ESP-JETA
ESP Journal of Engineering & Technology Advancements
ISSN: 2583-2646 / Volume 4 Issue 4 December 2024 / Page No: 161-165
Paper Id: JETA-V4I4P121 / Doi: 10.56472/25832646/JETA-V4I4P121

Original Article

Enhanced Recommendations Based on Health in E-Commerce Using Large Language Models

Aditi Choudhary
Computer Science, University of Southern California, California, Software Developer at Amazon, USA.

Received Date: 01 October 2024 Revised Date: 11 November 2024 Accepted Date: 02 December 2024

Abstract: *The rapid growth and increasing complexity of e-commerce platforms have driven the need for innovative methods to enhance product recommendations. This research builds on earlier studies that utilized machine learning algorithms and investigates how Large Language Models (LLMs) can further improve recommendation systems. By leveraging their superior natural language comprehension, LLMs can provide personalized recommendations tailored to individual customer inquiries, purchase histories, and product information. In this study, we propose a framework that integrates LLMs using Amazon Bedrock to optimize product recommendations in e-commerce. Our approach emphasizes understanding user intent dynamically and enhancing recommendation accuracy across different product categories. Additionally, the system includes a comprehensive knowledge base that adheres to established rules and regulations related to product recommendations, taking into account individual user health profiles. The findings reveal a significant increase in recommendation precision and overall customer satisfaction, highlighting the transformative potential of LLMs in reshaping recommendation strategies within the e-commerce sector.*

Keywords: *E-Commerce, Recommendation System, Large Language Models, Machine Learning, Amazon Bedrock, Personalization.*

I. INTRODUCTION

The evolution of recommendation systems has been transformative, shifting from simple rule-based approaches to more advanced machine learning techniques to tackle the demands of personalization and scalability. Traditional methods such as Collaborative Filtering (CF), Support Vector Machines (SVM), and Random Forest have played a pivotal role in improving recommendation accuracy. However, these models often struggled with challenges like sparse data and their inability to adapt to ever-changing user preferences. With the introduction of Large Language Models (LLMs) like GPT-4 and the emergence of platforms such as Amazon Bedrock, Google Vertex AI, and Microsoft Azure Cognitive Services, recommendation systems have entered a new era. LLMs bring the ability to analyze unstructured data, including user reviews, product descriptions, and conversational queries, allowing them to generate highly personalized and contextually relevant recommendations. Each LLM platform offers unique advantages. For instance, Google Vertex AI is well-regarded for its capability to handle complex queries and sustain coherent interactions over extended dialogues. Similarly, Microsoft Azure Cognitive Services excels at providing enterprise-grade tools for creating sophisticated virtual assistants and conversational agents. However, this paper emphasizes Amazon Bedrock for its exceptional performance in classification tasks and its seamless integration with the AWS ecosystem, making it a reliable choice for building robust and scalable recommendation systems. This research proposes an innovative recommendation system that leverages Amazon Bedrock's LLM capabilities in conjunction with a knowledge base specifically designed for health-aware product recommendations. This integration not only enhances the system's accuracy but also enables it to address user-specific health requirements by filtering and prioritizing products accordingly. By harnessing the power of Amazon Bedrock, the proposed system aims to deliver precise, scalable, and personalized recommendations tailored to the dynamic needs of e-commerce platforms.

II. E-COMMERCE PRODUCT RECOMMENDATIONS DESIGN AND IMPLEMENTATION

The product we're developing is essentially a web service powered by a Large Language Model (LLM) running behind the scenes. The user interface is built using React, which allows for a clean, intuitive, and responsive experience. This LLM is integrated with Amazon Bedrock, a platform that provides robust capabilities for working with foundational models and handling context-aware data through its Knowledge Bases. The core purpose of the product is to recommend health products tailored to users with specific disorders by leveraging advanced AI-driven insights and rules-based logic.

A. Ecommerce platform:

The web application enables users to input their health details and preferences. On the backend, the system processes this information using Amazon Bedrock's LLM to make sense of user data in context. Think of it as a conversation between the user and an intelligent assistant that understands their unique requirements, cross-references their inputs with a knowledge base, and suggests appropriate products.

