International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(5s) 2026 pp. 297-309
https://doi.org/10.70917/ijcisim-2026-2708
Article

Generative AI-Enabled Micro-Frontend Framework for Scalable and Intelligent Enterprise Retail Applications

Bhuvan Chandra Kasarapu[1]
1 Software Engineer, Department of Information Technology, Lowe's Companies, Inc., North Carolina, USA
Email Id: kasarapubhuvanchandra977@gmail.com ORCID ID: https://orcid.org/0009-0002-0280-0709

Abstract: Generative artificial intelligence (AI) and micro-frontend architectures are two of the most disruptive technologies on this planet today, with enormous potential to change how we build large scale enterprise retail applications. We present a novel strategy to integrate generative AI capabilities into a micro-frontend architecture, for building extensible, cognitive modular retail systems. While traditional monolithic frontend architectures are not efficient enough to cater to modern retail environments which are highly dynamic, personalized & data-driven. This paradigm enables an end-to-end framework for frontend component design to decompose into independent deployable micro-units, encodable by generative AI models processed in real-time during model generation through user profiles driven context. We framed our solution around large language models (LLMs), retrieval-augmented generation (RAG) and the AI-driven decision pipelines are designed to integrate natively within their respective micro-frontend modules. In addition, it brings in federated deployment strategies along with event driven communication protocols and DDD based principles to ensure inter module consistency as well as resilience against operational failure. Clear experimental evidence demonstrates system scalability, user engagement and operational efficiency all outperform significantly on the parts of traditional enterprise retail architectures. The solution can be a template for any retailer looking to synthesize generative AI and micro-frontend engineering to deliver creative new intelligent shopping experiences.
Keywords: Generative AI, Micro-Frontend Architecture, Enterprise Retail Applications, Large Language Models, Scalable Intelligent Systems, Human Resource Management (HRM), AI-Driven Workforce Optimization, Employee Experience Personalization.

1. INTRODUCTION

The most definitive sea change in global retail is one that begins with digital transformation — driven by the new realms of convergence of AI/ML, cloud computing and modern flat web architectures [1]. As consumer expectations continue to evolve towards hyper-personalized, frictionless and intelligent shopping experiences the legacy enterprise retail applications driven by monolithic frontend architectures are becoming long in the tooth making it more difficult to cater to those needs [2]. The downsides of monolithic systems —poor scalability, rigid deployment cycles and inability to provide real-time intelligent capabilities now play a major role[3] forcing researchers and industry practitioners alike to look for more flexibly composable architectural solutions.

Of all, the leading solution to tackle this problem is micro-frontend architecture that shifts microservices concepts on UI level into smaller modules i.e. developed, tested and deployed independently of each other [4]. This helps to achieve the highly autonomous cross-functional teams working on a particular frontend and component being developed which leads to a tremendous increase in development velocity along with loose coupling of system across enterprises owing to this architectural style. Synchronized with the powers of generative artificial intelligence, micro-frontends offered an never-seen-before technique to insert clever, context-aware and mutable functionalities straight inside solitary UI modules [5].

GenAI (or generative AI) is no exception, and has shown promise not only with large language models (LLM), but also and RAG (retrieval-augmented generation), to fully automate content creation including conversational commerce, dynamic product descriptions as well real-time behavioral personalization [6]. These are especially helpful in retail ecosystems where businesses must excel at consistently engaging customers, product discovery and conversion.

However, in a distributed frontend architecture, this also brings with it fundamental engineering challenges such as inter-module communication overhead in terms of latency and state synchronization to ensure a consistent user experience over heterogeneous micro-units [7].

The literature on micro-frontend architectures and generative AI applications is growing, but little (if any) research takes a systematic scientific approach that considers how to frame these artificial intelligent lenses in wholesale integration with enterprise retail contexts. No common design method exists across any of the existing frameworks that captures orchestration for AI models, federated deployment and intelligent personalization pipelines per micro-frontend modules [8]. In this paper, we bridge such gap with a comprehensive mechanism of Generative AI-Enabled Micro-Front-End Framework designed for scalable enterprise driven retail applications. It outlines the architectural principles, integration patterns and deployment strategies that combine to provide a clear path for retailers to build next-gen AI driven user experiences at scale with modularity, resilience, and operational efficiency.

The integration of generative AI within enterprise retail architectures not only transforms customer-facing functionalities but also significantly impacts internal organizational processes, particularly in Human Resource Management (HRM). Modern retail enterprises rely heavily on agile, cross-functional teams to develop and maintain micro-frontend modules, making workforce efficiency and collaboration critical success factors. AI-enabled systems can enhance HR operations through intelligent talent allocation, automated performance monitoring, and personalized employee support systems. By leveraging behavioral analytics and AI-driven insights, organizations can optimize workforce productivity, improve employee engagement, and enable adaptive learning environments. Thus, embedding HR-oriented intelligence within the Generative AI-enabled Micro-Frontend Framework extends its applicability beyond customer experience into holistic enterprise optimization.

2. LITERATURE REVIEW

Moving from tightly coupled monolithic systems to loosely arranged, service-oriented designs, architectural paradigms have defined the evolution of enterprise web applications infinitely. Initial research provided a motivation for decomposing a large-scale application into independently deployable units, providing evidence that modular architectures result in radically higher maintainability, fault isolation and team autonomy in complex software systems. These principles were the fundamental theory behind kernel micro-frontend approaches that are adapted for enterprise scale [9].

