Leveraging LLM-Powered Multi-Agent Systems to Enhance Customer Experience in Complex Product Domains

Marco Valentini[1,*], Antonio Ferrara[1], Tommaso Di Noia[1], Giuseppe Illuzzi[2] and Pierangelo Colacicco[2]

[1]Politecnico di Bari, Bari, Italy
[2]Natuzzi Italia Innovation Center, Santeramo in Colle, Italy

Abstract

Providing an enhanced customer experience in complex retail environments poses significant daily challenges, like guiding users through wide catalogs and offering personalized support for complex purchasing decisions. This paper introduces SofAgent, a Large Language Model-powered Multi-Agent System designed to provide intelligent customer assistance through the cooperation of a group of agents orchestrated by a manager. Key contributions include a Multi-Agent System design tailored for nuanced product inquiries and style advice; effective task decomposition to handle diverse customer needs, e.g., information retrieval and recommendation; and integration of Large Language Model reasoning with factual data to ensure accurate product information and personalized recommendations. The resulting system aims to improve customer engagement and operational efficiency in complex retail environments.

Keywords

Multi-Agent Systems, Large Language Models, Recommender Systems, Information Retrieval

1. Introduction

In modern retail, especially for businesses with large and customizable product catalogs, providing superior customer assistance is paramount. Customers experience information overload [11, 2], and struggle to navigate extensive options, articulate nuanced requirements, and envision how different items might complement each other in terms of style or function [1]. These decision-making obstacles can lead to frustration and lost sales. Although skilled human consultants can bridge this gap, their availability is limited and scaling such expertise across digital platforms remains a significant challenge.

Recent advancements in Large Language Models (LLMs) have shown promising results in powering intelligent conversational agents capable of understanding natural language and engaging in complex dialogues [4, 3]. However, when faced with intricate retail scenarios, such as high-end furniture shopping, which demands nuanced product knowledge, aesthetic judgement, and personalized configuration, individual LLMs may fall short. Effectively supporting such scenarios requires maintaining well-grounded factual accuracy, stylistically coherent recommendations aligned with functional requirements, and the ability to manage multi-turn interactions that integrate information retrieval and personalized suggestions [13] for complementary items. These multifaceted requirements usually exceed the capabilities of a single, monolithic model [7].

Multi-Agent Systems (MASs), made up of specialized LLM-powered agents working together, offer a powerful way to solve complex problems by breaking them into smaller tasks. They can translate diverse user needs into manageable requests, enabling more effective and personalized assistance. To explore the potential of MASs in addressing complex customer assistance scenarios, where product recommendations must jointly account for item search based on specific characteristics, user-defined constraints, complementarity across items, and compositional configurations, this paper introduces

IIR2025: 15th Italian Information Retrieval Workshop, 3th - 5th September 2025, Cagliari, Italy
*Corresponding author.
m.valentini7@phd.poliba.it (M. Valentini); antonio.ferrara@poliba.it (A. Ferrara); tommaso.dinoia@poliba.it (T. D. Noia); gilluzzi@natuzzi.com (G. Illuzzi); pcolacicco@natuzzi.com (P. Colacicco)
0009-0004-1652-0745 (M. Valentini); 0000-0002-1921-8304 (A. Ferrara); 0000-0002-0939-5462 (T. D. Noia)

SofAgent, a novel LLM-powered MAS designed to support the customer purchasing journey within rich and multifaceted product ecosystems. Despite their practical importance, such scenarios remain underexplored in the literature [17, 6, 20], particularly in domains like furniture retail where choices often span interdependent products, personalized configurations, and aesthetic coherence. SofAgent leverages a set of cooperative agents, orchestrated by a manager, to interpret nuanced user requests, retrieve accurate product information, and provide coherent and personalized style advice.