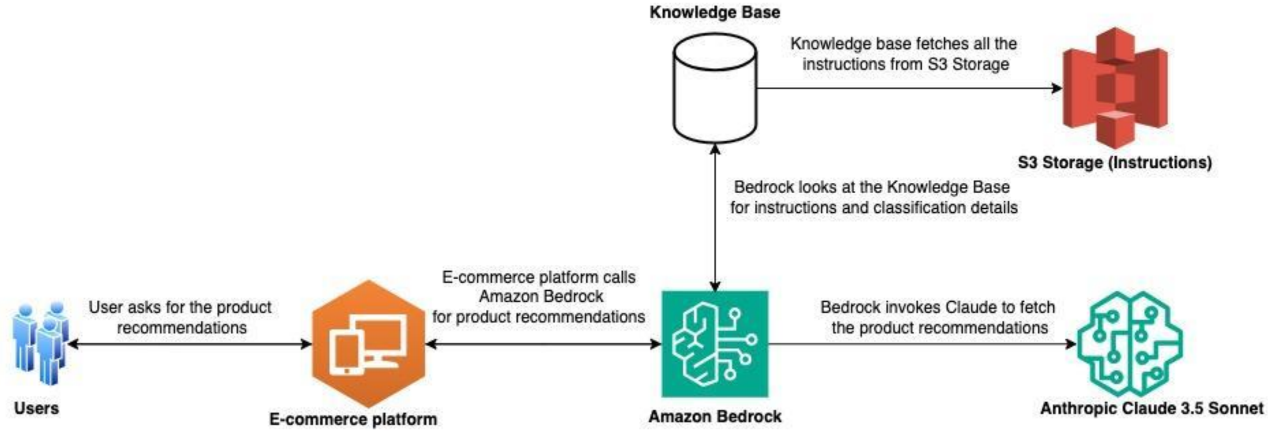

Figure 1: Ecommerce Products

B. Integration with Amazon Bedrock Knowledge Bases:

Amazon Bedrock Knowledge Bases form the backbone of our recommendation system. These knowledge bases have been trained specifically to understand segmentation rules for individuals with different health disorders. This ensures that the recommendations provided aren't just general or vague—they are rooted in actual data and rules created to meet health-specific needs. Some key capabilities provided by Bedrock Knowledge Bases include:

- Context-Aware Recommendations: They allow the LLM to consider context from previous interactions and user profiles to maintain a consistent flow.
- Source Attribution: This helps ensure transparency, as every recommendation can be traced back to its data source or rule.
- Dynamic Querying: Bedrock can even handle structured queries directly on the knowledge base, so there's no need for a separate database migration or setup.

C. How It Works

The RAG Workflow At the heart of the product lies the Retrieval Augmented Generation (RAG) workflow. This is how the system works step by step:

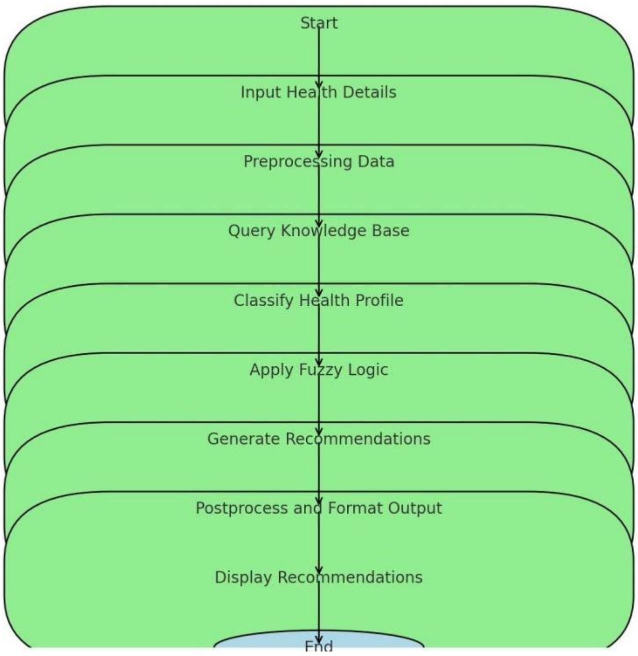

Figure 2: Implementation Chart

a) Pre-Processing:

The system first takes the user's input and any relevant company data (e.g., rules for disorders) and processes it.

This involves cleaning the data, tokenizing it, and preparing it for analysis. If the input data is unstructured (like free text), it's normalized into a format the LLM can interpret effectively.

b) Orchestration:

This is where the magic happens. The orchestration phase defines rules for the LLM to follow. For example, when a user describes their symptoms or condition, the system instructs the LLM to first consult the knowledge base to classify the user under a specific health disorder. Once classified, the system applies fuzzy logic to determine which health products align best with the user's condition. For example, a diabetic user would receive recommendations for low-sugar alternatives.

