

MARC: Multimodal and Multi-Task Agentic Retrieval-Augmented Generation for Cold-Start Recommender System^{*}

Seung Hwan Cho^{1,†}, Yujin Yang^{1,†}, Danik Baeck¹, Minjoo Kim¹, Young-Min Kim^{1,2,*},
Heejung Lee^{1,2} and Sangjin Park^{1,2}

¹Department of Industrial Data Engineering, Hanyang University, Republic of Korea

²School of Interdisciplinary Industrial Studies, Hanyang University, Republic of Korea

Abstract

Recommender systems (RS) are currently being studied to mitigate limitations during the cold-start conditions, by leveraging modality information or introducing Agent concepts based on the exceptional reasoning capabilities of Large Language Models (LLMs). Meanwhile, food and beverage recommender systems have traditionally used knowledge graph and ontology concepts due to the domain's unique data attribute and relationship characteristics. On this background, we propose MARC, a multimodal and multi-task cocktail recommender system based on Agentic Retrieval-Augmented Generation (RAG) utilizing graph database under cold-start conditions. The proposed system generates high-quality, contextually appropriate answers through two core processes: a task recognition router and a reflection process. The graph database was constructed by processing cocktail data from Kaggle, and its effectiveness was evaluated using 200 manually crafted questions. The evaluation used both LLM-as-a-Judge and human evaluation to demonstrate that answers generated via the graph database outperformed those from a simple vector database in terms of quality. The code is available at https://github.com/diddbwls/cocktail_rec_agentrag

Keywords

Recommender Systems, Agentic Retrieval-Augmented Generation, Knowledge Graph, Multimodal, Multi-Task, Cocktail

1. Introduction

In Natural Language Processing (NLP), Large Language Models (LLMs) are shifting research paradigms through their exceptional reasoning capabilities [1, 2]. LLMs demonstrate outstanding performance across diverse domains and tasks when combined with methodologies such as prompt engineering [3], fine-tuning [4], and Retrieval-Augmented Generation (RAG) [5]. Recent research in Recommender Systems (RS) have also leveraged these LLMs capabilities [6, 7, 8, 9], with studies introducing the concept of Agent being introduced [10, 11, 12]. It is important to note that the Agent concept is utilized extensively without clear boundaries in its definition [13]. This ambiguity can cause confusion in understanding research content. Therefore, this study bases its concepts on the definitions of Agent and Agentic RAG established by Wang et al. [14] and Singh et al. [15].

Recommender systems are information retrieval mechanisms that facilitate decision-making by providing internet users with information regarding their preferences and interests, in cases where the users lack the necessary information or expertise [16]. Following the widespread adoption of the internet, a substantial volume of diverse data has been generated online, thereby presenting users with the challenge of rapidly identifying the information they require. Consequently, recommender

RDGENAI '25: First International Workshop on Retrieval-Driven Generative AI at CIKM 2025, November 14th, 2025, Coex, Seoul, Republic of Korea

*Corresponding author.

[†]These authors contributed equally.

✉ shcho95@hanyang.ac.kr (S. H. Cho); yujinyang@hanyang.ac.kr (Y. Yang); dibaek@hanyang.ac.kr (D. Baeck);
kmj0921@hanyang.ac.kr (M. Kim); yngmnkim@hanyang.ac.kr (Y. Kim); stdream@hanyang.ac.kr (H. Lee);
psj3493@hanyang.ac.kr (S. Park)

ORCID 0000-0002-6914-901X (Y. Kim)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

systems have evolved to provide users with suitable content. Starting with recommender systems applying collaborative filtering methods between users [17], research has actively pursued performance enhancement from various perspectives, including modeling, scalability, and sequentiality [18, 19, 20]. These collaborative filtering-based methodologies have been shown to demonstrate excellent performance in warm-start characterized by abundant user-item interactions. Conversely, they have been observed to struggle to guarantee performance in cold-start with sparse interactions [21]. To mitigate this limitation, content-based research utilizing modality information such as text, images, and tables has been conducted [22, 23].

In application domains of recommender systems where data structures, attributes and relationships are critical, knowledge graph or ontology based methodologies are employed. A representative case is food and beverage recommender systems, where these approaches are used to accurately capture the intricate relationships between ingredients, recipes, containers, and other factors [24, 25, 26]. This suggests that while user interaction and history are important, meaningful recommendations can be made based solely on semantic similarity or relational information between content items. Moreover, knowledge graph has the advantage of being able to automatically explain the process of deriving answers, thereby enabling user understanding and reuse [24].

In this paper, we propose MARC, a multimodal and multi-task cocktail recommender system based on Agentic RAG using a graph database under cold-start conditions, aligned with recent research trends and domain specificity. The graph database used for RAG is cocktail data, which is based on the relationships between features. The proposed system has two main stages: Task Recognition Router and Reflection. Starting with the user’s image and text query (i.e., multimodal), the Router determines the task and performs retrieval through a configured algorithm based on this determination. The retrieved information is then evaluated by LLMs to verify that the data has been correctly fetched from and generated by the graph database. This Reflection continues until a configured threshold is exceeded. The resulting context from the Reflection is then fed into the model alongside the user’s query and prompt template. This enables the model to generate recommendations in response to the query. The main contributions of our study are as follows:

- We propose an Agentic RAG based recommender system that integrates multimodal and multi-task LLMs with Graph RAG for recommendations on cocktail data.
- We designed a Task Recognition Router and Reflection mechanism enabling multimodal and multi-task recommendations, thereby improving answer generation quality without requiring special fine-tuning.
- We demonstrate the effectiveness of the proposed method in the cocktail domain and confirm its potential as an interactive and explainable beverage recommender system.

