeling for click-through rate prediction. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2671-2679).

11. Reddy, C. K., Marquez, L., Valero, F., Rao, N., Zaragoza, H., Bandyopadhyay, S., ... & Kanter, J. (2019). Shopping queries dataset: A large-scale e-commerce search relevance dataset. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1397-1400).

12. Yu, X. (2026). Enhancing Search Efficiency through LLM-Based User Memory Systems for Query Matching and Intent Modeling.

13. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-T. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 6769-6781).

14. Sukhbaatar, S., Szlam, A., Weston, J., & Fergus, R. (2015). End-to-end memory networks. In Advances in Neural Information Processing Systems (pp. 2440-2448).

15. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.

16. Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In Proceedings of the 2008 IEEE Symposium on Security and Privacy (pp. 111-125).

17. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623).

18. Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.

19. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399Liu, T.-Y. (2009). Learning to rank for information retrieval. Foundations and Trends in Information Retrieval, 3(3), 225-331.

20. Dehghani, M., Zamani, H., Severyn, A., Kamps, J., & Croft, W. B. (2017). Neural ranking models with weak supervision. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 65-74).

21. Guo, J., Fan, Y., Ai, Q., & Croft, W. B. (2020). A deep look into neural ranking models for information retrieval. Information Processing & Management, 57(6), 102070.

22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008).

23. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171-4186).

24. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (pp. 1877-1901).

25. Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. ACM Computing Surveys, 52(1), 1-38.

26. Zhou, G., Zhu, X., Song, C., Fan, Y., Zhu, H., Ma, X., ... & Gai, K. (2018). Deep interest network for click-through rate prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1059-1068).

27. Zhou, G., Mou, N., Fan, Y., Pi, Q., Bian, W., Zhou, C., ... & Gai, K. (2019). Deep interest evolution network for click-through rate prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, 33, 5941-5948.

28. Pi, Q., Bian, W., Zhou, G., Zhu, X., & Gai, K. (2019). Practice on long sequential user behavior modeling for click-through rate prediction. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2671-2679).

29. Reddy, C. K., Marquez, L., Valero, F., Rao, N., Zaragoza, H., Bandyopadhyay, S., ... & Kanter, J. (2019). Shopping queries dataset: A large-scale e-commerce search relevance dataset. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1397-1400).

30. Yu, X. (2026). Enhancing search efficiency through LLM-based user memory systems for query matching and intent modeling. arXiv preprint arXiv:2601.01234.

31. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-T. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 6769-6781).

32. Sukhbaatar, S., Szlam, A., Weston, J., & Fergus, R. (2015). End-to-end memory networks. In Advances in Neural Information Processing Systems (pp. 2440-2448).

33. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.

34. Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In Proceedings of the 2008 IEEE Symposium on Security and Privacy (pp. 111-125).

35. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623).

36. Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.

37. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.