To explore the potential of MASs in meeting these challenges in customer assistance scenarios where recommendations with all together search of items from characteristics, constraints, complementarity, composition of products, which is a complex scenario underexplored in literature, this paper introduces SofAgent, a novel LLM-powered MAS designed to support the customer purchasing journey within complex product ecosystems. SofAgent leverages cooperative specialized agents, orchestrated by a manager, to effectively handle nuanced product inquiries and provide style advice. Key contributions of this work include: (1) a specialized MAS design tailored for complex product ecosystems like furniture retail, focusing on both detailed product information and stylistic guidance; (2) effective task decomposition strategies to handle diverse customer needs, including information retrieval, and personalized recommendations of complementary items; and (3) a robust integration of LLM reasoning capabilities with factual product data to ensure accuracy and mitigate the risk of hallucinations. While the system architecture is designed to be generalizable, this investigation focuses on the customer assistance scenario of Natuzzi, a premium Italian furniture brand, where users typically navigate extensive catalogs of sofas, furnishings, dimensions, materials, and configurations. In this context, SofAgent aims to understand customer needs expressed in natural language and efficiently guide them through product discovery and decision-making, thereby improving customer engagement and operational efficiency in sophisticated retail environments.

2. Background

The landscape of AI-driven conversational systems is being reshaped by Intelligent Agents based on LLMs and their organization into collaborative MASs. The power of LLM-powered agents stem from their advanced natural language processing and reasoning capabilities [16]. A key to their adaptability is In-Context Learning (ICL) [5], which enables the specialization of agents for distinct roles through simple instructions, such as in role-playing paradigms [10], avoiding costly retraining. In this paradigm, an agent's area of expertise is determined by its initial prompt, which defines the agent's role, relevant domain knowledge, and specific set of permitted actions, including tool-use protocols. Maintaining conversational state and enabling personalization are critical capabilities for LLM Agents. These are achieved through memory mechanisms that differentiate between the LLM's intrinsic short-term context window and externalized knowledge bases that help long-term understanding [19]. A crucial capability for the agents is the ability to use external tools via APIs [12]. This process is often governed by a structured framework like the Reason and Act (ReAct), which enables the agent to reason before acting [18]. Tool-use is essential for grounding responses in factual, real-time information from knowledge bases, ensuring the accuracy of product details and mitigating the hallucination issue [9].

However, while individual agents are powerful, tackling multifaceted problems, such as guiding a customer through a complex purchasing journey that requires information retrieval, search, and style advice, necessitates the orchestration of multiple, specialized capabilities [15]. This has led to the rise of MASs, where multiple specialized agents communicate and cooperate to achieve a common goal [6, 17]. Within a MAS, the principles of role-playing and ICL are leveraged to assign agents distinct competencies. This architecture allows for effective problem decomposition, where intricate user requests are broken into manageable sub-tasks handled by the appropriate expert agent, often orchestrated by a central manager [14]. This decentralization not only enhances robustness but is inherently suited for the diverse demands of advanced customer service. Despite these advancements, deploying LLM-based MASs effectively presents ongoing challenges, including ensuring consistent agent coordination, maintaining long-term context for personalization, and guaranteeing factual accuracy [8].

