PersonaRAG: Enhancing Retrieval-Augmented Generation Systems with User-Centric Agents

Saber Zerhoudi[1], Michael Granitzer[1]

[1]*University of Passau, Passau, Germany*

Abstract

Large Language Models (LLMs) struggle with generating reliable outputs due to outdated knowledge and hallucinations. Retrieval-Augmented Generation (RAG) models address this by enhancing LLMs with external knowledge, but often fail to personalize the retrieval process. This paper introduces PersonaRAG, a novel framework incorporating user-centric agents to adapt retrieval and generation based on real-time user data and interactions. Evaluated across various question answering datasets, PersonaRAG demonstrates superiority over baseline models, providing tailored answers to user needs. The results suggest promising directions for user-adapted information retrieval systems. Findings and resources are available at https://github.com/padas-lab-de/ir-rag-sigir24-persona-rag.

Keywords

User interactions, Retrieval-Augmented Generation (RAG), Personalized Information Retrieval, Multi-Agent RAG

1. Introduction

Large Language Models (LLMs) such as GPT-4 [2] and LLaMA 3 [3] have significantly advanced the field of natural language processing (NLP) by demonstrating impressive performance across various tasks and exhibiting emergent abilities that push the boundaries of artificial intelligence [4]. However, these models face challenges such as generating unreliable outputs due to issues like hallucination and outdated parametric memories [5].

Retrieval-Augmented Generation (RAG) models have shown promise in addressing these issues by integrating externally retrieved information to support more effective performance on complex, knowledge-intensive tasks [6]. Despite these advancements, the deployment of RAG systems within broader AI frameworks continues to face significant challenges, particularly in handling noise and irrelevance in retrieved data [7].

A key limitation of existing RAG systems is their inability to adapt outputs to users' specific informational and contextual needs. Personalized techniques in information retrieval, such as adaptive retrieval based on user interaction data and context-aware strategies, are increasingly recognized as essential for enhancing user interaction and satisfaction [8, 9]. These methods aim to refine the retrieval process dynamically, tailoring it more closely to individual user profiles and situational contexts [10].

The integration of agent-based systems with personalized RAG architectures presents a compelling avenue for research. Such systems utilize a multi-agent framework to simulate complex, adaptive interactions tailored to user-specific requirements [11]. By embedding intelligent, user-oriented agents within the RAG framework, these systems can evolve into more sophisticated tools that not only retrieve relevant information but also align it closely with the user's specific preferences and contexts in real-time. Importantly, the personalization strategy employed in these systems is fully transparent to the user, ensuring that the user is aware of how their information is being used to tailor the results.

In this study, we present PersonaRAG, an innovative methodology that extends traditional RAG frameworks by incorporating user-centric agents into the retrieval process. This approach addresses the previously mentioned limitations by promoting active engagement with retrieved content and utilizing dynamic, real-time user data to continuously refine and personalize interactions. PersonaRAG aims to enhance the precision and relevance of LLM outputs, adapting dynamically to user-specific needs while maintaining full transparency regarding the personalization process.

Our experiments, conducted using GPT-3.5, develop the PersonaRAG model and evaluate its performance across various question answering datasets. The results indicate that PersonaRAG achieves an improvement of over 5% in accuracy compared to baseline models. Furthermore, PersonaRAG demonstrates an ability to adapt responses based on user profiles and information needs, enhancing the personalization of results. Additional analysis shows that the principles underlying PersonaRAG can be generalized to different LLM architectures, such as Llama 3 70b and Mixture of Experts (MoE) 8x7b [12]. These architectures benefit from the integration of external knowledge facilitated by PersonaRAG, with improvements exceeding 10% in some cases. This evidence indicates that PersonaRAG not only contributes to the progress of RAG systems but also provides notable advantages for various LLM applications, signifying a meaningful step forward in the development of more intelligent and user-adapted information retrieval systems.

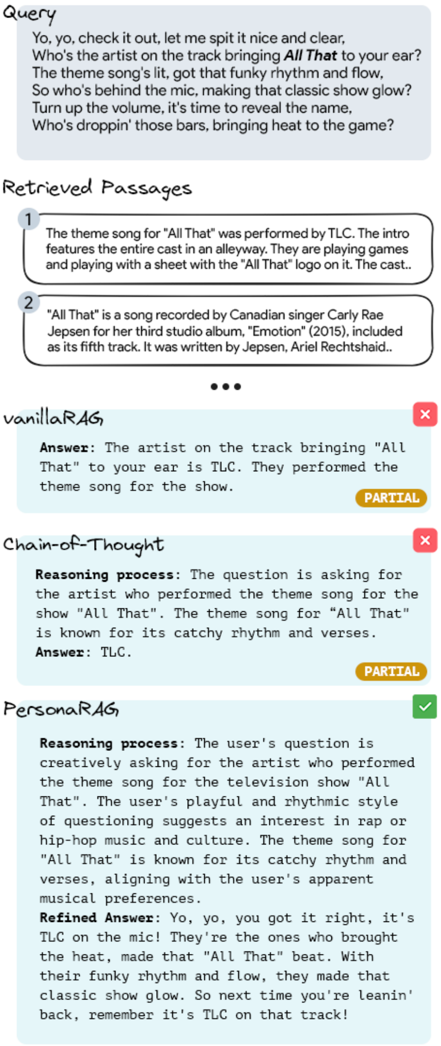
2. Related Work

Retrieval-Augmented Generation (RAG) systems have emerged as a significant advancement in natural language processing and machine learning, enhancing language models by integrating external knowledge bases to improve performance across various tasks, such as question answering, dialog understanding, and code generation [6, 13]. These systems employ dense retrievers to pull relevant information, which the language model then uses to generate responses. However, the development of RAG systems and their integration within broader artificial intelligence frameworks is an ongoing area of research, with several challenges and opportunities for improvement.

Recent developments in RAG systems have focused on refining these models to better handle the noise and irrelevant information often retrieved during the process. Xu et al.

Information Retrieval's Role in RAG Systems (IR-RAG) workshop at SIGIR, 2024, Washington D.C., USA

saber.zerhoudi@uni-passau.de (S. Zerhoudi); michael.granitzer@uni-passau.de (M. Granitzer)
0000-0003-2259-0462 (S. Zerhoudi); 0000-0003-3566-5507 (M. Granitzer)



Query

Yo, yo, check it out, let me spit it nice and clear,
Who's the artist on the track bringing ***All That*** to your ear?
The theme song's lit, got that funky rhythm and flow,
So who's behind the mic, making that classic show glow?
Turn up the volume, it's time to reveal the name,
Who's droppin' those bars, bringing heat to the game?

Retrieved Passages

1 The theme song for "All That" was performed by TLC. The intro features the entire cast in an alleyway. They are playing games and playing with a sheet with the "All That" logo on it. The cast..

2 "All That" is a song recorded by Canadian singer Carly Rae Jepsen for her third studio album, "Emotion" (2015), included as its fifth track. It was written by Jepsen, Ariel Rechtshaid..

•••

vanillaRAG ✗

Answer: The artist on the track bringing "All That" to your ear is TLC. They performed the theme song for the show.