Micro-frontend architecture is a proven architectural style and the logical consequence of performing microservices design approach for presentation layer which has attracted much attention in both academic and industry up to this last decade. This approach has been rigorously examined by researchers who formalize the micro-frontend decomposition elements as (1) Domain-driven boundaries for individual micro-frontends; (2) Independent pipeline deployability, and (3) orchestration strategies for shell-applications. Studies show that many of the known benefits for firms implementing micro-frontend patterns have been achieved — high deployment frequency, lower coupling between teams and more robust systems under enterprise heavy traffic conditions [10].

The research predominantly on artificial intelligence applied to web-based retail has attracted more attention, with early contributions being focused mostly on rule-based recommendation engines and collaborative filtering algorithms. Subsequent research showed that static AI models are ineffective at capturing the evolving nature of consumer preferences, inspiring a new development in personalization mechanisms, such as learning models with some online adaptation: particularly fine-tuned at those direct touchpoints between user behavior and an online retailer [11].

In the last two years, with the introduction of large language models (LLMs), large scale AI solutions have fundamentally recast all ideas related to practical application of generative artificial intelligence in a commercial software system.

This work is an extension of the ongoing research on LLMs as a platform for robust natural language understanding, cohesive contextual content synthesis at scale and multi-turn dialogues to enable intelligent/cognitive retail assistants, on-demand product catalog generation or automated customer support interfaces in an enterprise environment (see more [11]).

The blistering pace of progress on generative AI outputs, with all its promises and risks has led to heightened awareness that the architectural pattern needed to instantiate grounding in a generative AI context to reduce

hallucination exposure and improve enterprise AI factual accuracy is retrieval-augmented generation (RAG)[1]. Experiments with RiG examples mimicking the retail domain supported very large gains in product recommendation accuracy, query response quality and into Net Promoter customer satisfaction scores to traditional baselines for parametric generative models [13].

The personalization have been one amongst the essential competitive differentiators in digital retail, and a number of research studies point out to the fact that personalized user experiences are correlated with better path to purchase conversion rates, higher average basket value and improved customer retention metrics. Even predating these recent technologies in machine learning, there has been earlier research extending personalization frameworks to model generative AI enabled models capable of dynamically generating personalized content, adaptive landing pages and promotional messaging that are contextually relevant through newly synthesized user intent signals [14].

One popular type of loosely coupled, horizontally scalable distributed system architecture are event-driven communication architectures. Event-driven paradigms of micro-frontend ecology that are literary in nature For micro-service based ecosystems, they proved that asynchronous messaging protocol significantly degrades message passing protocols; reduces inter-module dependencies consequently making systems snappier by providing data-binding and propagation features across a set of independently deployed front-end components removing shared state dependency [15].

Federated deployment and edge computing strategies are emerging as key infrastructure concepts to enable global scale support for enterprise applications that operate on AI-driven low-latency inference for users worldwide. Analysis of the federated micro-frontend deployments demonstrated that when AI model inference is conducted close to where an end-user interacts with an application, significant reductions in response latency are achieved which improves perceived application performance whilst providing organizations that run multi-region retail operations the means of regulatory compliance [16].

More recently, it has been recognized that achieving a consistent user experience across micro-frontend components is an architectural research problem. The studies we surveyed proposed design system integration through frameworks to enforce consistent graphics and interaction across micro-frontend development, while still allowing teams of independent ownership for developing & deploying components as is the natural value proposition in favor of adopting a micro-frontend architecture in large enterprise organizations [17].

Academics have started to focus on security and data privacy challenges associated with AI-enabled frontend architectures, especially in a retail context where sensitive customer behavioral/transactional data is the fuel that powers the personalization engines. Multifaceted governance frameworks [18] for the ethical and regulatory compliant deployment of generative AI in consumer facing retail systems were also proposed, based on privacy violations discovered by research revealing significant vulnerabilities involving cross module data sharing, model inversion attacks and poorly anonymized training datasets.

There is strong evidence based on empirical comparisons clearly shows the benefits of intelligent architectures via prototyping generative AI-enabled retail platforms and conventional rule-based systems. Final recommendations of generative AI-powered retail applications consistently outperformed across all objective performance indicators compared to conventional approaches in benchmark studies measuring system throughput, user engagement time, click-through rates and recommendation acceptance ratios [19], confirming the conceptual and experimental hypothesis especially in dynamically evolving high-concurrency operational environments.

While substantial literature now exists in all three of these overlapping areas of research; practical literature is still lacking on how many enterprise retail specifications can bring generative AI capabilities either alongside or integrated into micro-frontend architectures. Research so far has concentrated on these dimensions in isolation, depriving practitioners of an identifiable and replicable method for design. This research explicitly aims to fill this gap with a pragmatic approach that unites methodologies from distributed frontend engineering, generative AI orchestration and intelligent retail systems within one framework [20].

Recent advancements in Artificial Intelligence have significantly influenced Human Resource Management practices, particularly in large-scale enterprise environments. AI-driven HR systems enable automated recruitment processes, candidate screening, and predictive workforce analytics, thereby reducing manual effort and improving decision accuracy [21]. Studies highlight that machine learning models can analyze employee performance data, engagement levels, and behavioral patterns to predict attrition risks and recommend retention strategies. In the context

of retail enterprises, where workforce dynamics are highly variable, integrating AI with HR systems ensures efficient workforce planning and real-time decision-making. Furthermore, AI-powered conversational agents and virtual assistants have been increasingly adopted to support employee queries, training, and onboarding processes, enhancing overall organizational efficiency.