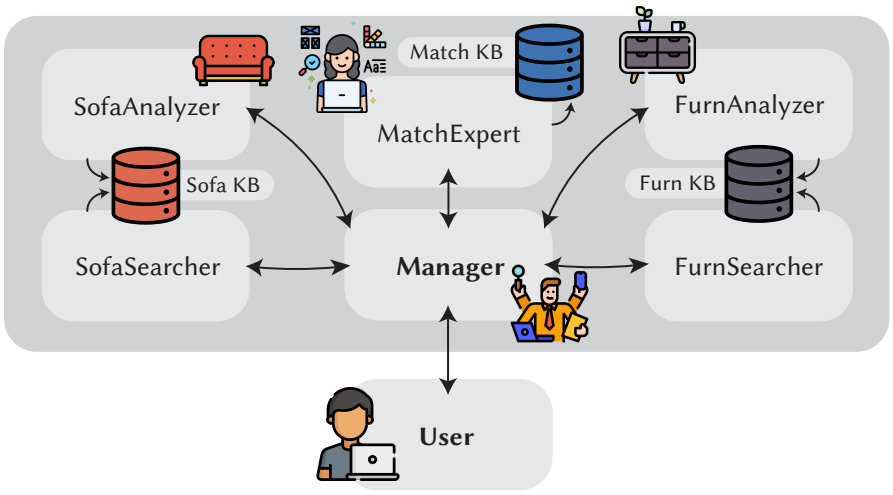


Figure 1: SofAgent Multi-Agent Architecture

3. Methodology

Aiming to provide expert customer assistance in complex retail environments, we design SofAgent, a hierarchical MAS with specialized agents orchestrated by a Manager, which acts as the primary user interface as depicted in Figure 1. Each agent's specific role, its allowed actions, and available tools are defined at initialization via meticulously crafted prompts, available in the Github repository, applying the ICL and role-playing paradigms. The workload is divided among the following key agents:

- **Manager Agent**: This is the central orchestrator, responsible for interpreting user intent, task decomposition, and response synthesis. It employs a two-stage reasoning process, i.e., thought and action to analyze the current situation and select the most suitable specialized agent for each sub-task. Finally, it synthesizes the findings from all agents into a single, coherent response for the customer. The agent also implements error handling and fallback strategies, such as searching for matches with similar input items if a direct match is not found.
- **Analyzer Agents**: These agents retrieve detailed factual information about specific products. **SofaAnalyzer** handles sofas (IDs, features, prices, configurations, seats), while **FurnAnalyzer** addresses other furniture, and is capable of iterative processing for multi-category requests.
- **Searcher Agents**: The **SofaSearcher** and **FurnSearcher** agents operate as intelligent search engines. They interpret natural language directives from the Manager and transform them into structured queries compatible with the product catalogs. They enable filtering by various attributes such as product features and price, with the SofaSearcher agent possessing the additional capability to query by seating capacity. These agents enhance system's capability to navigate large catalogs based on user-defined criteria.
- **MatchExpert Agent**: This agent acts as a virtual design consultant that recommends harmonious pairings for a given item based on precomputed recommendations using different criteria: generic stylistic collections, color themes, or spatial layouts. It plays a crucial role in providing valuable style advice and increasing cross-selling.

To ensure factual accuracy and mitigate the risk of hallucination, we ground the system's reasoning in a structured, external knowledge base that the specialist agents can access through a dedicated suite of tools. The knowledge base contains all the information about Natuzzi catalog: a detailed sofa catalog (**Sofa KB**), which is essential for understanding the characteristics of individual models, including design, materials, and functional features, a list of predefined sofa configurations, which provides concrete examples of assembled sofas, including their seating capacities and retail prices. Additionally, it contains a collection of other home furnishings (**Furn KB**), which are categorized and detailed with

key features, pricing, and material information. The knowledge base also contains contextual and matching data (**Match KB**), including information on stylistic compatibilities between items, color theme alignments, and predefined room layouts, allowing the systems to generate aesthetically informed, and personalized recommendations for customers.

4. Experiments

This section outlines the experimental methodology and key results of the evaluation of SofAgent, benchmarked against a single-agent baseline. The evaluation aims to assess (1) whether a multi-agent architecture improves the accuracy of information retrieval in response to user queries, and (2) whether integrating factual information from a structured knowledge base reduces the likelihood of hallucinations commonly exhibited by LLMs.

The implementation of the system is based on LangChain[1], and the code can be accessed on Github[2]. Both the SofAgent architecture and the baseline were powered by GPT-4o models, and have access to the same knowledge base data. Furthermore, both the systems have been initialized with a carefully crafted prompt instructing them to act as expert Natuzzi shopping assistants and precisely describing the scope and structure of the available knowledge base. The test set for the systems was designed to simulate realistic customer interactions and challenge the systems' ability to interpret, reason over, and retrieve information from a complex product catalog. In particular, we used Gemini 2.5 Pro to generate 120 synthetic user requests that reflect a diverse set of manually designed use cases and able to probe various system functionalities, including factual recall, multi-criteria filtering, and sophisticated product matching across a multifaceted knowledge base.

