Understanding Prompt Construction and Interaction Challenges in LLM-Based Multi-Robot Systems for Spatial Tasks

Xiaoran Yang
xyang49@ncsu.edu
Department of Computer Science
North Carolina State University
Raleigh, North Carolina, USA

Yang Zhan
yzhan3@ncsu.edu
Department of Computer Science
North Carolina State University
Raleigh, North Carolina, USA

Qiao Jin
qjin4@ncsu.edu
Department of Computer Science
North Carolina State University
Raleigh, North Carolina, USA

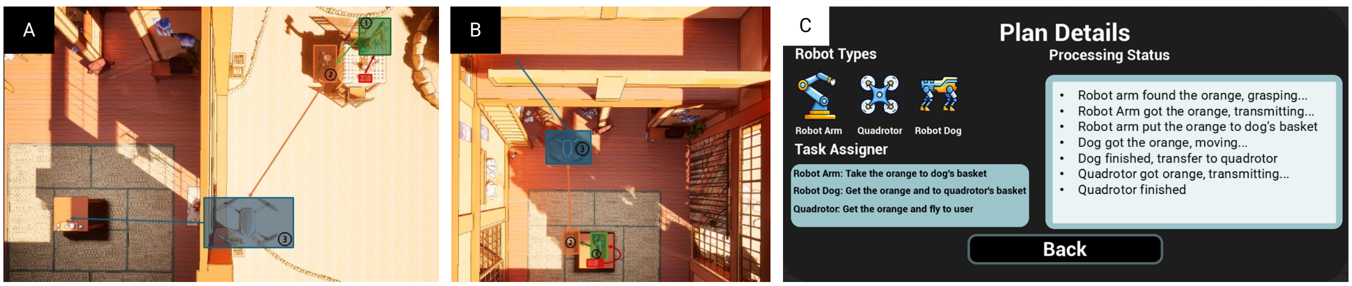

Figure 1: Two tasks (A, B) with three robots: a robotic arm (green), a robot dog (orange), and a drone (blue). The red box indicates the transported object. In A, food is retrieved from outside and delivered indoors. In B, a book is transported from a table to a second-floor bookshelf. C shows the planner interface, including robot types, task allocation, and real-time execution status.

Abstract

LLM-based multi-robot systems (MRS) can take a user's high-level goal and generate coordinated robot actions for spatial tasks. However, human-in-the-loop is still needed when the system benefits from humans providing clarifications or preferences, especially in settings where full autonomy is not yet reliable and the environment is highly uncertain. Motivated by this gap, we conducted a qualitative study (N = 12) in a VR-based multi-robot simulation environment to examine how users construct prompts for MRS and what challenges they perceive during task decomposition, allocation, and execution. Results show that users use prompts to specify robot roles, order, and end goals, and re-prompt when results are inefficient, or execution is not observable. We also found that gaps in workflow visibility, interface support, and coordination logic hinder users' understanding and their ability to intervene. We provide design implications for human-in-the-loop interfaces and prompt support in LLM-based MRS.

CCS Concepts

• **Human-centered computing** → **User studies**; **Virtual reality**; *Mixed / augmented reality*.

Keywords

Virtual Reality, Human Robot Interaction, Multi-Robot System

CHI EA '26, Barcelona, Spain

ACM ISBN 979-8-4007-2281-3/2026/04
https://doi.org/10.1145/3772363.3798326

ACM Reference Format:
Xiaoran Yang, Yang Zhan, and Qiao Jin. 2026. Understanding Prompt Construction and Interaction Challenges in LLM-Based Multi-Robot Systems for Spatial Tasks. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3772363.3798326