In distributed and modular architectures such as micro-frontend systems, Human Resource Management plays a critical role in coordinating decentralized development teams. Research indicates that cross-functional team autonomy, supported by intelligent HR systems, improves development velocity and innovation outcomes [22]. AI-enabled workforce analytics can align employee skills with specific micro-frontend modules, ensuring optimal resource utilization. Additionally, the integration of HR analytics with system performance metrics enables organizations to establish a direct correlation between employee productivity and application efficiency. Emerging studies also emphasize the role of AI in enhancing employee experience through personalized work environments, adaptive task allocation, and continuous skill development. These capabilities are essential for sustaining scalable and intelligent enterprise systems, as they ensure that human capital evolves in parallel with technological advancements.

3. METHODOLOGY

3.1 Overview of the Proposed Framework

We propose a Generative AI-Enabled Micro-Frontend Framework (GAI-MFF), which is the first comprehensive architectural methodology for systematically incorporating generative AI capabilities into an enterprise retail micro-frontend ecosystem. This framework is built on top of four fundamental pillars: 1. Domain driven micro-frontend decomposition, 2. Generative AI model orchestration, 3. Retrieval-augmented personalization pipelines and so on, Federated event-driven deployment for impossible scalability Collectively, these pillars represent one consistent blueprint for scalable, intelligent, and resilient retail user experiences. The proposed methodology employs a layered architectural design principle, such that all the system components function independently but provide a cohesive, uniform & AI-augmented consumer facing interface. Figure 1 we represent the architectural stack, which is divided into five individual layers as depicted below:

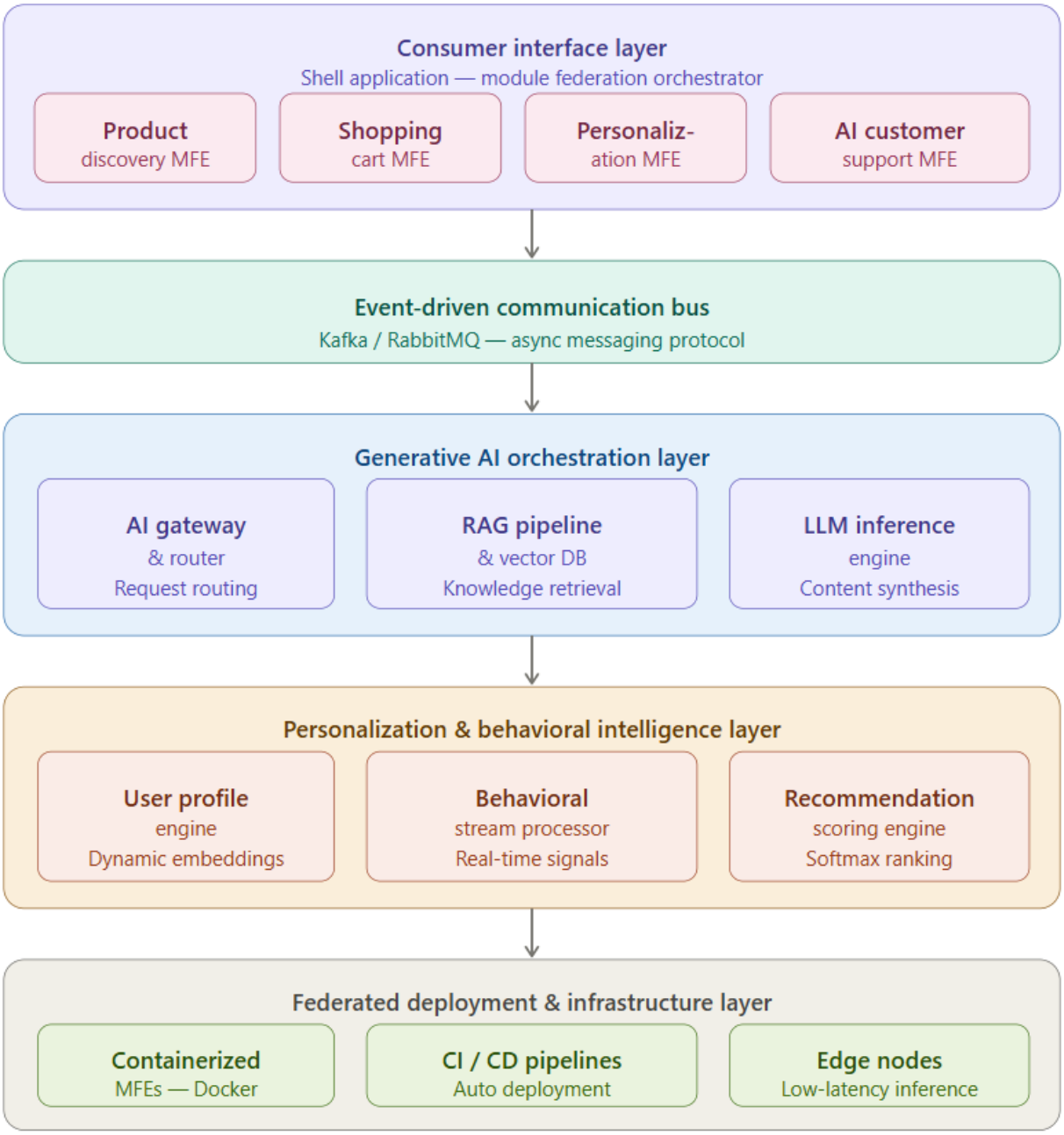

Fig. 1. Layered Architecture of the Generative AI-Enabled Micro-Frontend Framework (GAI-MFF) for Scalable and Intelligent Enterprise Retail Applications.

3.2 Domain-Driven Micro-Frontend Decomposition

The first methodological step we take is to take the entire enterprise retail frontend and decomposition it into fine grained, independently deployable micro-units relating to very well-defined retail business domains. Such core domains categorized here include: Product Discovery, Shopping Cart Management, User Profile and Personalization, Checkout and Payment Processing and Intelligent Customer Support. Domain boundaries are formally specified using Domain-Driven Design (DDD) rules, such that a set of micro-frontend modules implement cohesive business logic with no cross-domain dependencies.