Systems' performance was evaluated manually on test questions, followed by quantitative and qualitative analysis focused on key performance indicators, including:

1. **Accuracy:** The factual correctness and completeness of the information provided, verified against the ground truth represented by the knowledge base.
2. **Error Rate:** Incorrect responses were systematically classified into three distinct categories. **FIH (Factual Inaccuracy & Hallucination)** refers to cases where the system generated factually incorrect statements or fabricated information not supported by the knowledge base. **FMR (Failure to Meet User Requirements)** denotes instances where the system disregarded explicit user constraints or misinterpreted the query, thereby providing an answer that did not adequately address the user's specific need. **RGI (Response Generation Issues)** covers technical failures, such as exceeding token limits, that prevented the generation of a complete or coherent response.

The experimental findings offer compelling evidence for the effectiveness of SofAgent's multi-agent architecture. SofAgent's higher accuracy demonstrates that its workload division, consisting of delegating tasks to specialized agents, enhances the effectiveness of information retrieval compared to the monolithic baseline, which struggled to synthesize information from diverse sources. The analysis of error types, detailed in Table 1, provides an even starker contrast, with the baseline's failures dominated by Factual Inaccuracies and Hallucinations (FIH), underscoring the vulnerability of single-agent approaches when dealing with numerous data sources. In contrast, SofAgent's architecture brought reduction in this kind of errors, confirming that grounding agents in a factual knowledge base is a highly effective strategy for mitigating hallucinations. Crucially, SofAgent's predominant errors were Failures to Meet User Requirements (FMR), suggesting its challenges lie in refining information discovery, rather than the fundamental untrustworthiness caused by inventing facts.

Inter-System Agreement Analysis: An analysis of response concordance revealed a significant divergence: the two systems agreed for only 5.00% of the questions. Crucially, in the instances of disagreement, SofAgent provided the factually correct response 60.53% of the times. Conversely, the baseline was correct in only 10.53% of these cases, while a notable 28.07% of disagreements resulted

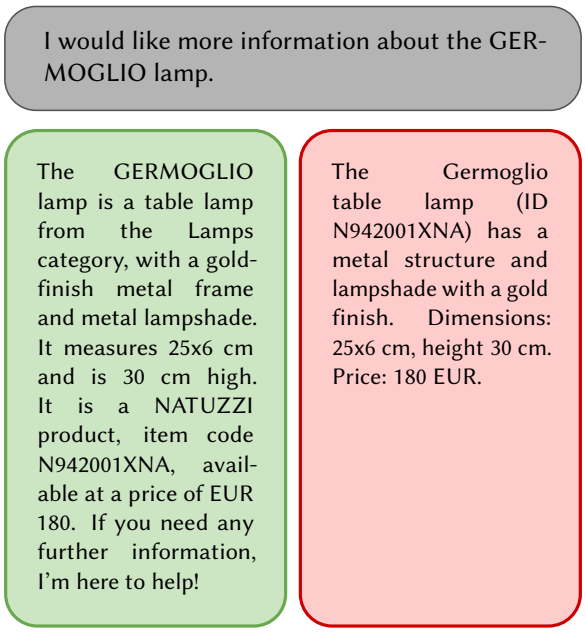
[1] https://www.langchain.com/
[2] https://github.com/sisinflab/SofAgent-IIR25

Table 1
Accuracy comparison between SofAgent and the single-agent baseline, with errors classified into three categories