PARTIAL

Chain-of-Thought ✗

Reasoning process: The question is asking for the artist who performed the theme song for the show "All That". The theme song for "All That" is known for its catchy rhythm and verses.
Answer: TLC.

PARTIAL

PersonaRAG ✓

Reasoning process: The user's question is creatively asking for the artist who performed the theme song for the television show "All That". The user's playful and rhythmic style of questioning suggests an interest in rap or hip-hop music and culture. The theme song for "All That" is known for its catchy rhythm and verses, aligning with the user's apparent musical preferences.
Refined Answer: Yo, yo, you got it right, it's TLC on the mic! They're the ones who brought the heat, made that "All That" beat. With their funky rhythm and flow, they made that classic show glow. So next time you're leanin' back, remember it's TLC on that track!

Figure 1: Illustrations of Various RAG Models. Vanilla RAG and Chain-of-Thought [1] use passive learning, while PersonaRAG involves user-centric knowledge acquisition.

[13] addressed this issue by employing natural language inference models to select pertinent sentences, thereby enhancing the RAG's robustness. Additionally, advancements have been made in adaptively retrieving information, with systems like those proposed by Jiang et al. [14] dynamically fetching passages that are most likely to improve generation accuracy.

Despite these improvements, RAG systems still face limitations, particularly in adapting their output to the user's specific profile, such as their information needs or intellectual knowledge. This limitation stems from the current design of most RAG systems, which do not typically incorporate user context or personalized information retrieval strategies [15]. Consequently, there exists a gap between the general effectiveness of RAG systems and their applicability in personalized user experiences, where context and individual user preferences play a crucial role.

Personalization in information retrieval is increasingly recognized as essential for enhancing user interaction and satisfaction [16]. Techniques such as user profiling, context-aware retrieval, and adaptive feedback mechanisms are commonly employed to tailor search results to individual users' needs. For instance, Jeong et al. [17] proposed adaptive retrieval strategies that dynamically adjust the retrieval process based on the complexity of the query and the user's historical interaction data. These personalized approaches not only improve user satisfaction but also increase the efficiency of information retrieval by reducing the time users spend sifting through irrelevant information.

The integration of personalized techniques with agent-based systems provides a promising pathway to augment the capabilities of RAG systems. Agent-based systems, particularly in the form of LLM-Based Multi-Agent Frameworks [18], enable the simulation of complex interactions that can lead to more nuanced and contextually appropriate outputs. By incorporating multi-agent systems into RAG frameworks, there is potential for developing more robust and adaptive retrieval mechanisms that can handle a broader range of queries and generate more accurate responses, closely tailored to the specific needs and contexts of individual users.

In conclusion, while significant progress has been made in enhancing the effectiveness and personalization of RAG systems, ongoing research is crucial to address their existing limitations and expand their applications. The integration of personalized information retrieval and agent-based enhancements represents a promising avenue for further enhancing the adaptability and accuracy of RAG systems, potentially leading to intelligent information retrieval tailored to the specific needs of users.

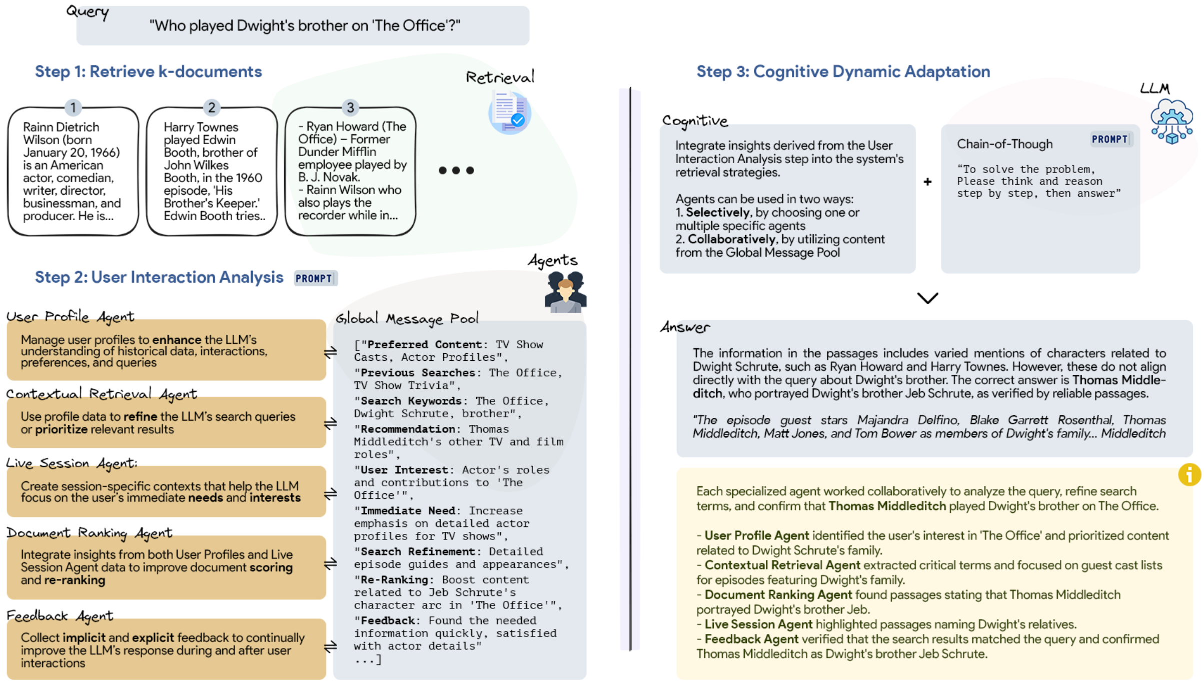
3. Methodology

In this section, we present the methodology underlying our PersonaRAG approach, which aims to enhance the ability of Language Large Models (LLMs) to actively engage with, understand, and leverage user profile information for personalized content generation. We begin by discussing the fundamental concepts of Retrieval-Augmented Generation (RAG) models (Section 3.1) and then introduce our PersonaRAG technique, which encourages LLMs to actively assimilate knowledge from live search sessions (Section 3.2).

3.1. Fundamentals of Retrieval-Augmented Generation (RAG) Models

State-of-the-art RAG models, as described in previous studies [19, 20, 21], employ retrieval systems to identify a set of passages $D = \{d_1, \ldots, d_n\}$ when given a query q. These passages are intended to enhance the generative capabilities of LLMs by providing them with contextually relevant information.

Early versions of RAG models typically employ a traditional retrieval-generation framework, in which the retrieved data set $D = \{d_1, \ldots, d_n\}$ is directly fed into LLMs to generate responses to the query q. However, these passages often contain irrelevant information, and the direct utilization approach in RAG has been shown to restrict the potential benefits of the RAG framework [22]. This limitation has sparked further discussion on how to improve LLMs by integrating retrieval results and outputs generated by the models themselves [23].



"Who played Dwight's brother on 'The Office'?"

Rainn Dietrich Wilson (born January 20, 1966) is an American actor, comedian, writer, director, businessman, and producer. He is...