2. Related Works

2.1. Beverage and Food Recommendation

Recommender systems based on food or beverages are not mainstream research compared to the movie domain, but they are steadily being studied, with research utilizing ontologies or knowledge graphs existing. Ontologies and knowledge graphs are representations that systematize information and relational data into a graph structure based on a knowledge system and schema of concept definitions and relationships through domain-specific vocabulary. They provide the ability to reflect constraints necessary for recommendation and offer explainability [27, 28, 29].

Chen et al. proposed a recommender system combining common-sense reasoning to detect emotions from inferences and colors based on a cocktail ontology knowledge base [30], while Ahlam et al. built an ontology based on the Canada Food Allergies and Intolerances Databases and integrated useful features such as food item selection, descriptions, and recommendations [31]. Haussmann et al. constructed a knowledge graph based on large-scale public food data and performed food recommendations [24]. Oliveira et al. constructed a wine ontology and confirmed that ontologies in recommender systems

influenced performance improvement [25]. Showafah et al. constructed an ontology containing knowledge about food and nutritional components along with nutritional requirements, and combined it with the TOPSIS method to provide optimal recommendations regarding nutritional balance and user preferences [32]. Chen et al. proposed a collaborative recipe knowledge graph, combining health suitability scoring between recipes and user preferences with an attention-based graph convolutional neural network to present a health-aware personalized recommendation model [33]. Gawrysiak et al. proposed WineGraph, which pairs food and wine using food and wine review data and augments FlavorGraph data [26].

2.2. Graph Retrieval Augmented Generation

Graph RAG is a method that addresses the limitations of vector-based RAG. By leveraging the semantic and structural relationships between nodes in a knowledge graph structure, it simultaneously improves performance and interpretability in multi-document reasoning and relationship-based question answering through search and inference [34]. Han et al. proposed GraphRAG, which constructs a graph from entities and relationships extracted from documents and integrates it as input for LLMs through subgraph search [34]. Hu et al. proposed GRAG, performing efficient subgraph exploration using a divide-and-conquer approach [35], while Shen et al. proposed GeAR, combining an agent framework with graph expansion [36]. Xiang et al. demonstrated through GraphRAG-Bench that graph structures are particularly effective for questions where relational information is crucial [37]. Agrawal et al. introduced query-based graph neural networks to enhance query-aware retrieval performance [38]. Liang et al. proposed a graph-based RAG specialized for the geospatial analysis domain, demonstrating domain-tailored applicability [39]. In summary, Graph RAG has complemented existing RAG in terms of accuracy, efficiency, and explainability, establishing itself as a core technology directly relevant to Agentic RAG design in this research.

2.3. LLM based Recommendation

In the early stages of recommender systems, Collaborative Filtering (CF) was applied to newsgroups to make recommendations based on user rating similarity [17], and content-based methodologies [18], Matrix Factorization (MF) [19], and deep learning-based methodologies [20] were introduced, leading to significant advancements. Building upon this prior research, this discussion will focus exclusively on the rapidly evolving recommender systems based on LLMs. Wang et al. proposed a zero-shot Next-Item Recommendation (NIR) prompting strategy integrating a three-step prompting approach to recommend movie ranking lists [6]. Lyu et al. proposed LLM-REC, integrating four unique prompting strategies, to enhance personalized text-based recommendations through text augmentation [7]. Tian et al. improved recommendation accuracy and relevance by processing multimodal information through LLMs and projecting it into an integrated latent space [9]. Finally, Wang et al., whose work is most similar to ours, aimed to enhance answer generation and recommendation quality by supplementing LLMs' hallucination issues with KG RAG [40]. Our research distinguishes itself by introducing an attempt to build an Agentic RAG framework that autonomously recognizes user queries and improves answer quality.

2.4. LLM as Recommendation Agent

Research utilizing LLM as agent has shifted beyond the existing static prompt-response paradigm, transforming it into a dynamic decision-making framework capable of managing complex systems through diverse sub-component configurations [41]. A typical LLM-based agent structure consists of four modules: Profile, Memory, Planning, and Action [40]. Research introducing agents based on this structural form is also underway in recommender systems, which can be categorized into recommender-oriented, interaction-oriented, and simulation-oriented approaches [42].

The recommender-oriented approaches focus on developing intelligent recommender systems with planning capabilities, making direct recommendations based on users' past behaviors [10, 8, 43]. Wang

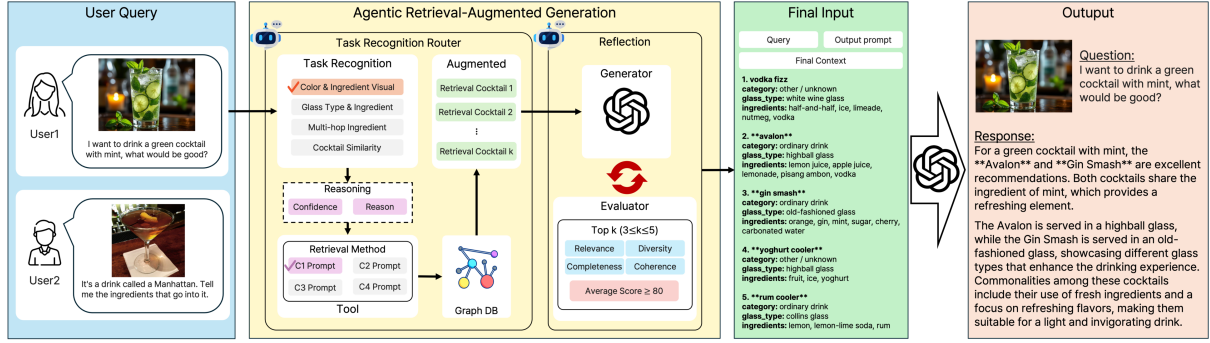


Figure 1: Overall workflow of the MARC

et al. proposed a self-inspiring algorithm to enhance planning capabilities, developing an agent capable of zero-shot personalized recommendations [10], while Shi et al. proposed a framework for learning planning capabilities at the macro level through reflector reflection and at the micro level through personalized recommendations via interactions between agent and critic [8]. Zhao et al. proposed a framework that sets LLMs as proxy users and recommends through tool learning, enabling the generation of recommendation lists aligned with preferences [43].