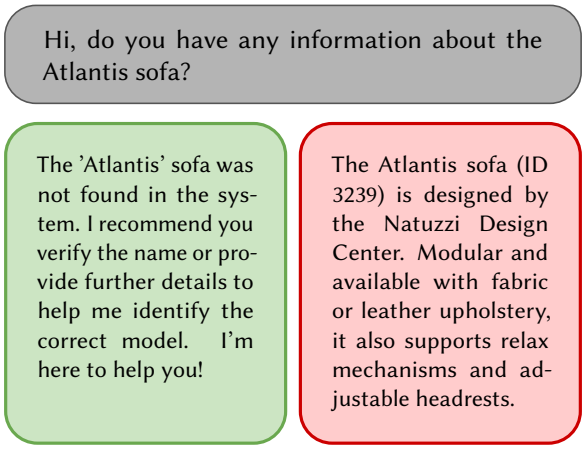
System	Accuracy	Error Rate	FIH	FMR	RGI
GPT-4o	8.33%	91.66%	88.33%	3.33%	0%
SofAgent	74.17%	25.83%	3.33%	21.67%	0.83%

in both systems providing different, factually incorrect responses, indicating the inherent difficulty of some queries. This analysis underscores that while direct agreement was rare, SofAgent was more reliable when the systems' outputs differed on factual content. It should be noted that a remaining case of disagreement, caused by a technical issue, was not considered in the previous percentages.

To illustrate the comparative performance of the systems, Figure 2 presents two distinct interaction scenarios that showcase key differences in their behavior. Figure 2(a) depicts a scenario involving a straightforward factual query. While both systems provide the correct information, SofAgent (green) delivers a more detailed and conversational response, better fulfilling the role of a helpful assistant, whereas the baseline's reply (red) is more terse. The second example, shown in Figure 2(b), highlights a more critical divergence in robustness. SofAgent correctly identifies that the requested item does not exist in the knowledge base and informs the user. In contrast, the single-agent baseline exhibits a severe hallucination, inventing a detailed but entirely incorrect description for the non-existent product.



(a) SofAgent more detailed answer



(b) SofAgent's robustness to hallucinations

Figure 2: Examples of system responses. SofAgent (green, left) vs baseline (red, right).

5. Conclusion

This paper introduced SofAgent, a MAS that demonstrated higher accuracy and a reduction in hallucinations compared to a single-agent approach for complex retail assistance. Its success stems from decomposing the workload into subtasks and distributing them among specialized agents, as well as grounding LLM reasoning in a factual, external knowledge base. Despite promising results, the experiments identified wide room for improvement. Therefore, future work will focus on refining agent coordination and retrieval logic, leveraging next-generation LLMs, and expanding capabilities to include multimodal interactions, aiming to create a more robust AI-driven customer experience.

Acknowledgments

The authors acknowledge the partial support by: “Progetto 2Next: Industry 4.0 per il Design Italiano”, “Huawei PhD Grant” and “2022LKJWHC - TRex-SE: Trustworthy Recommenders for Software Engineers”.

Declaration on Generative AI

During the preparation of this work, the author did not use any AI tool.

References

[1] Matteo Attimonelli, Danilo Danese, Daniele Malitesta, Claudio Pomo, Giuseppe Gassi, and Tommaso Di Noia. Ducho 2.0: Towards a more up-to-date unified framework for the extraction of multimodal features in recommendation. In Tat-Seng Chua, Chong-Wah Ngo, Roy Ka-Wei Lee, Ravi Kumar, and Hady W. Lauw, editors, Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024, pages 1075–1078. ACM, 2024.

[2] Salvatore Bufi, Vincenzo Paparella, Vito Walter Anelli, and Tommaso Di Noia. Legal but unfair: Auditing the impact of data minimization on fairness and accuracy trade-off in recommender systems. In Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2025, New York City, NY, USA, June 16-19, 2025, pages 114–123. ACM, 2025.

[3] Dario Di Palma, Giovanni Maria Biancofiore, Vito Walter Anelli, Fedelucio Narducci, and Tommaso Di Noia. Content-based or collaborative? insights from inter-list similarity analysis of chatgpt recommendations. In Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization, UMAP Adjunct 2025, New York City, NY, USA, June 16-19, 2025, pages 28–33. ACM, 2025.