The key here is that the system doesn't operate on simple matching logic but uses complex AI-driven reasoning powered by Bedrock's managed capabilities.

c) Post-Processing:

Finally, the recommendations are formatted into a clear and user-friendly response. Post-processing ensures that the output aligns with the user's expectations, whether it's a ranked list of products, detailed descriptions, or actionable insights.

D. Why Amazon Bedrock?

- Among other LLM platforms like Google Vertex AI and Microsoft Azure Cognitive Services, we chose Amazon Bedrock because of its seamless integration with AWS infrastructure and its unique capabilities in classification and RAG workflows. Bedrock simplifies the entire pipeline by providing a fully managed workflow. For example:
- If a user's health data includes structured inputs like test results or unstructured notes, Bedrock can automatically query and retrieve information from both formats.
- With in-built natural language query support, it avoids the need for setting up complex vector databases or additional backend infrastructure.

E. Real-World Example

Let's walk through a typical scenario. Suppose a user logs in and mentions they have diabetes. The system immediately uses the LLM to understand this input and consults the knowledge base for rules regarding diabetic users. Based on these rules, it classifies the user and applies fuzzy logic to recommend products like sugar-free snacks or insulin kits. The entire process, from the user entering data to receiving tailored recommendations, is completed in milliseconds.

F. The Human Touch

While the system is highly automated, the rules in the knowledge base are designed by experts—healthcare professionals, data scientists, and product managers—ensuring every recommendation aligns with real-world needs and ethical standards.

III. PROMPTING TECHNIQUE IN DESIGN

Prompting is a critical technique in Large Language Models (LLMs), serving as a mechanism to guide the model's responses by providing structured input examples. It leverages the model's pretrained knowledge to generate accurate and context-aware outputs. Several prompting techniques exist, each tailored for specific use cases and varying levels of complexity.

- Zero-Shot Prompting - Zero-shot prompting involves asking the model to perform a task without providing any prior examples. While this approach is useful for straightforward queries, it can lack precision in complex or nuanced tasks.
- Few-Shot Prompting - Few-shot prompting provides the model with a small number of examples to illustrate the desired response format. This technique enhances the model's understanding of the task and improves output accuracy compared to zero-shot prompting.
- Multishot Prompting - Multishot prompting takes the concept further by providing multiple examples and contexts, ensuring the model gains a deeper understanding of the task. This technique is particularly valuable for applications requiring high accuracy, nuanced responses, or complex decision-making.

A. Importance of Multishot Prompting

Multishot prompting is essential because it allows for:

- Contextual Understanding: The model processes multiple examples, helping it understand variations in user inputs.
- Error Reduction: By exposing the model to diverse scenarios, multishot prompting minimizes the likelihood of incorrect predictions.

- Scalability: It prepares the model to handle complex, multi-layered queries effectively.

Example: Health-Aware Recommendations Using Multishot Prompting

For a person with diabetes, dietary restrictions often guide product recommendations. Multishot prompting can refine the recommendation process by incorporating health constraints and fuzzy logic to classify products based on their nutritional content. Formula for Fuzzy Logic is as follows -

$$\text{Recommendation} = \{\text{HighlyRecommended, Recommended, LeastRecommended, Avoid} \mid \mu(\text{HighlyRecommended}), \mu(\text{Recommended}), \mu(\text{LeastRecommended}), \mu(\text{Avoid})\}\quad \text{Degree(Fat), Degree(Sugar)}$$

For a diabetic individual, the maximum sugar intake is 7g, and the maximum fat content is 4g. Using fuzzy logic, food items are classified as follows: The model evaluates the nutritional properties of food items, mapping them to categories based on their compliance with health requirements. For instance:

- A product with 6g of sugar and 3g of fat might fall under "Recommended."
- A product exceeding both limits would be classified as "Avoid."

By integrating multishot prompting and fuzzy logic, the recommendation system dynamically adapts to user-specific health conditions, providing tailored suggestions. Figure-3 illustrates this process, highlighting how the system leverages fuzzy logic to align recommendations with individual dietary needs.

C. Techniques in Multishot Prompting

- Stepwise Reasoning: Breaking down tasks into smaller steps, allowing the model to process each component sequentially.
- Contextual Variations: Providing examples with varied contexts to ensure robustness.
- Dynamic Prompts: Adjusting prompts dynamically based on user inputs to maintain relevance.

By adopting multishot prompting, the system ensures a robust and adaptable recommendation framework, catering to diverse user needs and enhancing the overall accuracy of the recommendations.