Interaction-oriented approaches enable tracking user preferences and explaining recommendation rationale through conversation [11, 44, 12]. Zeng et al. developed an interactive agent using LLMs and Answer Set Programming (ASP) that can request missing information [44], while Shu et al. proposed a human-centered recommendation framework based on a learn-execute-critique loop and a reflection mechanism to ensure alignment with user personality [11]. Huang et al. enhanced the agent component and integrated traditional recommender systems to build a multi-purpose interactive system [12].

The simulation-oriented approaches aim to simulate user behavior and item characteristics within recommender systems [45, 46, 47]. Zhang et al. proposed an LLM-based generative agent simulator equipped with user profiles, memory, and behavior modules [45]. Zhang et al. modeled user-item interactions and their relationships in a recommender system by treating users and items as agents and simulating their interactions through collaborative filtering [46]. Guo et al. proposed a framework that positions users and items within a knowledge graph in simulated recommendation scenarios and integrates into the simulation as natural language descriptions [47].

3. Proposed Method

The overall system proposed in this paper follows the structure and flow of Agentic RAG and is designed for cold-start conditions, where little or no interaction history exists between users and items, rather than for warm-start conditions with abundant interactions. First, we directly constructed a graph database based on cocktail relationships for RAG, which will be used to retrieve context for cocktail recommendations. Subsequently, the system receives the user’s cocktail modality information as input, recognizes the task type at the Router stage, and selects the retrieval algorithm. This enables graph database search and selects the top-k candidates. Finally, the Reflection stage sets an appropriate top-k for high-quality recommendations and derives the final context. Both stages are performed by LLMs. Figure 1 illustrates the overall workflow of the MARC. The following subsections provide detailed of each component, including graph database construction, the Router, Graph RAG, and Reflection.

3.1. Graph Database Construction

Graph Structure. The Kaggle cocktail dataset is publicly available and features are well-organized, making it easy to build a relational database. Based on this, we collected data including cocktail recipes, ingredients, and metadata. We parsed the ingredient lists and measurements from the raw data, normalized them, and consolidated duplicate ingredient names. For items that were missing or

Table 1
Node, Property, and Relation in Cocktail Knowledge Graph

Node	Property	Relation	
		Cypher	Definition
Cocktail	id, name, name_embedding, alcoholic, ingredients, drinkThumbnail, ingredientMeasures, description, image description, image description_embedding, instructions, instructions_embedding	-	Main Node
Ingredients	name, name_embedding	(Cocktail)-[:HAS_INGREDIENT {measure: "quantity"}]->(Ingredient)	Represents the relationship between cocktails and ingredients, storing the quantity of each ingredient using the measure property.
Category	name, name_embedding	(Cocktail)-[:CATEGORY]->(Category)	Connects the category to which the cocktail belongs.
GlassType	name, name_embedding	(Cocktail)-[:HAS_GLASSTYPE]->(GlassType)	Indicates the type of glass in which the cocktail is served.

did not exist during this process (e.g., visual image descriptions of cocktails, cocktail explanations), we manually constructed them. Specifically, to express the visual characteristics of each cocktail image as text, we used the Qwen2.5-VL-7B [48] to generate descriptions. These descriptions include visual elements such as color, texture, and garnish.

A graph database was constructed based on preprocessed data. The graph consists of nodes and relationships, enabling structured information to be effectively retrieved and input as context. Nodes comprise Cocktail, Ingredients, Category, and GlassType. Except for Cocktail, the remaining nodes possess only two properties: name and name_embedding, which is an embedding of the name property. Relationships between nodes are expressed centered around the main node, Cocktail. The cocktail knowledge graph consists of 4 node types and 3 relation types. The central node, Cocktail, possesses 12 properties. Ingredient, Category, and GlassType nodes connect to it. This graph structure effectively represents the complex relationships within cocktails, enabling multi-hop search and relationship-based recommendations. It particularly facilitates the discovery of new cocktails through common usage patterns among ingredients and supports multi-faceted searches via category and glass type. A detailed explanation of the graph structure can be found in Table 1.

Vector Representation. Vector representations are used in various processes during search, including seed node selection, to calculate similarity and select candidates. To achieve this, embedding vectors are generated, and the resulting values are stored as properties of graph nodes. Cocktail, Category, and GlassType were embedded to calculate semantic similarity between item names, while image description was embedded to calculate visual similarity. This enables real-time calculation of cosine similarity between stored embedding values and user queries, allowing for effective retrieval of semantically related cocktails.

3.2. Agentic RAG: Task Recognition Router, Graph RAG, Reflection

The main stages of the Agentic RAG used in this study are the Task Recognition Router and Reflection. Task recognition is the Router that determines which search method to use based on the user’s question intent. Reflection is the process where LLMs evaluate the quality of search results based on the results themselves, then decide whether to expand the candidate pool by adjusting the top-k value if necessary.