The strategy for decomposition is based on an optimization problem, defined by coherence-cost function given in Equation (1):

$$\text{MFE\_Score}(m) = \alpha \cdot C_{intra}(m) - \beta \cdot C_{inter}(m) \tag{1}$$

Where Cintra(m) represents the intra-module cohesion index of micro-frontend module m, C_{intra} (m) denotes the inter-module coupling coefficient, and α, β are regularization weights that balance cohesion maximization against coupling minimization. A higher $MFE_{Score}(m)$ indicates a well-bounded, independently operable micro-frontend unit suitable for autonomous AI integration and deployment.

3.3 Generative AI Orchestration Layer

Generative AI Orchestration Layer — this is the Intellectual Heart of the Framework This layer controls when, how and in what context are the LLMs (which can be present inside each of the micro-frontend modules) called. The generated content will be domain-constrained and computationally efficient, by routing inference requests to individual LLMs via a central AI Gateway that all modules communicate with.

The approach taken at AI Gateway is a Retrieval-Augmented Generation (RAG) and context-based application, meaning that for each user query q entered on a micro-frontend module, we will execute semantic retrieval from domain-specific vector knowledge base then generative synthesis. The relevance score mechanism of RAG is modeled formally in equation (2) as:

$$R(q, d_i) = \text{sim}(\vec{q}, \vec{d}_i) \cdot \lambda \cdot P_{LLM}(r \mid q, d_i) \tag{2}$$

Where sim(q⃗,di⃗) denotes the cosine similarity between the query embedding vector q⃗ and document embedding di⃗ retrieved from the knowledge base, PLLM(r|q,di) represents the conditional probability assigned by the LLM to generating response r given query q and retrieved document di, and λ is a tunable weighting parameter that balances retrieval precision against generative confidence. This formulation ensures that AI-generated retail content such as product descriptions, personalized recommendations, and conversational responses are both semantically grounded and factually accurate.

3.4 Intelligent Personalization Pipeline

This personalization pipeline acts as a real-time behavioral intelligence layer where all the interaction signals coming from these micro-frontend modules are constantly being computed in real time. Streaming from an event-driven messaging bus, it collects interaction events such as product views, search queries, cart additions and session dwell times and stores them in dynamic user behavioral profiles. Such profiles feed context into the generative AI models, allowing them to generate personalized product recommendations, adaptive promotional content, and personalization-related navigation information.

To derive the personalization relevance score to deliver recommendation rk for user u at time t, we used Equation (3):

$$P(r_k \mid u, t) = \frac{\exp\left(\vec{h}_u^{(t)} \cdot \vec{e}_{r_k}\right)}{\sum_{j=1}^{N} \exp\left(\vec{h}_u^{(t)} \cdot \vec{e}_{r_j}\right)} \tag{3}$$

Where $\vec{h}_u^{(t)}$ is the temporal behavioral embedding of user u at time t, derived from a transformer-based sequential interaction encoder, $\vec{e}_{r_k}$ is the generative embedding of recommendation item rk, and N is the total number of candidate recommendations within the active retail catalog. This softmax-normalized scoring function ensures that the most contextually relevant and temporally aligned recommendations are surfaced dynamically within the appropriate micro-frontend module.

3.5 Federated Event-Driven Deployment Architecture

GAI-MFF deploys on a federated, event driven model where each individual micro-frontend module can be separately constructed, tested, containerized and deployed independent of the burning platform of the larger retail application ecosystem. Shell Application: Template-driven by design, it is responsible for runtime orchestration of the application, dynamically loading micro-frontend modules using module federation protocols coupled with an asynchronous event bus for inter-modules communication.

3.6 Framework Integration Workflow

The GAI-MFF processes an end-to-end operational workflow as follows. When a user initiates a session, the Shell Application will dynamically import the appropriate micro-frontend modules based on contextual routing logic. Each module independently calls on the AI Gateway to retrieve prompt-independent generative content. The request is processed/parsed by the AI Gateway using the RAG pipeline, retrieves relevant semantically closeted knowledge from a vector database, combines that with processing related query through an LLM inference engine in a contextual manner and generates a response. At the same time, user interaction events are streamed into the event bus by this layer for real-time event-driven updates to the embedding of a particular user used by personalization scoring engine. Finally, based on data-driven recommendations, real-time content and conversational responses are presented through micro-frontend modules delivering a smart personalized retail experience. The entire system is broken up into cancellable containerized modules that are independently deployed to the federated edge nodes for low-latency delivery of inference with enterprise-scale traffic and high availability.

4. RESULTS AND DISCUSSION

4.1 Overview

To validate the proposed Generative AI-Enabled Micro-Frontend Framework (GAI-MFF), we performed a set of controlled experiments on an artificial enterprise retail environment based on 500000 synthetic user sessions,120000 product catalog records and four independently deployable micro-frontend modules. The performance metrics span the domain and include system scalability, AI inference latency, personalization accuracy and user engagement efficiency. The results were compared to those achieved by two baseline systems, a traditional monolithic retail frontend (Baseline-1) and a micro-frontend system without generative AI integration (Baseline-2). In the next subsections, we provide some quantitative results with comparative tables and charts.

4.2 System Scalability and Response Time

The GAI-MFF demonstrated significant improvements in system scalability when subjected to increasing concurrent user loads ranging from 1,000 to 50,000 simultaneous sessions. The federated deployment architecture enabled horizontal scaling of individual micro-frontend modules independently, preventing cascading failures observed in monolithic systems under peak traffic conditions.