Harry Townes played Edwin Booth, brother of John Wilkes Booth, in the 1960 episode, 'His Brother's Keeper.' Edwin Booth tries..

- Ryan Howard (The Office) – Former Dunder Mifflin employee played by B. J. Novak.
- Rainn Wilson who also plays the recorder while in...

Manage user profiles to **enhance** the LLM's understanding of historical data, interactions, preferences, and queries

Use profile data to **refine** the LLM's search queries or **prioritize** relevant results

Create session-specific contexts that help the LLM focus on the user's immediate **needs** and **interests**

Integrate insights from both User Profiles and Live Session Agent data to improve document **scoring** and **re-ranking**

Collect **implicit** and **explicit** feedback to continually improve the LLM's response during and after user interactions

```
["Preferred Content: TV Show
Casts, Actor Profiles",
"Previous Searches: The Office,
TV Show Trivia",
"Search Keywords: The Office,
Dwight Schrute, brother",
"Recommendation: Thomas
Middleditch's other TV and film
roles",
"User Interest: Actor's roles
and contributions to 'The
Office'",
"Immediate Need: Increase
emphasis on detailed actor
profiles for TV shows",
"Search Refinement: Detailed
episode guides and appearances",
"Re-Ranking: Boost content
related to Jeb Schrute's
character arc in 'The Office'",
"Feedback: Found the needed
information quickly, satisfied
with actor details"
...]
```

Integrate insights derived from the User Interaction Analysis step into the system's retrieval strategies.

Agents can be used in two ways:
1. **Selectively**, by choosing one or multiple specific agents
2. **Collaboratively**, by utilizing content from the Global Message Pool

```
"To solve the problem,
Please think and reason
step by step, then answer"
```

The information in the passages includes varied mentions of characters related to Dwight Schrute, such as Ryan Howard and Harry Townes. However, these do not align directly with the query about Dwight's brother. The correct answer is **Thomas Middleditch**, who portrayed Dwight's brother Jeb Schrute, as verified by reliable passages.

"The episode guest stars Majandra Delfino, Blake Garrett Rosenthal, Thomas Middleditch, Matt Jones, and Tom Bower as members of Dwight's family... Middleditch

Each specialized agent worked collaboratively to analyze the query, refine search terms, and confirm that **Thomas Middleditch** played Dwight's brother on The Office.

- **User Profile Agent** identified the user's interest in 'The Office' and prioritized content related to Dwight Schrute's family.
- **Contextual Retrieval Agent** extracted critical terms and focused on guest cast lists for episodes featuring Dwight's family.
- **Document Ranking Agent** found passages stating that Thomas Middleditch portrayed Dwight's brother Jeb.
- **Live Session Agent** highlighted passages naming Dwight's relatives.
- **Feedback Agent** verified that the search results matched the query and confirmed Thomas Middleditch as Dwight's brother Jeb Schrute.

Figure 2: Overview of Our PersonaRAG Model showcasing the dynamic interaction among specialized agents within the system, facilitated by a global message pool for structured communication. The diagram illustrates the flow from user query input through various agents, including User Profile, Context Retrieval, Session Analysis, Document Ranking, and Feedback Agents, highlighting their contributions to real-time adaptation and personalized content generation by integrating live user data and feedback for continuous improvement and contextually relevant search experiences.

3.2. PersonaRAG: RAG with User-Centric Agents

Drawing from the principles of adaptive learning and user-centered design, we develop a new PersonaRAG architecture to enable IR systems to dynamically learn from and adapt to user behavior in real-time. As shown in Figure 2, PersonaRAG introduces a three-step pipeline: retrieval, user interaction analysis, and cognitive dynamic adaptation. Unlike traditional IR models that statically respond to queries, PersonaRAG focuses on leveraging live user data to continually refine its understanding and responses without the need for manual retraining.

3.2.1. User Interaction Analysis

To understand user behavior from live interactions, PersonaRAG treats the IR system as a cognitive structure capable of receiving, interpreting, and acting upon user feedback [24]. Mimicking human learning behaviors, we establish four distinct agents within the system dedicated to analyzing user interactions from different perspectives: engagement tracking, preference analysis, context understanding, and feedback integration. These agents' roles are detailed in Section 3.2.2.

3.2.2. Cognitive Dynamic Adaptation

Following adaptive learning principles, we employ a dynamic adaptation mechanism to assist the IR system in utilizing real-time user data for continuous improvement. This mechanism facilitates the integration of insights gained from User Interaction Analysis into the system's retrieval processes. Specifically, we prompt the system to adjust its query responses based on an initial understanding of the user's needs and refine these responses as more user data becomes available. This approach not only personalizes the search results but also helps in correcting any misalignments or errors in real-time.

PersonaRAG employs a highly specialized agent architecture, with each agent focusing on a specific aspect of the information retrieval process. All agents utilize in-context learning, i.e., prompting, to perform their designated tasks. This role specialization allows for the efficient decomposition of complex user queries into manageable tasks [25]. To foster this, we engage the IR system as five specialized agents to analyze user interactions based on retrieved data. At present, the focus is on the functionality and interaction of these agents rather than their individual performance metrics.

User Profile Agent This component manages and updates user profile data, incorporating historical user interactions and preferences [26, 27]. It monitors how users interact with search results, such as click-through rates and navigation paths. The User Profile Agent helps the system understand what captures user interest and leads to deeper engagement, enabling personalized search experiences.

Contextual Retrieval Agent This agent is responsible for the initial retrieval of documents based on the user's current query. It accesses both a traditional search index and a more dynamic context-aware system that can consider broader aspects of the query environment. It utilizes user profile data to modify and refine search queries or to prioritize search results. For instance, if a user consistently engages more with certain types of documents or topics, the retrieval agent can boost those document types in the search results, ensuring that the most relevant information is presented to the user.

Live Session Agent This agent analyzes the current session in real-time, observing user actions such as clicks, time spent on documents, modifications to the query, and any feedback provided. It creates a session-specific context model that captures the user's immediate needs and interests. The real-time data collected by this agent is used to adjust the ongoing session, potentially re-ranking search results or suggesting new queries based on the user's behavior and preferences. Additionally, the Live Session Agent updates the user profile with new insights gleaned from the session, allowing for a more personalized and efficient search experience in future interactions.

Document Ranking Agent This agent is responsible for re-ranking the documents retrieved by the Contextual Retrieval Agent. It integrates insights from both the User Profile Agent and the Live Session Agent to score and order the documents more effectively. By considering the user's historical preferences and their current session behavior, the Document Ranking Agent ensures that the most relevant and valuable documents are presented to the user in a prioritized manner. This agent continuously adapts its ranking algorithms based on the feedback received from the user and the insights provided by the other agents in the system.

Feedback Agent This agent gathers implicit and explicit feedback during and after user interactions. Implicit feedback includes behavioral data like time spent on documents, click counts, and navigation patterns. Explicit feedback involves direct user input on document relevance and quality, collected through ratings, surveys, or comments. The agent uses this information to train and refine models for other agents, particularly the Document Ranking Agent. This process enhances the system's ability to anticipate user needs and deliver relevant documents based on accumulated feedback and insights.