Task Recognition Router. The Task Recognition Router analyzes the user’s natural language query and classifies query into the most suitable task among four defined tasks (C1–C4). Each task employs a search algorithm optimized for a specific aspect of cocktail recommendation and is defined as follows:

- **C1 (Color-Ingredient Visual Search):** Processes questions related to color keywords ("red", "golden", "blue", etc.) or visual appearance ("elegant", "layered", "beautiful"). The Router detects based on the association between color and ingredients, and visual appeal.
- **C2 (Glass Type with Ingredient Matching):** Processes queries specifying a particular glass type ("highball glass", "martini glass", etc.) and considers combinations of glass and ingredients.
- **C3 (Multi-hop Ingredient Expansion):** Performs multi-hop graph exploration when only an ingredient list is provided ("cocktails made with whiskey and vermouth").
- **C4 (Cocktail Similarity and Alternative):** When a specific cocktail is explicitly targeted ("cocktails similar to Manhattan"), performs relationship-based similarity search.

The Router utilizes an LLMs to analyze the intent and keywords of user queries, returning task recognition results in JSON format along with confidence scores and reasons. This enables the automatic selection and execution of the most suitable search algorithm for each query.

Graph RAG. We implemented graph based retrieval algorithms optimized for each task. Beyond simply using graphs as data storage, the algorithms perform relationship based retrieval by leveraging the structural characteristics and relational information of graph database.

- **C1:** Visual feature-based ingredient matching and relationships between cocktails
- **C2:** Glass type and ingredient relationships
- **C3:** Multi-hop relationships between ingredients and cocktails
- **C4:** Ingredient and recipe complexity relationships

This effectively satisfies diverse user search intents and enables the discovery of semantically related cocktails. Each recognized task employs a specific retrieval method, with detailed retrieval approaches specified in the Appendix A.1.

Reflection. Reflection is a reflective reasoning process that evaluates the quality of retrieved cocktail candidates and selects the optimal recommendation result. Reflection process is as follows:

1. **Multi-Candidate Generation:** Each search algorithm generates top-k candidates. At this stage, different hyperparameter settings (similarity threshold, expansion depth, etc.) are applied to ensure diversity.
2. **Quality Assessment:** LLMs evaluate each candidate set. The evaluation criteria include the following: Relevance to user query (Relevance score), Diversity of recommendation results (Diversity score), Completeness of information (Completeness score), Consistency of explanation (Coherence score).
3. **Iterative Refinement:** If the set quality threshold is not met, LLMs gradually expand the top-k value (a search parameter) to perform a re-search (top-k, top-(k+1), top-(k+2), ...). This process repeats up to three times, improving the diversity and quality of candidate cocktails in each iteration compared to the previous results.
4. **Final Selection:** The result with the highest overall score is ultimately selected from a set of candidates. LLMs present the recommended result to the user along with the reason for the selection.

This Reflection goes beyond simply returning search results, ensuring high-quality recommendations that accurately capture user intent and fit the context. It enables meaningful recommendations even in cold-start conditions by leveraging the relational information within the graph structure.

4. Experiment and Results

4.1. Dataset

The dataset used for constructing the graph database was created by combining Kaggle datasets¹ and removing duplicates. Additionally, image descriptions for each cocktail were generated using Qwen-2.5-VL-7B. Description data corresponding to the cocktail descriptions was also manually constructed. Items for which reliable descriptions could not be obtained were excluded, resulting in a final dataset comprising 436 cocktails.

For evaluation, a separate question dataset consisting of 200 manually designed queries was constructed. These queries were evenly distributed across four task types (50 per task), including 140 multimodal questions and 60 text-only questions. Each task contains 35 items with images, accounting for approximately 70% of the total items per task. The images used in the experiment were separately constructed to avoid overlap with images in the database.

4.2. Experiment Configuration

Implementation Details. The graph database for Graph RAG was constructed using Neo4j², and embeddings were processed using OpenAI’s text-embedding-3-small³. All LLMs corresponding to each task were processed using GPT-4o-mini.

Metrics. LLM-as-a-Judge is a method for evaluating and improving LLMs, where LLMs assess the outputs of other models as if they were being evaluated by humans [49]. The approach of utilizing LLMs as evaluators in recommender systems was first proposed by Zhang et al. [50], who demonstrated its validity through comparative validation against human evaluations for the same items. While existing evaluation methods have limitations in terms of cost, time, and scalability, LLMs enable efficient and reproducible large-scale evaluations based on their exceptional language understanding and reasoning capabilities. Therefore, this study adopts the LLM-as-a-Judge approach, considering these advantages. The experiment was conducted across four categories: Persuasiveness, Transparency, Accuracy, and Satisfaction, with participants asked to rate each on a scale of 1 to 5 points. Simultaneously, we conduct human evaluation as performed in prior studies, using the same categories and scale for assessment. Detailed descriptions of the evaluation categories and each scale are provided in Appendix A.3.

Table 2

LLM-as-a-judge scores by evaluation models and methods. Bold text indicates the highest score in each category.