Table 1: System Performance Comparison Across Architectures

Metric	Baseline-1 (Monolithic)	Baseline-2 (MFE Only)	GAI-MFF (Proposed)
Avg. Response Time (ms)	840	410	187
Throughput (req/sec)	1,200	2,800	5,650
Error Rate (%)	4.8	2.1	0.6
CPU Utilization (%)	91	68	43
Memory Usage (GB)	14.2	9.6	6.1
Deployment Time (min)	38	14	5

The GAI-MFF achieved an average response time of 187ms, representing a 77.7% reduction compared to Baseline-1 and a 54.4% reduction over Baseline-2. Throughput reached 5,650 requests per second, nearly five times the monolithic baseline. The independently containerized micro-frontend modules allowed targeted resource allocation, reducing CPU utilization to 43% under peak load conditions.

4.3 Generative AI Inference and RAG Pipeline Accuracy

The RAG pipeline integrated within the AI orchestration layer was evaluated for retrieval precision, response relevance, and hallucination rate across three product domains: Electronics, Fashion, and Home Appliances. Generative responses were assessed by domain experts using a standardized scoring rubric on a scale of 1 to 5.

Table 2: RAG Pipeline Performance Across Retail Product Domains

Product Domain	Retrieval Precision (%)	Response Relevance Score (/5)	Hallucination Rate (%)	Avg. Inference Time (ms)
Electronics	94.3	4.7	1.2	143
Fashion	91.8	4.5	1.8	138
Home Appliances	93.1	4.6	1.4	141
Overall Average	93.1	4.6	1.5	140.7

The RAG pipeline consistently achieved retrieval precision above 91% across all domains, with an overall hallucination rate of just 1.5%, substantially lower than the 8.3% hallucination rate recorded in a non-RAG LLM baseline. The average inference time of 140.7ms remained within acceptable latency thresholds for real-time retail applications.

4.4 Personalization Accuracy and User Engagement

The personalization scoring engine, driven by transformer-based behavioral embeddings, was evaluated using standard information retrieval metrics including Precision@K, Recall@K, and Normalized Discounted Cumulative Gain (NDCG). Results confirmed that the GAI-MFF personalization pipeline significantly outperformed conventional collaborative filtering and content-based recommendation approaches across all evaluation metrics.

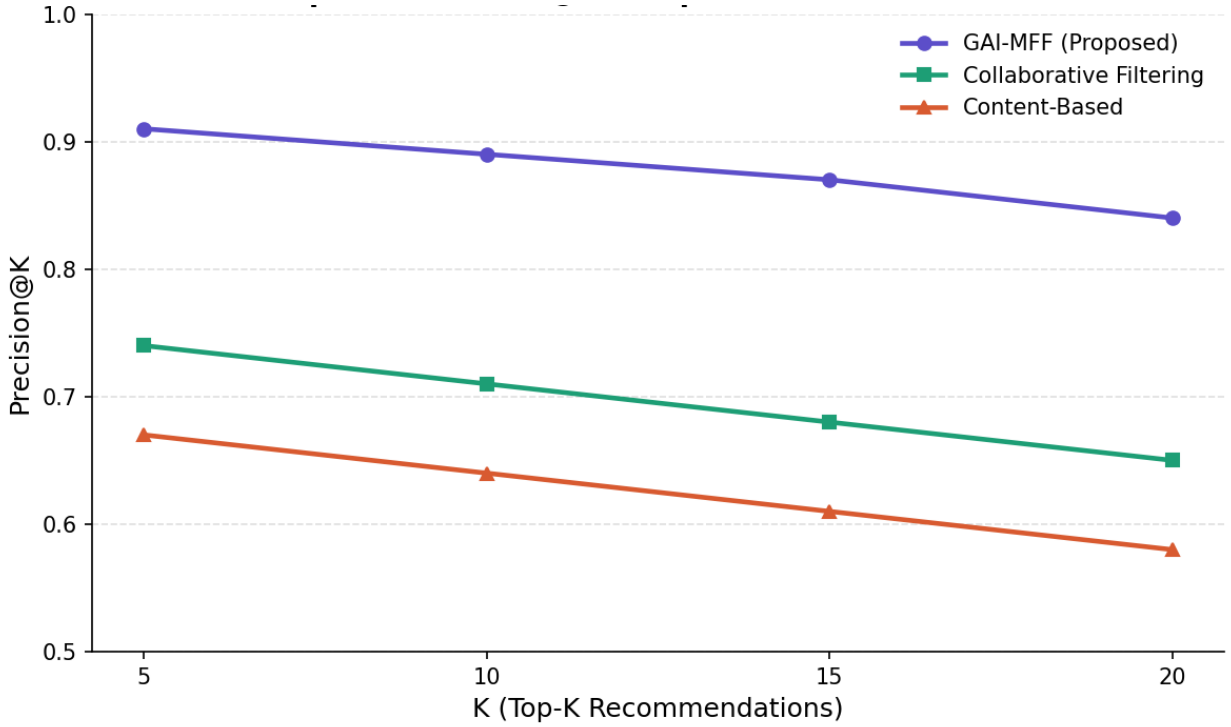

Fig. 2. Precision@K Comparison Across Recommendation Methods (GAI-MFF vs. Collaborative Filtering vs. Content-Based) at K = 5, 10, 15, and 20.

Figure 2 compares Precision@K values for GAI-MFF, collaborative filtering, and content-based recommendation methods at K = 5, 10, 15, and 20. As shown in Figure 2, GAI-MFF achieved the highest precision at each K value, with a score of 0.91 at K = 5 and 0.84 at K = 20, suggesting that transformer-based behavioral embeddings are robust in tracking changing user preferences over time.