By dynamically integrating insights from the User Profile Agent, Contextual Retrieval Agent, Live Session Agent, Document Ranking Agent, and Feedback Agent into the IR processes, PersonaRAG not only adapts to immediate user needs but also evolves over time to better anticipate and meet user expectations. This multi-agent approach enables PersonaRAG to embody a truly adaptive and user-focused information retrieval system, leveraging specialized agents to analyze user interactions from different behavioral perspectives and deliver highly personalized and contextually relevant search experiences. The inclusion of the Document Ranking Agent ensures that the most pertinent documents are identified and presented to users, further enhancing the system's ability to effectively satisfy user information needs.

3.3. PersonaRAG Operational Workflow

The PersonaRAG framework employs a structured workflow that allows for sequential and parallel processing of tasks, ensuring clarity and consistency in communication between agents through well-defined data structures and protocols [28]. The process involves the User Profile Agent, Contextual Retrieval Agent, Live Session Agent, Document Ranking Agent, and Feedback Agent working together to refine search queries, prioritize relevant results, and improve document scoring and re-ranking based on user profile, session-specific contexts, and feedback.

PersonaRAG's modular design allows for flexibility in the system setup, enabling researchers to focus on the most relevant aspects of the user's profile, session, and feedback data. Agents work collaboratively by utilizing content from the Global Message Pool, which serves as a central hub for inter-agent communication [28], eliminating inefficiencies and enabling agents to access or update information as required.

The Feedback Agent collects and analyzes implicit and explicit user feedback to generate insights into the effectiveness of retrieval strategies and document relevance. This feedback is used to make dynamic adjustments to the system, refining retrieval methods and altering the weighting of user profile factors. Through this iterative process, PersonaRAG continuously adapts and improves its performance, enhancing the accuracy and user satisfaction of the retrieval results [29].

4. Experimental Setups

In this section, we present the experimental setup employed in our study, including the datasets, baseline models, evaluation metrics, and implementation details. We also provide an overview of the prompts used in our experiments.

4.1. Datasets

Our experiments are conducted on three widely used single-hop benchmark datasets in the field of Information Retrieval (IR): NaturalQuestions (NQ) [30], TriviaQA [31], and WebQuestions (WebQ) [32]. NQ is a well-known dataset in Natural Language Understanding (NLU), consisting of structured questions and corresponding Wikipedia pages annotated with long and short answers. TriviaQA comprises question-answer pairs collected from trivia and quiz-league websites, while WebQ consists of questions selected using the Google Suggest API, with answers being entities in Freebase.

Table 1 summarizes the datasets used in our initial study. Due to the high cost of using language models and the large number of API calls required, we randomly sampled 500 questions from each raw dataset to create more manageable subsets for our experiments. While this sampling approach limits the scope of our study, it allows us to conduct an initial investigation into the performance of different RAG systems on these datasets. We acknowledge that future work with larger sample sizes and more comprehensive experiments will be necessary to draw definitive conclusions. Nonetheless, we believe this preliminary study provides valuable insights into the relative strengths and weaknesses of the tested RAG approaches.

4.2. Models

We compare PersonaRAG with several baseline models, including prompt learning and RAG models. The prompt templates used in user interaction analysis and dynamic adaptation are presented in Section 4.4. Initially, the question-answering (QA) instruction is fed to ChatGPT to conduct

Dataset	#Query	#Corpus	Sampling Rate
NQ	8,757	79,168	5.7%
TriviaQA	8,837	78,785	5.7%
WebQ	2,032	3,417	24.6%

Table 1
Summary of datasets. Each dataset consists of randomly sampled 500 questions from the raw dataset.

the vanilla answer generation model. Following the work of Wei et al. [33], the Chain-of-Thought model is implemented, which generates question rationale results to produce the final results. Additionally, the Guideline model serves as a baseline, generating problem-solving steps and guiding Language Models (LLMs) to generate the answer.

For the RAG-based baselines, two models are implemented: vanilla RAG and Chain-of-Thought, which include utilizing raw retrieved passages (CoT with Passage) and refining the passages as notes (CoT with Note). The vanilla RAG model directly feeds the top-ranked passages to the LLM. The Chain-of-Note model [1] is also implemented, which refines and summarizes the retrieved passages for generation. Inspired by Self-RAG Asai et al. [34], the Self-Rerank model is conducted, which filters out unrelated contents without fine-tuning LLMs.

4.3. Evaluation Metrics

When evaluating adaptive models, it is crucial to consider both task performance and user-centric adaptability simultaneously, along with their trade-offs. Therefore, the results are reported using different metrics, some of which measure effectiveness and others measure efficiency.

For effectiveness, accuracy is used, following the standard evaluation protocol in the field of Information Retrieval (IR) [35, 36, 34]. Accuracy assesses whether the predicted answer contains the ground-truth answer. Both the outputs of the Language Learning Model (LLM) and golden answers are converted to lowercase, and string matching (StringEM) is performed between each golden answer and the model prediction to calculate accuracy.

To evaluate user-centric adaptability, the BLEU-2 score is measured to assess the text similarity between different RAG and baseline setups and how well the generated answers resemble each other. This metric provides insights into the system's ability to generate consistent and coherent responses across various configurations. Additionally, the average sentence length and the average number of syllables of the answers from different RAG setups are reported as a post-hoc analysis. These measures validate whether the RAG system effectively adjusts its responses based on user knowledge levels, ensuring that the generated answers are tailored to the user's understanding and expertise.

Combining these evaluation strategies provides a comprehensive view of both the effectiveness and user-centric adaptability of the RAG system. The accuracy metric ensures that the system generates correct answers, while the BLEU-2 score and post-hoc analysis of sentence length and syllable count confirm the system's ability to adapt to user knowledge levels. As the understanding of user needs and system capabilities evolves, it is essential to continuously refine these metrics to maintain the RAG system's effectiveness in delivering personalized, context-aware responses that cater to the diverse requirements of users in the field of IR.

4.4. Implementation Details

For a fair comparison and following the work of Mallen et al. [35] and Trivedi et al. [37], the same retriever, a term-based sparse retrieval model known as BM25 [38], is used across all different models. The retrieval model is implemented using the OpenMatch toolkit [39]. For the external document corpus, the KILT-Wikipedia corpus preprocessed by Petroni et al. [40] is used, and the top-k relevant documents are retrieved.

Regarding the LLMs used to generate answers, the Llama 3 model instruct (ref) with 70b parameters, Mixture of Experts (MoE) 8x7b (ref), and the GPT-3.5 model (`gpt-3.5-turbo-0125`) are employed. For the retrieval-augmented LLM design, the implementation details from Trivedi et al. [37] are followed, which include input prompts, instructions, and the number of test samples for evaluation (e.g., 500 samples per dataset).

