Conversational Recommender System Using a Combination of Fine-Tuned GPT-4o and Retrieval-Augmented Generation for Laptop Recommendations

1st Fathan Askar
School of Computing
Telkom University
Bandung, Indonesia
askarfathan@student.telkomuniversity.ac.id

2nd Z. K. A. Baizal*
School of Computing
Telkom University
Bandung, Indonesia
baizal@telkomuniversity.ac.id

Abstract*—Conversational recommender systems (CRS) have revolutionized personalized recommendations in recommender systems by using interactive and adaptive decision-making, particularly in complex domains (e.g., laptops). Existing CRS provides interaction between the system and the user through Form-based Layouts and Natural Language. Natural language-based interactions are typically constructed using Conventional Natural Language Processing (C-NLP) methods. While both interactions have shown certain successes, they also have limitations. Form-based layouts restrict users from expressing their preferences freely because of their rigid and structured nature. On the other hand, C-NLP allows for more dynamic interactions but relies heavily on domain-specific datasets and still struggles to interpret complex user requirements. To tackle these issues, we propose the development of a CRS using Large Language Models (LLMs). Specifically, we combined a Fine-Tuned GPT-4o model and the retrieval technique of Retrieval-Augmented Generation (RAG). LLMs, with their extensive pre-training, can process sophisticated conversations, infer subtle user preferences, and deliver precise recommendations without requiring intensive additional training. At the same time, RAG's retrieval technique effectively incorporates large-scale datasets, ensuring the system maintains relevance and scalability. In evaluation, we compare three models: 1) RAG, 2) Fine-Tuned GPT-4o, and 3) Combined Model (Fine-Tuned GPT-4o + RAG's Retrieval Technique). Results show that the Combined Model excels with an average Hit Rate of 1, Precision of 0.9031, NDCG of 0.9865, and the highest user satisfaction. These outcomes confirm that the approach resolves limitations and ensures scalability.

Keywords*—Conversational recommender system, large language models, laptop recommendation, fine-tuned GPT-4o, retrieval-augmented generation

I. INTRODUCTION

Modern consumers increasingly rely on technology to support their daily needs, including making decisions on what products to purchase, such as laptops. Consumers often face hardships when choosing options that suit their requirements and priorities among the many available products. Designing a reliable recommender system to simplify the selection process can contribute to resolving this matter [1]. One viable option that is available is Conversational Recommender Systems (CRS), which deliver customized and comprehensive recommendations through interactive dialogues. CRS is a type of recommender system that facilitates user interaction with the system using natural language through conversation, enabling them to convey preferences and provide additional information effectively [2]. This dialogue-based approach creates opportunities to address more complex user needs. Advances in Large Language Models (LLMs), such as ChatGPT's GPT-4o model, have demonstrated extraordinary capabilities in understanding conversations and generating responses in natural language, establishing it as a suitable foundation for building a more advanced and effective CRS [3].

This research proposes the development of a CRS using Fine- Tuned GPT-4o and Retrieval-Augmented Generation (RAG) retrieval technique for laptop recommendation. We chose the Fine-Tuned GPT-4o model because of its advantages over previous LLMs such as GPT-3.5. The advantages of the GPT- 4o model are higher reasoning ability and better efficiency in handling conversations between the model and the user [4]. The Fine-Tuned GPT-4o model also shows strong adaptability for fine-tuning, which makes it suitable for personalized domains such as laptops.

We also use RAG to help the Fine-Tuned GPT-4o model efficiently cope with large-scale datasets through dynamic retrieval. This combination overcomes the main challenges of previous methods, such as rigid interaction models and difficulty in capturing complex user preferences by utilizing Fine-Tuned GPT-4o and makes the model more scalable by using RAG. The Fine-Tuned GPT-4o and RAG work together to create a robust and scalable CRS [5].

Unlike Form-based Layouts or Conventional Natural Language Processing (C-NLP), the proposed CRS leverages LLMs to address the main challenges of previous interaction methods, such as the limited user flexibility in expressing preferences in Form-based Layouts, as well as avoiding the requirement for comprehensive domain-specific training and addressing the inability to capture complex or implicit user preferences in C-NLP [6][7]. The proposed system is expected to capture user needs expressed in complex natural language, so that it can produce more accurate recommendations. Main contributions of this paper are as follows:

- We introduce a Conversational Recommender System that combines Fine-Tuned GPT-4o and Retrieval-

Augmented Generation to resolve the issues faced by recent methods by utilizing the LLM's ability to comprehend complex conversations and RAG's efficient integration of large-scale datasets for personalized and scalable recommendations.