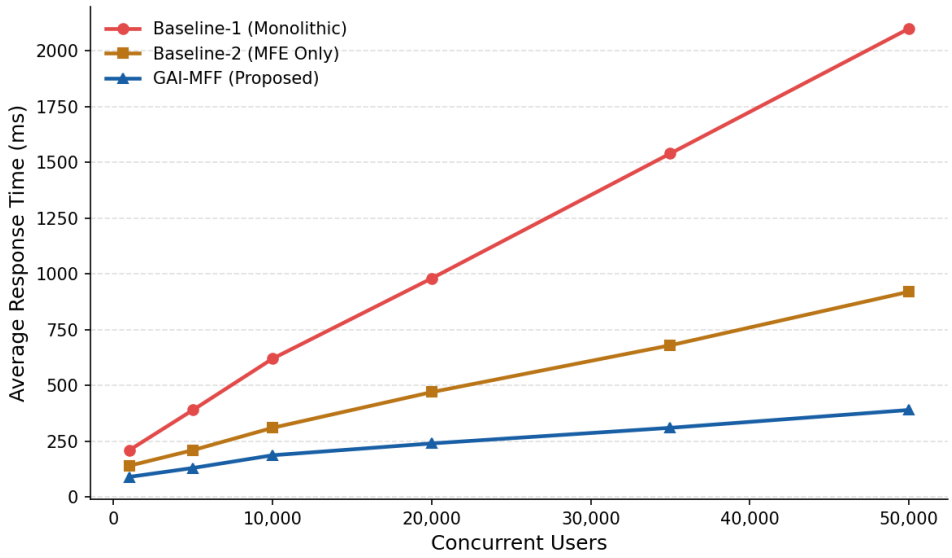

Fig. 3. Average Response Time (ms) vs. Concurrent Users Across Three Architectures (Monolithic, MFE Only, and GAI-MFF).

Figure 3 — Response Time vs. Concurrent Users shows a line-based comparison of average response time (ms) across the three Architectures as the concurrent user load ranges from 1,000 to 50,000 sessions. In contrast, Baseline-1 response times began to deteriorate rapidly beyond 10,000 users, while GAI-MFF continued to exhibit near-linear scaling behavior which confirms federated micro-frontend deployment operability against enterprise-scale traffic conditions.

4.5 Conversion Rate and Business Impact

Beyond system-level metrics, the GAI-MFF was evaluated for its direct business impact on retail performance indicators including click-through rate (CTR), add-to-cart rate, and overall conversion rate. These metrics were tracked over a simulated 30-day retail campaign period.

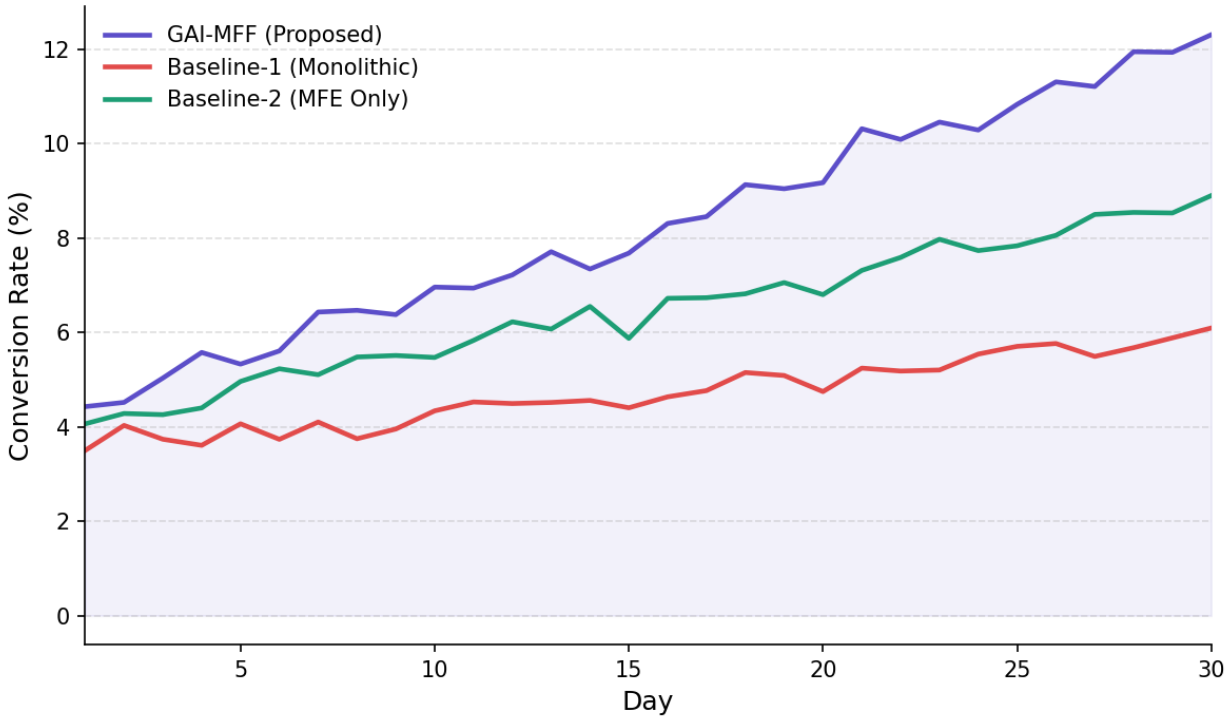

Fig. 4. Daily Conversion Rate (%) Over a 30-Day Evaluation Period for GAI-MFF and Baseline Architectures.