Evaluation	Method	Persuasiveness	Transparency	Accuracy	Satisfaction	Average
GPT-4o-mini ⁴	w/o Graph, Reflection	3.64	4.45	3.41	3.63	3.78
	w/o Reflection	4.06	4.71	3.73	4.06	4.14
	w/o Graph	3.84	4.57	3.59	3.85	3.96
	MARC (Ours)	4.12	4.70	3.73	4.11	4.17
GPT-5 ⁴	w/o Graph, Reflection	1.93	2.84	1.73	1.69	2.05
	w/o Reflection	2.16	3.01	1.99	1.92	2.27
	w/o Graph	1.97	2.82	1.81	1.78	2.09
	MARC (Ours)	2.19	2.97	1.98	1.93	2.27
Human	w/o Graph, Reflection	4.16	4.21	4.13	3.82	4.08
	w/o Reflection	4.39	4.47	4.28	4.17	4.33
	w/o Graph	4.36	4.40	4.37	4.22	4.34
	MARC (Ours)	4.35	4.46	4.45	4.26	4.38

¹<https://www.kaggle.com/datasets/ai-first/cocktail-ingredients>, <https://www.kaggle.com/datasets/aadyasingh55/cocktails>, <https://www.kaggle.com/datasets/joakark/cocktail-ingredients-and-instructions>

²<https://neo4j.com>

³<https://platform.openai.com/docs/models/text-embedding-3-small>

⁴<https://openai.com/api/>

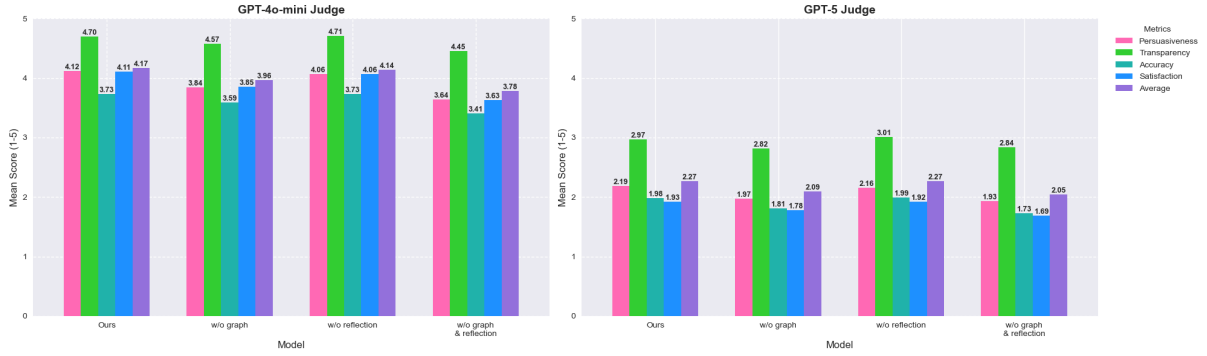


Figure 2: Visualization of Evaluation Results using LLM-as-a-Judge (Left: GPT-4o-mini, Right: GPT-5)

4.3. Results

The quantitative performance of MARC is reported using two evaluation metrics. First, we present the results of the LLM-as-a-Judge based automated evaluation, followed by the results of the Human Evaluation conducted using the same metrics. We performed ablation tests on 200 identical responses generated by GPT-4o-mini, comparing MARC (Ours), vector RAG only (w/o Graph), without Reflection (w/o Reflection), and vector RAG only and without Reflection (w/o Graph, Reflection) approaches. To ensure fairness, 13 questions that yielded no search results due to database limitations were excluded. For the remaining 187 questions, we performed LLM-as-a-Judge and the evaluation models used were GPT-4o-mini and GPT-5, cross-validated against each other.

For LLM-as-a-Judge in both evaluation models, MARC achieved the highest overall average score. In the GPT-4o-mini evaluation, MARC achieved an average of 4.17 points, while other approaches scored 3.78, 3.96, and 4.14 points, showing relative improvements ranging from +0.39 to +0.03. The GPT-5 evaluation also showed a relative improvement, with MARC recording an average of 2.27 points compared to other approaches averaging 2.05, 2.09, and 2.27 points, corresponding to gains of +0.22 to +0.00. Comparing the evaluation models, GPT-5 was relatively conservative. This result is consistent with the findings of Abdoli et al. [51], which reported that GPT-5 tends to assign conservative scores and shows higher variability compared to other GPT versions. Nevertheless, the fact that the Transparency metric consistently outperformed GPT-5 results suggests that the graph structure more clearly reveals the causal pathways underlying recommendations, thereby enhancing the traceability of explanations (why-explanation clarity).

Human evaluation was conducted by four individuals. A total of 80 samples were randomly shuffled and presented, consisting of 10 questions per task extracted for all approaches. Similar to LLM-as-a-Judge, three cases where search results were unavailable due to database limitations were excluded, resulting in a final evaluation of 154 questions. Human evaluation results showed MARC achieved an average of 4.38 points, while the other approaches scored 3.78, 3.96, and 4.14 points, showing relative improvements ranging from +0.30 to +0.04. In the human evaluation, while the overall average and metrics such as Transparency, Accuracy, and Satisfaction generally increased, Persuasiveness showed a slight decline compared to w/o Reflection and w/o Graph. This indicates that non-structural elements such as writing style, tone, and summary density influence the impression of the response on actual humans, suggesting the need for additional calibration.

We compared all the approaches, MARC was the best overall in all the evaluation models. However, the results suggest that the Graph RAG and database contributed most significantly to performance improvement. The incorporation of the Reflection further enhanced overall quality but showed marginal or inconsistent effects in certain metrics. This tendency can be interpreted from two perspectives. First, regarding top-k selection, the hyperparameter was fixed at 3 (tested within the range of 3–5), which ensures stable performance but may have limited the potential benefit of the Reflection loop. If top-k = 1 had been applied, a stronger positive effect of Reflection could likely have been observed, since the Reflection would have played a more decisive role in expanding candidate diversity. Second, the

threshold for response quality was also set empirically at 80 and was not systematically tuned. A stricter threshold could have provided more refined filtering and thus higher-quality responses. Therefore, while both modules jointly contribute to performance, the graph structure is the dominant factor driving improvement, and optimizing Reflection-related hyperparameters remains an important direction for future work.