- We validated the efficiency of the suggested approach through evaluations using Hit Rate, Precision, NDCG metrics, as well as real-world validation through user feedback.
- With the combination of Fine-Tuned GPT-4o and RAG method, we establish a strong foundation for future advancements in CRS.

This article is made up of 5 sections. Section I provides an introduction to this research, including the motivation and objectives of this research. In Section II, we review the main concepts (such as LLM, CRS, LLM in CRS, and RAG) and Related Work of this research. Then, Section III describes the system flow and outline the design of the Combined Model. We outline the evaluation approach and share results in Section IV, analyzing the system's effectiveness. Section V concludes the paper and highlights directions for future research.

II. Background and Related Work

A. Large Language Models

Large Language Models (LLMs) are Artificial Intelligence models that have been trained to understand user input and also to generate text that matches the user's request [8]. These models are made up of millions to billions of parameters that have been trained on large datasets [9]. LLMs are able to produce content that is usually hard to tell apart from human-generated text, as well as recognize context and language variations [10]. ChatGPT is one of the most popular LLMs recently.

ChatGPT's GPT-4o model is one of the implementations of LLMs that is designed to generate humanized responses from a machine. GPT-4o was developed by OpenAI, a United States based research organization leady by Sam Altman at the time of writing. This model uses the transformer architecture that was already introduced in the original Generative Pre-Trained Transformer (GPT) language model [11]. Impressively, this model is able to do many things including responding to different types of questions, provide information, and communicate with users [12].

B. Conversational Recommender System

A recommendation system that provides interaction capabilities between the user and the system is a conversational recommendation system. With this, the user can provide specifications on the recommendations that are more in-depth, making the recommendations from the system more accurate. In addition, the system can also ask questions or clarifications to the user to gain a better understanding of the user's wishes [13]. The quality of recommendations can be improved compared to other recommendation systems as the system can communicate with the user to better understand their preferences [14][15].

The interaction between the system and user can enhance the preference gathering process, allow the user to provide feedback and enable them to ask questions about the recommendations. Interest in CRS has increased significantly in recent years. Major advances in the field of natural language processing, the presence of voice-controlled home assistants, and the greater use of chatbot technology have led to this development [16].

C. LLMs in CRS

In recent years, the use of LLMs in CRS has become an interesting field of research [3][17][18]. By combining LLM and CRS, the system can become smarter in terms of understanding conversations [14]. With this combination, the system can understand more difficult conversations and questions, thus provide better recommendations.

Zhang [19] claims that there are a number of advantages for recommender systems when LLMs are included in CRS. First, LLMs make user modeling more detailed and dynamic. Second, by producing more organic responses, LLMs can boost user involvement in the discussion and foster system confidence. Third, LLMs can enhance policy learning in CRS, which allows greater flexibility in response to user's preferences changing. Finally, LLMs are able to increase transparency and get better user experience by generating more informative and accurate responses. This all shows that the usage of LLMs greatly benefits CRS.

D. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is an approach in NLP where it combines the power of LLMs with external information to generate responses that are more accurate [20]. RAG utilizes relevant data from external sources to enhance the quality of the text generated by the LLMs. RAG is divided into three primary components which are Retrieval (methods to retrieve relevant documents or information based on the user's query), Augmentation (adding retrieved information to the user's query to enhance its relevance), and Generation (response of the LLM based on the augmented input) [21].

The retrieval technique in RAG can be used to retrieve relevant data on a large scale, such as metadata or documents, with high efficiency using tools like FAISS (Facebook AI Similarity Search) [22]. This allows the system to incorporate large amounts of external data into the LLM without the need to fine-tune all the data directly into the model.

E. Related Work

Previous research in the field of recommender systems has utilized interaction methods such as Form-based Layouts and C-NLP to develop CRS. Ayundhita et al. [6] developed a CRS with Form-based Layouts as the interaction method for recommending laptops, leveraging a mapping of users' functional needs to the technical specifications of products using ontology. This system shows success in helping users choose laptops based on functional needs, but Form-based Layouts have limitations in allowing users to express broader specifications and in understanding natural language. Baizal et al. [13] also developed a CRS using Form-based Layout to recommend products based on product functional requirements. The system maps users' functional needs to technical specifications using ontology, guiding users through a structured series of functional queries. However, like the study before, the usage of Form-based Layouts limits the user to predefined queries, limiting the users from making broader or nuanced user preferences. In contrast, Cherian et al. [7] developed SpecMaster, a CRS for laptops that uses a chatbot