IV. COMPARISION ON VARIOUS LARGE LANGUAGE MODELS IN DESIGN

The design of a recommendation system using LLMs requires a careful evaluation of the capabilities and features of different models. Key LLMs used in industry include OpenAI's GPT series, Google's Gemini, Anthropic's Claude, and Amazon's Titan models available via Bedrock. Each model brings distinct strengths to specific applications:

- OpenAI's GPT Series: Renowned for its versatility, the GPT series excels in generating human-like text and handling a wide range of applications. It is particularly effective in creative tasks such as content generation, summarization, and conversational interfaces. However, its resource-intensive nature may limit scalability in enterprise-level applications.
- Google's Gemini: Gemini is designed for maintaining contextual coherence and handling complex, nuanced queries. Integrated into Google's ecosystem, it offers exceptional capabilities for applications such as conversational AI, search optimization, and multilingual processing. However, it is best suited for organizations heavily relying on Google's infrastructure.
- Anthropic's Claude: Claude is recognized as a seamless choice for intelligent and unmatched performance. It emphasizes safe and responsible usage, making it suitable for sensitive applications like healthcare and education. Its ability to provide superior contextual understanding and robust performance makes it ideal for designing advanced recommendation systems across various domains.
- Amazon's Titan Models via Bedrock: Titan models, accessible through Amazon Bedrock, are optimized for enterprise scalability, supporting diverse foundational tasks such as text classification, summarization, and sentiment analysis. Their seamless integration with AWS infrastructure makes them an ideal choice for large-scale recommendation systems, particularly in retail and e-commerce domains.

These LLMs offer varying trade-offs in terms of accuracy, scalability, and integration. While OpenAI's GPT provides unparalleled generative capabilities, Claude stands out for its intelligence and performance, making it a top choice for advanced systems. Titan models, on the other hand, excel in scalability and practical implementation in enterprise systems. The choice of an LLM depends on the specific requirements of the application, including the need for integration, resource availability, and the target audience.

V. CONCLUSION

The integration of Large Language Models (LLMs) into e-commerce recommendation systems signifies a paradigm shift, offering unparalleled personalization, context-aware decision-making, and scalability. By leveraging platforms like

Amazon Bedrock and incorporating a knowledge base tailored to user-specific requirements, this study has highlighted the ability of LLMs to surpass traditional approaches in accuracy and user satisfaction. The innovative use of techniques such as multishot prompting and fuzzy logic ensures that recommendations align with individual needs, particularly in health-aware scenarios.

While the advancements presented in this paper demonstrate a significant leap forward, the evolving landscape of AI and LLM technologies opens avenues for further exploration. Future work should aim to incorporate real-time feedback loops, extend knowledge bases to multilingual and cross-cultural contexts, and optimize computational efficiency to reduce resource demands. Such enhancements will ensure that recommendation systems not only meet but exceed the expectations of a global and diverse user base, making AI-driven recommendations indispensable for the future of e-commerce.

VI. REFERENCES

[1] Roy, K., Choudhary, A., & Jayapradha, J. (2017). Product Recommendations Using Data Mining and Machine Learning Algorithms. ARPN Journal of Engineering and Applied Sciences.

[2] Brown, T., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. OpenAI Research.

[3] Teytaud, O., & Sarrut, D. (2001). Kernel-Based Image Classification. ICANN.

[4] Yang, H.-W., Pan, Z.-G., & Zhang, M.-M. (2004). Machine Learning-Based Intelligent Recommendation in Virtual Mall.

[5] Tsuji, K., Yoshikane, F., & Sato, S. (2014). Book Recommendation Using Machine Learning Methods. Advanced Applied Informatics.

[6] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems.

[7] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Association for Computational Linguistics.

[8] Hendrycks, D., & Gimpel, K. (2016). Gaussian Error Linear Units (GELUs). arXiv preprint arXiv:1606.08415.

[9] Radford, A., Wu, J., Child, R., et al. (2019). Language Models are Unsupervised Multitask Learners. OpenAI.

[10] Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research.

[11] Zhou, Y., Zhang, X., & Zhang, W. (2021). Explainable Recommendation: Recent Progress and Emerging Challenges. IEEE Transactions on Knowledge and Data Engineering.

[12] Amazon Web Services. (2023). Amazon Bedrock: Building AI Applications with Generative Models. AWS Documentation.

[13] Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

[14] Bubeck, S., Eldan, R., Li, Y., et al. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

[15] Google Cloud. (2023). Introducing Gemini AI: The Next Generation of Conversational AI. Google Cloud Documentation.

[16] Anthropic. (2023). Claude AI: Ethical and Contextually Aware AI Solutions. Anthropic Research.