5. Conclusion

This study proposes MARC for cold-start by constructing a graph database based on cocktail data. The Task Recognition Router formalizes user queries into tasks, and responses are generated based on the threshold set in Reflection. Results from LLM-as-a-Judge and human evaluation show that our approach achieved overall balanced improvements. Although the human evaluation revealed a slight decrease in Persuasiveness, implying that linguistic style and prompt formulation still influence how users perceive the responses. Addressing these linguistic and stylistic factors will be an important direction for future prompt refinement. Limitations include: applied only to cocktail data, being restricted to cold-start conditions, utilizing only a graph database, employing a non-refined threshold, and relying on text-based similarity without image embeddings. Future research will focus on other domains (e.g. movies), warm-start conditions, implementing a Multi-agent-based RAG and reflection, and developing an end-to-end system using image embeddings.

References

- [1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.
- [2] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, A survey of large language models, *arXiv preprint arXiv:2303.18223* 1 (2023).
- [3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, *OpenAI blog* 1 (2019) 9.
- [4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models., *ICLR* 1 (2022) 3.
- [5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, Retrieval-augmented generation for knowledge-intensive nlp tasks, *Advances in neural information processing systems* 33 (2020) 9459–9474.
- [6] L. Wang, E.-P. Lim, Zero-shot next-item recommendation using large pretrained language models, *arXiv preprint arXiv:2304.03153* (2023).
- [7] H. Lyu, S. Jiang, H. Zeng, Y. Xia, Q. Wang, S. Zhang, R. Chen, C. Leung, J. Tang, J. Luo, Llm-rec: Personalized recommendation via prompting large language models, *arXiv preprint arXiv:2307.15780* (2023).
- [8] Large language models are learnable planners for long-term recommendation, 2024.
- [9] Mmrec: Llm based multi-modal recommender system, *IEEE*, 2024.
- [10] Y. Wang, Z. Jiang, Z. Chen, F. Yang, Y. Zhou, E. Cho, X. Fan, X. Huang, Y. Lu, Y. Yang, Recmind: Large language model powered agent for recommendation, *arXiv preprint arXiv:2308.14296* (2023).
- [11] Y. Shu, H. Zhang, H. Gu, P. Zhang, T. Lu, D. Li, N. Gu, Rah! recsys–assistant–human: A human-centered recommendation framework with llm agents, *IEEE Transactions on Computational Social Systems* 11 (2024) 6759–6770.
- [12] X. Huang, J. Lian, Y. Lei, J. Yao, D. Lian, X. Xie, Recommender ai agent: Integrating large language models for interactive recommendations, *ACM Transactions on Information Systems* 43 (2025) 1–33.
- [13] B. Bent, The term ‘agent’ has been diluted beyond utility and requires redefinition, *arXiv preprint arXiv:2508.05338* (2025).

- [14] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, A survey on large language model based autonomous agents, *Frontiers of Computer Science* 18 (2024) 186345.
- [15] A. Singh, A. Ehtesham, S. Kumar, T. T. Khoei, Agentic retrieval-augmented generation: A survey on agentic rag, *arXiv preprint arXiv:2501.09136* (2025).
- [16] Y. Wang, W. Ma, M. Zhang, Y. Liu, S. Ma, A survey on the fairness of recommender systems, *ACM Transactions on Information Systems* 41 (2023) 1–43.
- [17] Grouplens: An open architecture for collaborative filtering of netnews, 1994.
- [18] Item-based collaborative filtering recommendation algorithms, 2001.
- [19] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, *Computer* 42 (2009) 30–37.
- [20] Deep neural networks for youtube recommendations, 2016.
- [21] M. Volkovs, G. Yu, T. Poutanen, Dropoutnet: Addressing cold start in recommender systems, *Advances in neural information processing systems* 30 (2017).
- [22] X. Li, B. Chen, L. Hou, R. Tang, Ctrl: Connect tabular and language model for ctr prediction, *CoRR* (2023).
- [23] Where to go next for recommender systems? id-vs. modality-based recommender models revisited, 2023.
- [24] FoodKG: a semantics-driven knowledge graph for food recommendation, Springer, 2019.
- [25] L. Oliveira, R. Rocha Silva, J. Bernardino, Wine ontology influence in a recommendation system, *Big Data and Cognitive Computing* 5 (2021) 16.
- [26] WineGraph: A Graph Representation for Food-Wine Pairing, Springer, 2024.
- [27] The construction of domain ontology and its application to document retrieval, Springer, 2004.
- [28] R.-C. Chen, Y.-H. Huang, C.-T. Bau, S.-M. Chen, A recommendation system based on domain ontology and swrl for anti-diabetic drugs selection, *Expert Systems with Applications* 39 (2012) 3995–4006.
- [29] Y. Chen, L. Wu, M. J. Zaki, Bidirectional attentive memory networks for question answering over knowledge bases, *arXiv preprint arXiv:1903.02188* (2019).
- [30] Y.-H. Chen, T.-h. Huang, D. C. Hsu, J. Y.-j. Hsu, Colorcocktail: an ontology-based recommender system, *Proceedings of 20th American Association for Artificial Intelligence. Menlo Park, USA: AAAI Press* (2006) 79–82.
- [31] A. Al, J. Fiaidhi, S. Mohammed, Ontology-based food recommendation system for seniors: A peer to peer networking approach, *International Journal of Advanced Science and Technology* 123 (2019) 41–46. doi:10.33832/ijast.2019.123.05.
- [32] M. Showafah, S. W. Sihwi, Ontology-based daily menu recommendation system for complementary food according to nutritional needs using naïve bayes and topsiis, *International Journal of Advanced Computer Science and Applications* 12 (2021).
- [33] Y. Chen, Y. Guo, Q. Fan, Q. Zhang, Y. Dong, Health-aware food recommendation based on knowledge graph and multi-task learning, *Foods* 12 (2023) 2079.
- [34] H. Han, Y. Wang, H. Shomer, K. Guo, J. Ding, Y. Lei, M. Halappanavar, R. A. Rossi, S. Mukherjee, X. Tang, Retrieval-augmented generation with graphs (graphrag), *arXiv preprint arXiv:2501.00309* (2024).
- [35] Y. Hu, Z. Lei, Z. Zhang, B. Pan, C. Ling, L. Zhao, Grag: Graph retrieval-augmented generation, in: *North American Chapter of the Association for Computational Linguistics*, 2024. URL: <https://api.semanticscholar.org/CorpusID:270062608>.
- [36] Z. Shen, C. Diao, P. Vougiouklis, P. Merita, S. Piramanayagam, E. Chen, D. Graux, A. Melo, R. Lai, Z. Jiang, Gear: Graph-enhanced agent for retrieval-augmented generation, *arXiv preprint arXiv:2412.18431* (2024).
- [37] Z. Xiang, C. Wu, Q. Zhang, S. Chen, Z. Hong, X. Huang, J. Su, When to use graphs in rag: A comprehensive analysis for graph retrieval-augmented generation, *arXiv preprint arXiv:2506.05690* (2025).
- [38] V. Agrawal, F. Wang, R. Puri, Query-aware graph neural networks for enhanced retrieval-