4.5. Prompts Used in PersonaRAG

This subsection presents the prompt templates employed in the construction of the PersonaRAG model. The prompts utilized in the User Interaction Analysis and Cognitive Dynamic Adaptation components are detailed below. The prompt templates used by the baseline models are available in the project repository [1]. In the templates, `{question}` represents the input question, `{global_memory}` the Global Message Pool, while `{passages}` denotes the retrieved passages. Additionally, `{cot_answer}` is populated with the output generated by the Chain-of-Thought model.

The placeholder `{user_profile_answer}` is filled with the response produced by the User Profile agent model. Respectively, `{contextual_answer}` corresponds to the Contextual Retrieval agent model, `{live_session_answer}` to the Live Session agent model, `{document_ranking_answer}` to the Document Ranking agent model, and `{feedback_answer}` to the Feedback agent model.

4.5.1. Prompts Used in User Interaction Analysis

User Profile Agent

```
Your task is to help the User Profile Agent
improve its understanding of user preferences
based on ranked document lists and the shared
global memory pool.

Question: {question}
Passages: {passages}
Global Memory: {global_memory}

Task Description:
From the provided passages and global memory
pool, analyze clues about the user's search
preferences. Look for themes, types of
documents, and navigation behaviors that reveal
user interest. Use these insights to recommend
how the User Profile Agent can refine and expand
the user profile to deliver better-personalized
results.
```

[1] https://github.com/padas-lab-de/ir-rag-sigir24-persona-rag

Contextual Retrieval Agent

You are a search technology expert guiding the Contextual Retrieval Agent to deliver context-aware document retrieval.

Question: {question}
Passages: {passages}
Global Memory: {global_memory}

Task Description:
Using the global memory pool and the retrieved passages, identify strategies to refine document retrieval. Highlight how user preferences, immediate needs, and global insights can be leveraged to adjust search queries and prioritize results that align with the user's interests. Ensure the Contextual Retrieval Agent uses this shared information to deliver more relevant and valuable results.

Live Session Agent

Your expertise in session analysis is required to assist the Live Session Agent in dynamically adjusting results.

Question: {question}
Passages: {passages}
Global Memory: {global_memory}

Task Description:
Examine the retrieved passages and information in the global memory pool. Determine how the Live Session Agent can use this data to refine its understanding of the user's immediate needs. Suggest ways to dynamically adjust search results or recommend new queries in real-time, ensuring that session adjustments align with user preferences and goals.

Document Ranking Agent

Your task is to help the Document Ranking Agent prioritize documents for better ranking.

Question: {question}
Passages: {passages}
Global Memory: {global_memory}

Task Description:
Analyze the retrieved passages and global memory pool to identify ways to rank documents effectively. Focus on combining historical user preferences, immediate needs, and session behavior to refine ranking algorithms. Your insights should ensure that documents presented by the Document Ranking Agent are prioritized to match user interests and search context.

Feedback Agent

You are an expert in feedback collection and analysis, guiding the Feedback Agent to gather and utilize user insights.

Question: {question}
Passages: {passages}
Global Memory: {global_memory}

Task Description:
Using the retrieved passages and global memory pool, identify methods for collecting implicit and explicit user feedback. Suggest ways to refine feedback mechanisms to align with user preferences, such as ratings, surveys, or behavioral data. Your recommendations should guide the Feedback Agent in updating other agents' models for more personalized and relevant results.

Global Message Pool

You are responsible for maintaining and enriching the Global Message Pool, serving as a central hub for inter-agent communication.

Question: {question}
Agent Responses: {agent_responses}
Existing Global Memory: {global_memory}

Task Description:
Using the responses from individual agents and the existing global memory, consolidate key insights into a shared repository.
Your goal is to organize a comprehensive message pool that includes agent-specific findings, historical user preferences, session-specific behaviors, search queries, and user feedback. This structure should provide all agents with meaningful data points and strategic recommendations, reducing redundant communication and improving the system's overall efficiency.

4.5.2. Prompts Used in Cognitive Dynamic Adaptation

Chain-of-Thought

To solve the problem, Please think and reason step by step, then answer.

Question: {question}
Passages: {passages}
Reasoning process:
1. Read the given question and passages to gather relevant information.
2. Write reading notes summarizing the key points from these passages.
3. Discuss the relevance of the given question and passages.
4. If some passages are relevant to the given question, provide a brief answer based on the passages.
5. If no passage is relevant, directly provide the answer without considering the passages.

Answer:

Cognitive Agent

Your task is to help the Cognitive Agent enhance its understanding of user insights to continuously improve the system's responses.

Question: {question}
Initial Response: {cot_answer}

Method	Setting	Top-3			Top-5		
		WebQ	TriviaQA	NQ	WebQ	TriviaQA	NQ
w/o RAG	`gpt-3.5-turbo-0125`	59.61	97.36	43.90	62.43	97.36	41.46
	Guideline	36.53	42.10	17.07	47.21	36.84	21.95
vanillaRAG		38.46	78.94	36.58	50.14	81.57	41.46
Self-Refined	Chain-of-Thought (CoT)	57.69	89.47	39.02	67.51	89.47	41.46
	Chain-of-Note (CoN)	57.17	81.57	48.78	65.15	92.10	48.78
	Self-Rerank (SR)	32.63	81.57	43.90	40.26	84.21	51.21
PersonaRAG		**63.46**	**94.73**	**49.02**	**67.50**	**89.47**	**48.78**

Table 2
Overall Accuracy Performance Comparison Using Top-3 and Top-5 Passages. PersonaRAG results are reported in **bold**.

```
User Insights from Interaction Analysis:
User Profile Agent: {user_profile_answer},
Contextual Retrieval Agent: {contextual_answer},
Live Session Agent: {live_session_answer},
Document Ranking Agent:
{document_ranking_answer},
Feedback Agent: {feedback_answer}

Task Description:
Verify the reasoning process in the initial
response for errors or misalignments. Use
insights from user interaction analysis
to refine this response, correcting any
inaccuracies and enhancing the query answers
based on user profile. Ensure that your refined
response aligns more closely with the user's
immediate needs and incorporates foundational or
advanced knowledge from other sources.

Answer:
```

5. Experimental Results and Analyses

In this section, we show the overall experimental results and offer in-depth analyses of our method.

5.1. Main Results

Table 2 summarizes the primary findings for PersonaRAG across various single-hop question answering datasets. The approach was evaluated against multiple baseline models, including large language models (LLMs) without retrieval-augmented generation (RAG), the conventional RAG model, and self-refined variants, such as utilizing raw retrieved passages (CoT with Passage) or refining passages into notes (CoT with Note).

PersonaRAG demonstrated superior performance compared to most of the baseline models, achieving significant improvements over the conventional RAG (i.e., vanillaRAG) of over 10%, particularly on the WebQ dataset. It also consistently outperformed the ChatGPT-3.5 model, except on TriviaQA, which we suspect is part of the model's training dataset. These results suggest PersonaRAG's capability to guide LLMs in extracting relevant information through active learning techniques.

Specifically, the performance of RAG models was assessed using the top 3 and 5 ranked passages. While other RAG models generally benefited from more passages, PersonaRAG maintained consistent performance with either 3 or 5 passages, suggesting that 3 passages were adequate for generating accurate answers. PersonaRAG agents played a crucial role in efficiently extracting the necessary information regarding the user's information need to achieve these improvements.