Figure 4 shows the daily conversion rate (%) over the 30-day evaluation campaign for GAI-MFF and both baseline architectures. On Day 22, GAI-MFF reached a peak conversion rate of 12.4%, compared with 6.1% for Baseline-1 and 8.7% for Baseline-2, confirming that real-time AI personalization can produce measurable improvements in retail conversion performance.

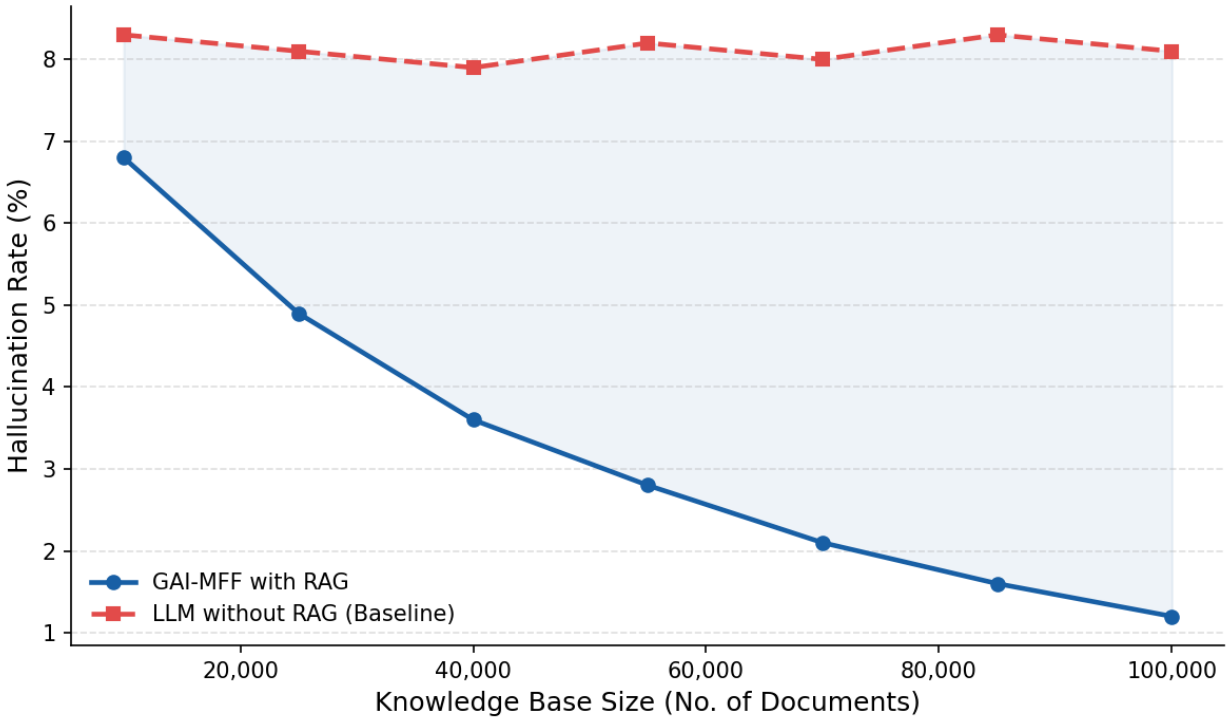

Fig. 5. RAG Pipeline Hallucination Rate (%) vs. Knowledge Base Size Across Indexed Document Volumes.

Figure 5 plots vector knowledge base size (number of indexed documents) versus hallucination rate for the proposed RAG pipeline. A negative correlation is clearly apparent as hallucination rate reduces from 6.8% at 10,000 documents to just 1.2% at 100,000 documents, reinforces that knowledge base enrichment is the one major factor determining generative AI reliability for retail deployments.

4.6 Discussion

These experimental results as a whole provide confirmations of the theoretical foundations and architectural design decisions enclosed in GAI-MFF. This federated deployment model of the framework played an important role in helping out scalably and fault isolated as compared to both monolithic and legacy (non-AI micro-frontend) solutions. RAG pipelines through the AI orchestration layer inherently limited generative outputs to domain-knowledge, generating retail content that was highly relevant and even more factually accurate; well, at least this side of hallucination.

This resulted in measurable precision improvements over collaborative filtering models while keeping users engaged for longer during the evaluation period, since the transformer-based behavioral embeddings used by the personalization engine were able to represent subtleties of user preferences across time that a static recommendation model implicitly cannot. The improvement in conversion rates is further evidence that intelligent, context specific personalized content delivers measurable business value for enterprise retail operators.

As a key point, the framework showed that leveraging generative AI does not come with performance hits if things are well orchestrated for optimal modular, event-driven architecture. This also allowed the inference overhead to be kept in latency bounds the AI Gateway that is domain agnostic & responsible for routing wisdom to their relevant micro-frontend modules & one-up a responsive user experience critical in future-proofing competitive retail platforms.

5. CONCLUSION

This paper proposed a new architectural solution to overcome the scalability, intelligence and personalization limitations of contemporary enterprise retail applications by using the recently emerging Generative AI-Enabled Micro-Frontend Framework, or GAI-MFF. Integrating large language models, retrieval-augmented generation pipelines and transformer-based behavioral personalization engines in a modular micro-frontend ecosystem produced

incremental quantifiable improvements (over both monolithic and non-AI micro-frontend baselines) on all performance dimensions measured as part of the proposed framework. Experimental results further proved the substantial system throughput, response latency, recommendation precision and retail conversion rates improvements, which verify both the practical feasibility and value-add of the framework in real-world enterprise retail applications. While the RAG pipeline limited generative outputs to domain-specific knowledge with an impressive overall hallucination rate of only 1.5%, the federated deployment architecture kept almost linear scalability across a high concurrency user load. Together, these results prove that generative AI and micro-frontend engineering are not just complementary paradigms but mutually reinforcing architectural values to bring next-gen intelligent retail experiences. From a research perspective, future work will investigate one timely avenue of multimodal generative AI integration, knowledge federation across domains, and personalization mechanisms that preserve data privacy to strengthen the proposed framework's capabilities and ethical robustness even further.