- augmented generation, arXiv preprint arXiv:2508.05647 (2025).
- [39] J. Liang, S. Hou, H. Jiao, Y. Qing, A. Zhao, Z. Shen, L. Xiang, H. Wu, Geographrag: A graph-based retrieval-augmented generation approach for empowering large language models in automated geospatial modeling, *International Journal of Applied Earth Observation and Geoinformation* 142 (2025) 104712.
 - [40] S. Wang, W. Fan, Y. Feng, S. Lin, X. Ma, S. Wang, D. Yin, Knowledge graph retrieval-augmented generation for llm-based recommendation, arXiv preprint arXiv:2501.02226 (2025).
 - [41] S. G. Patil, T. Zhang, X. Wang, J. E. Gonzalez, Gorilla: Large language model connected with massive apis, *Advances in Neural Information Processing Systems* 37 (2024) 126544–126565.
 - [42] Q. Peng, H. Liu, H. Huang, Q. Yang, M. Shao, A survey on llm-powered agents for recommender systems, arXiv preprint arXiv:2502.10050 (2025).
 - [43] Let me do it for you: Towards llm empowered recommendation via tool learning, 2024.
 - [44] Automated interactive domain-specific conversational agents that understand human dialogs, Springer, 2024.
 - [45] On generative agents in recommendation, 2024.
 - [46] Agentcf: Collaborative learning with autonomous language agents for recommender systems, 2024.
 - [47] T. Guo, C. Liu, H. Wang, V. Mannam, F. Wang, X. Chen, X. Zhang, C. K. Reddy, Knowledge graph enhanced language agents for recommendation, arXiv preprint arXiv:2410.19627 (2024).
 - [48] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, Qwen2. 5-vl technical report, arXiv preprint arXiv:2502.13923 (2025).
 - [49] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, Judging llm-as-a-judge with mt-bench and chatbot arena, *Advances in neural information processing systems* 36 (2023) 46595–46623.
 - [50] Large language models as evaluators for recommendation explanations, 2024.
 - [51] S. Abdoli, R. Cilibrasi, R. Al-Shikh, Understanding ai evaluation patterns: How different gpt models assess vision-language descriptions, arXiv preprint arXiv:2509.10707 (2025).

A. Appendix

A.1. Specific Graph Retrieval Algorithm

In this part, we described the topics covered in section 3.2. In particular, we have thoroughly covered the methods and processes for retrieving graph database, including examples.

C1: Color-Ingredient Visual Search Algorithm. We perform hybrid search combining graph structure and embeddings, focusing on the visual characteristics of cocktails. We extract the cocktail name, ingredients, category, and color from the user query. We then embed the user query to calculate cosine similarity with the image description embeddings, selecting the top-k cocktail nodes as seeds for graph exploration. This effectively reduces the overall graph search space. Next, it searches for ingredient nodes semantically associated with color keywords extracted from the user query (e.g., "red", "orange"). It calculates the similarity between the color keyword embedding and the ingredient name embedding to identify the most relevant ingredient, then expands the graph to include cocktails containing that ingredient. After integrating the initial seed and the expanded cocktails, we remove duplicates and recalculate the similarity between the image description embeddings of all candidates and the user query. Finally, we sort them by similarity and select the top-k cocktails. By leveraging both ingredient and image description information, we recommend cocktails that best match the user’s visual intent.

C2: Glass Type-Ingredient Matching. We explore the graph structure centered around glass type and ingredients. From user queries explicitly mentioning a glass type, extract the GlassType, Cocktail, Ingredients, and Category. First, filter Cocktail nodes matching the extracted GlassType. Then perform Ingredient matching. At this step, search for cocktails matching all ingredients. If insufficient results

are found, gradually relax the ingredient condition by excluding one ingredient at a time, starting with the last ingredient extracted from the user query. This iteration continues until k candidate cocktails are secured. Finally, the final cocktail is selected by sorting the similarity scores between the extracted cocktail's image description embedding and the user query in descending order. This approach provides results aligned with the user intent to explore cocktails with similar ingredients within the same glass type.