Furthermore, on the WebQ dataset, PersonaRAG achieved accuracy scores of 63.46% and 67.50% using Top-3 and Top-5 passages, respectively, surpassing the vanillaRAG model by 25% and 17.36%, and nearly all other baseline models (except for Chain-of-Thought using Top-5, which performed equally). On the NQ dataset, PersonaRAG maintained similarly robust performance with scores of 49.02% and 48.78%, outperforming all baselines (except for Chain-of-Thought and Self-Rerank (SR) using Top-5). This pattern was further validated by experiments on other datasets, with results showing that PersonaRAG consistently outperforms conventional RAG models with the capability of providing an answer tailored to the user's interaction and information need. The comprehensive understanding it provides contributes to the generation of accurate and user-centric answers across various question complexities.

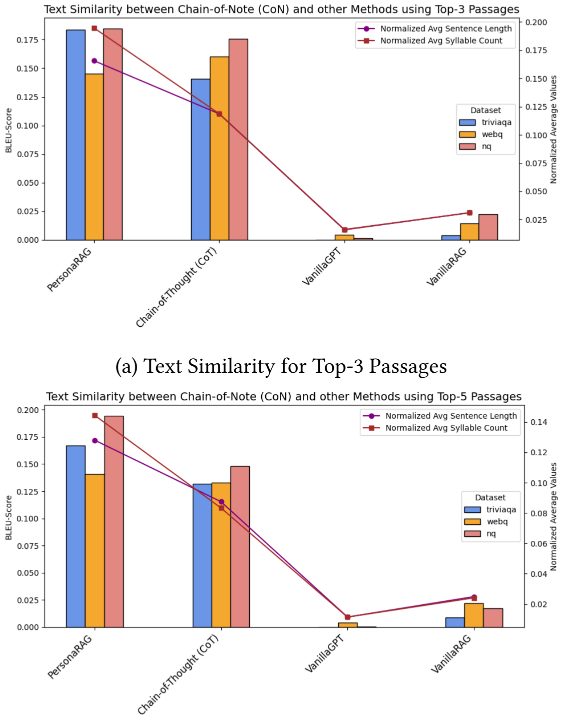
5.2. Comparative Analysis of RAG Configurations

Further experiments explored PersonaRAG's adaptive capabilities (Figure 3). BLEU-2 scores compared outputs from Chain-of-Note (consistently best outside PersonaRAG) with other methods. PersonaRAG showed higher similarity scores, indicating its ability to generate responses that address user needs rather than just summarizing input. Additionally, PersonaRAG provides personalized answers tailored to user profiles, extending beyond mere information provision.

The Chain-of-Note approach demonstrated comparable performance to the Chain-of-Thought approach, implying that both techniques effectively extract pertinent information from the retrieved passages and adapt it to align with the user's information need.

In contrast, vanillaGPT and vanillaRAG outputs differed significantly from the Chain-of-Note approach, indicating that counterfactual cognition often leads to diverse outcomes rather than focusing solely on query-relevant content. This suggests LLMs can construct knowledge from multiple perspectives and customize responses based on user understanding.

Post-hoc analyses of average sentence length and syllable count across RAG configurations provided insights into the system's ability to adapt responses to user comprehension levels. These observations highlight PersonaRAG's capacity to synthesize knowledge from various perspectives and tailor responses to different levels of user expertise.



(a) Text Similarity for Top-3 Passages

(b) Text Similarity for Top-5 Passages

Figure 3: Text Similarity between Chain-of-Note (CoN) and Other Methods Using BLEU-2 Score for Evaluation, with Normalized Average Sentence Length and Average Syllable Count.

Method	**WebQ**	**TriviaQA**	**NQ**
LLaMA3-70B			
w/o RAG	45.25	82.17	38.95
vanillaRAG	55.14	85.02	40.37
Chain-of-Thought	60.52	88.72	45.10
Chain-of-Note	62.67	89.37	48.25
Self-Rerank	54.25	84.50	47.77
PersonaRAG	**66.09**	**92.12**	**50.85**
MoE-8x7b			
w/o RAG	38.24	75.82	34.26
vanillaRAG	48.44	80.25	38.50
Chain-of-Thought	54.12	85.46	42.37
Chain-of-Note	55.98	87.55	45.14
Self-Rerank	52.50	83.04	44.96
PersonaRAG	**61.35**	**90.30**	**49.27**

Table 3
Overall Accuracy Performance Comparison for Top-5 Passages using LLaMA3-70B and MoE-8x7b.

5.3. Analysis on Generalization Ability

This experiment evaluates the quality of knowledge construction using different large language models (LLMs). As illustrated in Table 3, the PersonaRAG outcomes are used to prompt open-source LLMs, specifically LLaMA3-70B and MoE-8x7b, to generate accurate answers.

Compared to LLMs without retrieval-augmented generation (w/o RAG), vanilla RAG and Chain-of-Note often exhibit lower performance. This result suggests that retrieved passages can act as noise, adversely affecting model performance even after refinement through note generation. One primary reason for this behavior is that both LLaMA3-70B and MoE-8x7b struggle to efficiently analyze and identify relevant knowledge due to limitations in their processing capacities.

In contrast, the PersonaRAG method provides notable performance improvements: over 8% for LLaMA3-70B and more than 10% for MoE-8x7b across all datasets, underscoring its effectiveness. The PersonaRAG methodology distinguishes itself from the Chain-of-Note approach by offering a cognitive framework that connects retrieved passages with prior knowledge. This framework models the instructor's (GPT-3.5) reasoning process, guiding the student models (LLaMA3-70B and MoE-8x7b) to better understand knowledge retrieved from passages. The results demonstrate that the LLMs are capable of selecting appropriate passages to build more accurate responses, highlighting the benefits of the PersonaRAG approach for improving generalization.

Question: Who stole the Mona Lisa from the Louvre in 1911?
Answer: "Vincenzo Peruggia"

Chain-of-Thought:
1. Read the given question and passages to gather relevant information.
2. Write reading notes summarizing the key points.
3. Discuss the relevance of the given question and passages.
4. If some passages are relevant to the given question, provide a brief answer based on the passages.
5. If no passage is relevant, directly provide the answer without considering the passages.

Step 1: Passage Retrieval

1. The Mona Lisa was stolen from the Louvre Museum while the museum was closed for cleaning. Witnesses reported that a tall, stout individual was carrying what appeared to be a large panel covered with a horse blanket. Two years later, Vincenzo Peruggia was arrested in Florence.
2. On August 22, 1911, the Mona Lisa was stolen from the Louvre. Vincenzo Peruggia claimed he stole the painting to return it to Italy.
3. In 1911, Vincenzo Peruggia, a Louvre employee, stole the Mona Lisa.

Step 2: User Interaction Analysis

User Insights from Interaction Analysis:
- User Profile Agent: The user is interested in art heists and historical mysteries.
- Contextual Retrieval Agent: The user frequently searches for art history topics.
- Live Session Agent: The user reads about unsolved mysteries and historic crimes.
- Document Ranking Agent: The user favors detailed, chronological accounts.
- Feedback Agent: The user interacts with articles about famous art heists.