[4] Dario Di Palma, Giovanni Servedio, Vito Walter Anelli, Giovanni Maria Biancofiore, Fedelucio Narducci, Leonarda Carnimeo, and Tommaso Di Noia. Beyond words: Can chatgpt support state-of-the-art recommender systems? In Kevin Roitero, Marco Viviani, Eddy Maddalena, and Stefano Mizzaro, editors, Proceedings of the 14th Italian Information Retrieval Workshop, Udine, Italy, September 5-6, 2024, volume 3802 of CEUR Workshop Proceedings, pages 13–22. CEUR-WS.org, 2024.

[5] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024, pages 1107–1128. Association for Computational Linguistics, 2024.

[6] Jiabao Fang, Shen Gao, Pengjie Ren, Xiuying Chen, Suzan Verberne, and Zhaochun Ren. A multi-agent conversational recommender system. CoRR, abs/2402.01135, 2024.

[7] Antonio Ferrara, Marco Valentini, Paolo Masciullo, Antonio De Candia, Davide Abbattista, Riccardo Fusco, Claudio Pomo, Vito Walter Anelli, Giovanni Maria Biancofiore, Ludovico Boratto, and Fedelucio Narducci. DIVAN: deep-interest virality-aware network to exploit temporal dynamics in news recommendation. In Proceedings of the Recommender Systems Challenge 2024, RecSysChallenge 2024, Bari, Italy, October 14-18, 2024, pages 12–16. ACM, 2024.

[8] Diego Gosmar and Deborah A. Dahl. Hallucination mitigation using agentic AI natural language-based frameworks. CoRR, abs/2501.13946, 2025.

[9] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Trans. Inf. Syst., 43(2):42:1–42:55, 2025.

[10] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL:

communicative agents for "mind" exploration of large language model society. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023.

[11] Alberto Carlo Maria Mancino, Salvatore Bufi, Angela Di Fazio, Antonio Ferrara, Daniele Malitesta, Claudio Pomo, and Tommaso Di Noia. Datarec: A python library for standardized and reproducible data management in recommender systems. In Nicola Ferro, Maria Maistro, Gabriella Pasi, Omar Alonso, Andrew Trotman, and Suzan Verberne, editors, Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025, pages 3478–3487. ACM, 2025.

[12] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. Augmented language models: a survey. Trans. Mach. Learn. Res., 2023, 2023.

[13] Vincenzo Paparella, Vito Walter Anelli, Franco Maria Nardini, Raffaele Perego, and Tommaso Di Noia. Post-hoc selection of pareto-optimal solutions in search and recommendation. In Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos, editors, Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023, pages 2013–2023. ACM, 2023.

[14] Qiyao Peng, Hongtao Liu, Hua Huang, Qing Yang, and Minglai Shao. A survey on llm-powered agents for recommender systems. CoRR, abs/2502.10050, 2025.

[15] Marco Valentini. Cooperative and competitive llm-based multi-agent systems for recommendation. In Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto, editors, Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part V, volume 15576 of Lecture Notes in Computer Science, pages 204–211. Springer, 2025.

[16] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. Frontiers Comput. Sci., 18(6):186345, 2024.

[17] Zhefan Wang, Yuanqing Yu, Wendi Zheng, Weizhi Ma, and Min Zhang. Macrec: A multi-agent collaboration framework for recommendation. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang, editors, Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024, pages 2760–2764. ACM, 2024.

[18] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023.

[19] Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents. CoRR, abs/2404.13501, 2024.

[20] Han Zhou, Xingchen Wan, Ruoxi Sun, Hamid Palangi, Shariq Iqbal, Ivan Vulic, Anna Korhonen, and Sercan Ö. Arik. Multi-agent design: Optimizing agents with better prompts and topologies. CoRR, abs/2502.02533, 2025.