C3: Multi-hop Ingredient Expansion Search. C3 discovers new cocktails across up to 3 hops by exploring ingredient relationships through multi-hop graph traversal. First, it extracts cocktails and ingredients from the user query. At the 1-hop level, it searches for Cocktail nodes directly connected to the input ingredient via a HAS_INGREDIENT relationship. Only cocktails containing at least n matching ingredients are selected. At the 2-hop level, it analyzes ingredient patterns common to cocktails discovered in the first hop. For example, it uncovers hidden association patterns between ingredients, such as "Cocktails using mint and lime also commonly use rum." Only cocktails sharing at least n common ingredients proceed to the next step. At the 3-hop level, it explores new cocktails based on the discovered common ingredients. For each new cocktail found at the 3-hop level, it calculates the following two scores:

- **Extension strength:** How many of the common ingredients found in 2-hop are included in this cocktail?
- **Ingredient bonus:** How many of the original ingredients entered by the user are included in this cocktail?

The total score obtained by summing the counts from these two components is used to sort the results in descending order. In case of a tie, cocktails with higher expansion strength are ranked higher. The candidate cocktails selected through this multi-hop expansion are finally ranked in descending order by calculating the cosine similarity between the user query and the image description embedding of each cocktail. This multi-hop search recommends cocktails related by ingredients that are difficult to discover through direct ingredient matching. It uses cosine similarity to recommend cocktails, considering both the structural characteristics of the graph and semantic similarity.

C4: Material-based Similar Recipe Cocktail Recommendation. It comprehensively analyzes graph-based similarity and ingredient complexity to recommend alternative cocktail recipes. First, it extracts the target cocktail and ingredients from the user query as seeds. Prioritize exploring Cocktail nodes matching the seed cocktail extracted from the database. If the seed cocktail cannot be found in the user query, embed the entire query and return cocktails with high cosine similarity to the name_embedding as potential seeds. Then, explore other cocktails sharing ingredients identical to those in the seed cocktail. For each cocktail, the number of ingredients shared with the seed cocktail is calculated, while also determining the total number of ingredients in that cocktail. During the complexity analysis phase, only cocktails where the difference in ingredient count between the seed cocktail and the candidate cocktail is within m are selected. Additionally, only cocktails sharing at least x ingredients are retained. The final ranking is sorted in descending order based on the number of shared ingredients. In case of a tie, the top- k simpler cocktails with fewer ingredients are selected. This graph relationship enables the recommendation of alternative cocktails with similar ingredient compositions and appropriate complexity. C4 focuses on structural similarity rather than semantic similarity, providing practical recommendations that consider actual manufacturability and taste similarity.

A.2. Task-specific Prompt Template

This section presents the original text of the prompt used in the framework of this study. The prompts for each process and task within the framework are written in English, accompanied by brief descriptions of each template. Only the Reflection prompt is attached below and detailed prompts are available on the publicly released GitHub repository.

Listing 1: Reflection Prompt Template

```
REFLECTION_PROMPT_TEMPLATE = """You are a quality evaluation expert for a cocktail recommender system.
Based on the user's query and the retrieved cocktail candidates, evaluate the quality according to the
following four criteria.

## User Query
{user_query}

## Retrieved Cocktail Candidates ({num_results} items)
{search_results}

## Evaluation Criteria
1. **Relevance**: How relevant are the search results to the user's query? (0-100)
   - Do they reflect the user's desired characteristics (color, flavor, ingredients, style, etc.)?
   - Do they satisfy the core requirements of the query?

2. **Diversity**: How diverse are the recommended cocktails? (0-100)
   - Do they offer different styles, flavors, and ingredient combinations?
   - Do they provide a broad range of options rather than monotonous results?

3. **Completeness**: How comprehensive are the recommendations? (0-100)
   - Do they sufficiently provide the information the user is seeking?
   - Is there a possibility that better alternatives are missing?

4. **Coherence**: How logically consistent are the recommendations? (0-100)
   - Are the reasons for recommendation clear and valid?
   - Do the recommendations form a harmonious set overall?

## Output Format (JSON)
{
  "relevance": 85,
  "diversity": 70,
  "completeness": 80,
  "coherence": 90,
  "overall_score": 81.25,
  "feedback": "High relevance but low diversity. Many cocktails have similar styles; providing more varied
    options would be better.",
  "suggestions": [
    "Increase diversity in color or base spirits",
    "Consider adding alcoholic/non-alcoholic options"
  ],
  "should_retry": False
}

Each score should be between 0 and 100, with overall_score as the average of the four scores.
Set should_retry to true if the overall score is below 80.
"""
```

Reflection Template. This template performs reflection based on generated responses to expand the number of retrieved cocktails and enhance context. It is the second core step proposed in this study. If the set threshold is not exceeded, the process repeats until it is, ensuring high-quality information retrieval and enabling input into LLMs' context.

A.3. Detailed Description of Evaluation Metrics

The experiment was conducted using the following items:

- **Persuasiveness:** "This explanation is convincing to me."
- **Transparency:** "Based on this explanation, I understand why this movie is recommended."
- **Accuracy:** "This explanation is consistent with my interests."
- **Satisfaction:** "I am satisfied with this explanation."

For each item, LLMs instructed to evaluate on a scale of 1 to 5 points. 1 point was set as Strongly Disagree, 2 points as Disagree, 3 points as Neutral, 4 points as Agree, and 5 points as Strongly Agree. In other words, 1 point is the lowest score, and 5 points is the highest score. These evaluation items were applied identically to the Human Evaluation process.