Step 3: Cognitive Dynamic Adaptation

The user has a consistent interest in famous art heists and historical mysteries. They are drawn to unsolved mysteries and historic crimes. They favor detailed, chronological accounts, so the final response should highlight theft motives, the method, and the investigation leading to the arrest, providing comprehensive and authoritative sources to satisfy their engagement with articles on art heists.
Answer: Vincenzo Peruggia, a Louvre employee, stole the Mona Lisa from the Louvre Museum on August 21, 1911. He claimed he stole the painting to return it to Italy, and the theft remained unsolved for two years before he was arrested in Florence in December 1913. The painting was later recovered and returned to the Louvre.

Table 4
PersonaRAG Case Study.

5.4. Case Study

Finally, we randomly sample one case in Table to demonstrate the effectiveness of PersonaRAG.

The user interaction analysis mechanism effectively generates comprehensive results by integrating foundational and advanced insights from user data. Retrieved passages provide critical clues for answering questions, while agent analyses summarize and illustrate the applicability of external information to user queries. The cognitive dynamic adaptation module refines initial chain-of-thought responses using these insights, generating accurate answers. For example, including knowledge about the ”theft of the Mona Lisa in 1911,” ”Vincenzo Peruggia,” and ”Florence” enhances the reasoning process’s precision and detail. This demonstrates PersonaRAG’s effectiveness in helping IR agents combine external knowledge with intrinsic user data to produce well-informed responses.

6. Conclusion

This paper proposes PersonaRAG, which constructs the retrieval-augmentation architecture incorporating user interaction analysis and cognitive dynamic adaptation. PersonaRAG builds the user interaction agents and dynamic cognitive mechanisms to facilitate the understanding of user needs and interests and enhance the system capabilities to deliver personalized, context-aware responses with the intrinsic cognition of LLMs.

Furthermore, PersonaRAG demonstrates effectiveness in leveraging external knowledge and adapting responses based on user profiles, knowledge levels, and information needs to support LLMs in generation tasks without finetuning. However, this approach requires multiple calls to the LLM’s API, which can introduce additional time latency and increase API calling costs when addressing questions. The process involves constructing the initial Chain-of-Thought, processing the User Interaction Agents results, and executing the Cognitive Dynamic Adaptation to generate the final answer. Furthermore, the inputs to LLMs in this approach tend to be lengthy due to the inclusion of extensive retrieved passages and the incorporation of user needs, interests, and profile construction results. These factors can impact the efficiency and cost-effectiveness of the PersonaRAG approach in practical applications of Information Retrieval (IR) systems.

Future research will aim to optimize the process by reducing API calls and developing concise representations of user profiles and retrieved information without compromising response quality. We also plan to explore more user-centric agents to better capture writing styles and characteristics of RAG users/searchers. This will enhance the system’s ability to understand and adapt to individual preferences, improving personalization and relevance in IR tasks.

Acknowledgments

This work has received funding from the European Union’s Horizon Europe research and innovation program under grant agreement No 101070014 (OpenWebSearch.EU, https://doi.org/10.3030/101070014).

References

[1] W. Yu, H. Zhang, X. Pan, K. Ma, H. Wang, D. Yu, Chain-of-note: Enhancing robustness in retrieval-augmented language models, CoRR abs/2311.09210 (2023). URL: https://doi.org/10.48550/arXiv.2311.09210. doi:10.48550/ARXIV.2311.09210.

[2] OpenAI, GPT-4 technical report, CoRR abs/2303.08774 (2023). URL: https://doi.org/10.48550/arXiv.2303.08774. doi:10.48550/ARXIV.2303.08774. arXiv:2303.08774.

[3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, CoRR abs/2302.13971 (2023). URL: https://doi.org/10.48550/arXiv.2302.13971. doi:10.48550/ARXIV.2302.13971. arXiv:2302.13971.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[5] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, P. Fung, A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, in: J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, A. A. Krisnadhi (Eds.), Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023, Association for Computational Linguistics, 2023, pp. 675–718. URL: https://doi.org/10.18653/v1/2023.ijcnlp-main.45. doi:10.18653/V1/2023.IJCNLP-MAIN.45.

[6] P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

[7] J. Chen, H. Lin, X. Han, L. Sun, Benchmarking large language models in retrieval-augmented generation, in: M. J. Wooldridge, J. G. Dy, S. Natarajan (Eds.), Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February

20-27, 2024, Vancouver, Canada, AAAI Press, 2024, pp. 17754–17762. URL: https://doi.org/10.1609/aaai.v38i16.29728. doi:10.1609/AAAI.V38I16.29728.

[8] J. Teevan, S. T. Dumais, E. Horvitz, Personalizing search via automated analysis of interests and activities, SIGIR Forum 51 (2017) 10–17. URL: https://doi.org/10.1145/3190580.3190582. doi:10.1145/3190580.3190582.

[9] K. Sugiyama, K. Hatano, M. Yoshikawa, Adaptive web search based on user profile constructed without any effort from users, in: S. I. Feldman, M. Uretsky, M. Najork, C. E. Wills (Eds.), Proceedings of the 13th international conference on World Wide Web, WWW 2004, New York, NY, USA, May 17-20, 2004, ACM, 2004, pp. 675–684. URL: https://doi.org/10.1145/988672.988764. doi:10.1145/988672.988764.

[10] G. Adomavicius, B. Mobasher, F. Ricci, A. Tuzhilin, Context-aware recommender systems, AI Mag. 32 (2011) 67–80. URL: https://doi.org/10.1609/aimag.v32i3.2364. doi:10.1609/AIMAG.V32I3.2364.

[11] M. J. Wooldridge, An Introduction to MultiAgent Systems, Second Edition, Wiley, 2009.

[12] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de Las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mixtral of experts, CoRR abs/2401.04088 (2024). URL: https://doi.org/10.48550/arXiv.2401.04088. doi:10.48550/ARXIV.2401.04088.

[13] F. Xu, W. Shi, E. Choi, RECOMP: improving retrieval-augmented lms with compression and selective augmentation, CoRR abs/2310.04408 (2023). URL: https://doi.org/10.48550/arXiv.2310.04408. doi:10.48550/ARXIV.2310.04408.

[14] Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, G. Neubig, Active retrieval augmented generation, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, Association for Computational Linguistics, 2023, pp. 7969–7992. URL: https://doi.org/10.18653/v1/2023.emnlp-main.495. doi:10.18653/V1/2023.EMNLP-MAIN.495.

[15] H. Zamani, W. B. Croft, Embedding-based query language models, in: B. Carterette, H. Fang, M. Lalmas, J. Nie (Eds.), Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval, ICTIR 2016, Newark, DE, USA, September 12- 6, 2016, ACM, 2016, pp. 147–156. URL: https://doi.org/10.1145/2970398.2970405. doi:10.1145/2970398.2970405.