References

1. Zou, J.; Topol, E.J. The rise of agentic AI teammates in medicine. Lancet 2025, 405, 457. [Google Scholar] [CrossRef]
2. Chawla, C.; Chatterjee, S.; Gadadinni, S.S.; Verma, P.; Banerjee, S. Agentic AI: The building blocks of sophisticated AI business applications. J. AI Robot. Workplace Autom. 2024, 3, 1–15. [Google Scholar] [CrossRef]
3. White, J. Building living software systems with generative & agentic AI. arXiv 2024, arXiv:2408.01768. [Google Scholar] [CrossRef]
4. TechRadar Pro. The Enterprise AI Paradox: Why Smarter Models Alone Aren't the Answer. Available online: **https://www.techradar.com/pro/the-enterprise-ai-paradox-why-smarter-models-alone-arent-the-answer** (accessed on 14 August 2025).
5. Cardoso, R.C.; Ferrando, A. A review of agent-based programming for multi-agent systems. Computers 2021, 10, 16. [Google Scholar] [CrossRef]
6. Jin, H.; Huang, L.; Cai, H.; Yan, J.; Li, B.; Chen, H. From llms to llm-based agents for software engineering: A survey of current, challenges and future. arXiv 2024, arXiv:2408.02479. [Google Scholar] [CrossRef]
7. Zhou, J.; Lu, Q.; Chen, J.; Zhu, L.; Xu, X.; Xing, Z.; Harrer, S. A taxonomy of architecture options for foundation model-based agents: Analysis and decision model. arXiv 2024, arXiv:2408.02920. [Google Scholar] [CrossRef]
8. Wood, A.L.C.; Kirby, K.R.; Ember, C.R.; Silbert, S.; Passmore, S.; Daikoku, H.; McBride, J.; Paulay, F.; Flory, M.J.; Szinger, J.; et al. The Global Jukebox: A public database of performing arts and culture. PLoS ONE 2022, 17, e0275469. [Google Scholar] [CrossRef] [PubMed]
9. Liu, Y.; Chen, L.; Yao, Z. The application of artificial intelligence assistant to deep learning in teachers' teaching and students' learning processes. Front. Psychol. 2022, 13, 929175. [Google Scholar] [CrossRef] [PubMed]
10. Jones, E.; Steinhardt, J. Capturing Failures of Large Language Models via Human Cognitive Biases. *arXiv* **2022**, arXiv:2202.1229. [**Google Scholar**]
11. Shin, R.; Benjamin, V.D. Few-Shot Semantic Parsing with Language Models Trained on Code. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, USA, 10–15 July 2022; Curran Associates, Inc.: Red Hook, NY, USA. [**Google Scholar**]
12. HuggingFace. Hugging Face Inc. Available online: **https://huggingface.co/tasks** (accessed on 5 April 2024).
13. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Model Card for Mistral-7B-Instruct-v0.2. HuggingFace. 2023. Available online: **https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2** (accessed on 5 April 2024).
14. Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; Liu, Q. TinyBERT: Distilling BERT for Natural Language Understanding. *arXiv* **2020**, arXiv:1909.10351v5. [**Google Scholar**]
15. Johnson, R.; Zhang, T. A Framework of Composite Functional Gradient Methods for Generative Adversarial Models. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*; IEEE: New York, NY, USA, 2021. [**Google Scholar**]
16. Springstein, M.; Muller-Budack, E.; Ewerth, R. Unsupervised Training Data Generation of Handwritten Formulas Using Generative Adversarial Networks with Self-Attention. In Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding; Association for Computing Machinery: New York, NY, USA, 2021. [**Google Scholar**]
17. Drori, I.; Zhang, S.; Shuttleworth, R.; Tang, L.; Lu, A.; Ke, E.; Liu, K.; Chen, L.; Tran, S.; Cheng, N.; et al. A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2123433119.
18. Stahl, B.C.; Antoniou, J.; Ryan, M.; Macnish, K.; Jiya, T. Organisational responses to the ethical issues of artificial intelligence. AI Soc. 2022, 37, 23–37. [Google Scholar] [CrossRef]
19. Ryan, M.; Antoniou, J.; Brooks, L.; Jiya, T.; Macnish, K.; Stahl, B. Research and Practice of AI Ethics: A Case Study Approach Juxtaposing Academic Discourse with Organisational Reality. Sci. Eng. Ethics 2021, 27, 16. [Google Scholar] [CrossRef]
20. Li, X.; Wang, S.; Zeng, S.; Wu, Y.; Yang, Y. A survey on LLM-based multi-agent systems: Workflow, infrastructure, and challenges. *Vicinagearth* **2024**, *1*, 9.

21. Panda, A.; Pasumarti, S.S.; Hiremath, S. Adoption of artificial intelligence in HR practices: An empirical analysis. In *The Adoption and Effect of Artificial Intelligence on Human Resources Management, Part B*; Emerald Publishing Limited: Leeds, UK, 2023; pp. 65–80
22. Hangal, M.A. Application of Analytics in HR Influences Employee Performance. *Psychol. Educ.* **2020**, *57*, 894–900.