[16] M. R. Ghorab, D. Zhou, A. O'Connor, V. Wade, Personalised information retrieval: survey and classification, User Model. User Adapt. Interact. 23 (2013) 381–443. URL: https://doi.org/10.1007/s11257-012-9124-1. doi:10.1007/S11257-012-9124-1.

[17] S. Jeong, J. Baek, S. Cho, S. J. Hwang, J. C. Park, Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity, CoRR abs/2403.14403 (2024). URL: https://doi.org/10.48550/arXiv.2403.14403. doi:10.48550/ARXIV.2403.14403.

[18] Y. Li, Y. Zhang, L. Sun, Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents, CoRR abs/2310.06500 (2023). URL: https://doi.org/10.48550/arXiv.2310.06500. doi:10.48550/ARXIV.2310.06500.

[19] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, Q. Guo, M. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, CoRR abs/2312.10997 (2023). URL: https://doi.org/10.48550/arXiv.2312.10997. doi:10.48550/ARXIV.2312.10997.

[20] Y. Huang, J. Huang, A survey on retrieval-augmented text generation for large language models, arXiv preprint arXiv:2404.10981 (2024).

[21] S. Siriwardhana, R. Weerasekera, T. Kaluarachchi, E. Wen, R. Rana, S. Nanayakkara, Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering, Trans. Assoc. Comput. Linguistics 11 (2023) 1–17. URL: https://transacl.org/ojs/index.php/tacl/article/view/4029.

[22] J. Chen, H. Lin, X. Han, L. Sun, Benchmarking large language models in retrieval-augmented generation, in: M. J. Wooldridge, J. G. Dy, S. Natarajan (Eds.), Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, AAAI Press, 2024, pp. 17754–17762. URL: https://doi.org/10.1609/aaai.v38i16.29728. doi:10.1609/AAAI.V38I16.29728.

[23] K. Wu, E. Wu, J. Zou, How faithful are rag models? quantifying the tug-of-war between rag and llms' internal prior, arXiv preprint arXiv:2404.10198 (2024).

[24] R. C. Atkinson, R. M. Shiffrin, Human memory: A proposed system and its control processes, in: K. W. Spence, J. T. Spence (Eds.), Psychology of Learning and Motivation, volume 2 of *Psychology of Learning and Motivation*, Elsevier, 1968, pp. 89–195. URL: https://doi.org/10.1016/s0079-7421(08)60422-3. doi:10.1016/S0079-7421(08)60422-3.

[25] A. Sharma, S. Kumar, Semantic web-based information retrieval models: a systematic survey, in: Data Science and Analytics: 5th International Conference on Recent Developments in Science, Engineering and Technology, REDSET 2019, Gurugram, India, November 15–16, 2019, Revised Selected Papers, Part II 5, Springer, 2020, pp. 204–222.

[26] A. Kacem, Personalized Information Retrieval based on Time-Sensitive User Profile. (Recherche d'Information Personalisée basée sur un Profil Utilisateur Sensible au Temps), Ph.D. thesis, Paul Sabatier University, Toulouse, France, 2017. URL: https://tel.archives-ouvertes.fr/tel-01707423.

[27] A. Singh, A. Sharma, A multi-agent framework for context-aware dynamic user profiling for web personalization, in: Software Engineering: Proceedings of CSI 2015, Springer, 2019, pp. 1–16.

[28] S. Hong, X. Zheng, J. Chen, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, Metagpt: Meta programming for multi-agent collaborative framework, CoRR abs/2308.00352 (2023). URL: https://doi.org/10.48550/arXiv.2308.00352.

doi:10.48550/ARXIV.2308.00352.

[29] D. K. Limbu, A. M. Connor, R. Pears, S. G. MacDonell, Contextual relevance feedback in web information retrieval, in: I. Ruthven (Ed.), Proceedings of the 1st International Conference on Information Interaction in Context, IIiX 2006, Copenhagen, Denmark, October 18-20, 2006, ACM, 2006, pp. 138–143. URL: https://doi.org/10.1145/1164820.1164848. doi:10.1145/1164820.1164848.

[30] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. P. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M. Chang, A. M. Dai, J. Uszkoreit, Q. Le, S. Petrov, Natural questions: a benchmark for question answering research, Trans. Assoc. Comput. Linguistics 7 (2019) 452–466. URL: https://doi.org/10.1162/tacl_a_00276. doi:10.1162/TACL_A_00276.

[31] M. Joshi, E. Choi, D. S. Weld, L. Zettlemoyer, Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension, in: R. Barzilay, M. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, Association for Computational Linguistics, 2017, pp. 1601–1611. URL: https://doi.org/10.18653/v1/P17-1147. doi:10.18653/V1/P17-1147.

[32] J. Berant, A. Chou, R. Frostig, P. Liang, Semantic parsing on freebase from question-answer pairs, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2013, pp. 1533–1544. URL: https://aclanthology.org/D13-1160/.

[33] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

[34] A. Asai, Z. Wu, Y. Wang, A. Sil, H. Hajishirzi, Self-rag: Learning to retrieve, generate, and critique through self-reflection, CoRR abs/2310.11511 (2023). URL: https://doi.org/10.48550/arXiv.2310.11511. doi:10.48550/ARXIV.2310.11511.

[35] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, H. Hajishirzi, When not to trust language models: Investigating effectiveness of parametric and non-parametric memories, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, Association for Computational Linguistics, 2023, pp. 9802–9822. URL: https://doi.org/10.18653/v1/2023.acl-long.546. doi:10.18653/V1/2023.ACL-LONG.546.

[36] J. Baek, S. Jeong, M. Kang, J. C. Park, S. J. Hwang, Knowledge-augmented language model verification, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, Association for Computational Linguistics, 2023, pp. 1720–1736. URL: https://doi.org/10.18653/v1/2023.emnlp-main.107. doi:10.18653/V1/2023.EMNLP-MAIN.107.

[37] H. Trivedi, N. Balasubramanian, T. Khot, A. Sabharwal, Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, Association for Computational Linguistics, 2023, pp. 10014–10037. URL: https://doi.org/10.18653/v1/2023.acl-long.557. doi:10.18653/V1/2023.ACL-LONG.557.

[38] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3, in: D. K. Harman (Ed.), Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994, volume 500-225 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 1994, pp. 109–126. URL: http://trec.nist.gov/pubs/trec3/papers/city.ps.gz.

[39] S. Yu, Z. Liu, C. Xiong, Z. Liu, Openmatch-v2: An all-in-one multi-modality plm-based information retrieval toolkit, in: H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe, B. Poblete (Eds.), Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, ACM, 2023, pp. 3160–3164. URL: https://doi.org/10.1145/3539618.3591813. doi:10.1145/3539618.3591813.

[40] F. Petroni, A. Piktus, A. Fan, P. S. H. Lewis, M. Yazdani, N. D. Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, V. Plachouras, T. Rocktäschel, S. Riedel, KILT: a benchmark for knowledge intensive language tasks, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, Association for Computational Linguistics, 2021, pp. 2523–2544. URL: https://doi.org/10.18653/v1/2021.naacl-main.200. doi:10.18653/V1/2021.NAACL-MAIN.200.

s