1 Introduction

A multi-robot collaborative system consists of two or more robots that work together to assist users in completing tasks [8, 28]. Unlike traditional single-robot systems, MRS offers advantages such as distributed perception, parallel execution, and coordinated cooperation. These systems have therefore been widely studied for solving complex tasks [17]. They have shown superior performance over single robot solutions in domains including navigation, object manipulation, exploration, search and rescue, and education [6, 7, 26, 32]. However, MRS still faces challenges in correctly decomposing user instructions into subtasks and assigning them to suitable robots [11, 12]. The rapid advancement of large language models (LLMs) provides a promising avenue for addressing this problem. LLMs possess rich world knowledge and strong capabilities in complex reasoning and natural language understanding [30, 36]. Systems such as COHERENT [18] demonstrate that LLMs can interpret users' natural language instructions, decompose them into subtasks, allocate responsibilities among heterogeneous robots, and dynamically adjust plans based on feedback. This reduces the barrier to interaction and operation. Despite these advances, deploying LLM-based MRS in real-world spatial tasks still involves interaction challenges underexplored. Much of the existing work focuses on system-level coordination and performance [16, 22, 31], human involvement is often limited to goal specification and outcome review [34]. Although involving humans may slow MRS due to extra communication overhead [27], keeping people in the loop can expand task scope without full autonomy, complement robot capabilities, increase control and help the system adapt to uncertain, changing environments [5]. Specifically, physical actions are constrained by three-dimensional geometry and embodied motion, and real-world environments are often only partially observable. As a result, planning frequently relies on assumptions that cannot be fully verified in advance. Additionally, robots operating in parallel share space and resources, which can lead to conflicts and interdependencies. These situations often require users to make preference-based decisions and risk trade-offs. These characteristics motivate us to examine LLM-based MRS from a human perspective and ask: **(RQ1) How should prompts be constructed for LLM-based MRS applied to spatial tasks? (RQ2) What challenges do users face when interacting with LLM-based MRS?**

To address these research questions, we conducted an exploratory user study using a VR-simulated MRS as the research platform [18]. This simulation provided participants with a concrete interactive context in which they could experience LLM-based multi-robot task planning, allowing us to observe how participants constructed and revised prompts during their interactions with the system. In the study, participants completed spatial tasks by issuing natural language instructions in a scenario involving three collaborating robots. Semi-structured interviews were conducted after the task was completed. The main contributions of this work are threefold: (1) We provide empirical insights into how users construct prompts to guide LLM-based MRS in spatial tasks; (2) We identify user-centered interaction challenges that influence how users remain human in the loop during task decomposition, allocation, and execution; (3) We discuss design implications to better support human-in-the-loop with MRS.

2 Methods

To investigate how users construct prompts for LLM-based MRS applied in spatial tasks and the challenges they encounter when interacting with MRS, we conducted task-based user studies in a VR simulation, analyzed semi-structured interviews using thematic analysis, and derived design implications addressing our RQs.

2.1 Participants and Setting

The study was approved by the university's ethics review board. Twelve participants (aged 22–32, M = 25.25, SD = 2.89) were recruited via internal university mailing lists and received gift cards as compensation. Participants varied in their familiarity with AI tools (2 high, 4 moderate, 6 low) and MRS (3 none, 3 basic, 6 high).

2.2 Multi-Robot System

We used a VR-based multi-robot simulation testbed to study how users prompt and intervene in an LLM-based MRS. We built the VR simulation in Unreal Engine 5 and ran it on Oculus Quest 3S headsets. As shown in Fig 1, the system consists of an interaction interface (Fig. 1C) and virtual scenes (Fig. 1A, Fig. 1B). In the virtual scenes, we constructed a cabin environment and deployed a 6 DoF robotic arm, a quadruped robot dog, and a quadrotor drone, both inside and outside the cabin. The robotic arm handles object-grasping tasks, the quadruped robot dog handles ground transportation tasks, and the quadrotor drone handles aerial transportation tasks. The interaction interface serves as the entry point to the task planner. Users initiate the task planner by issuing natural language instructions and use this interface to monitor which robots are currently executing tasks, the subtasks generated by the task planner, and the system's real-time operational status. The Task Planner coordinates task decomposition and allocation to enable efficient multi-robot collaboration within each robot's capabilities. Upon receiving a user instruction, it interprets user intent, identifies target objects, and broadcasts relevant information to robots' environment sensing modules for object detection. Each robot includes a task module, an environment sensing module, a feedback module, and a switch module. Following a standard MRS workflow [27], tasks are decomposed into subtasks and dynamically allocated based on robots' states and environmental perception. Robots execute assigned tasks using a [target, path] specification, while continuously reporting execution status. The LLM centrally controls task planning, switching, and coordination, while perception runs locally on each robot. Based on real-time feedback, the Task Planner adjusts execution order and assignments to improve robustness and efficiency, and returns planning results to the interface for visualization.

2.3 Task Design

The purpose of the tasks was to explore users' interaction needs with MRS in the context of problem solving [15]. The task design was informed by benchmark tasks used in prior work on LLM-based MRS [3, 9, 22, 33, 37], with object manipulation and retrieval serving as the core task paradigm and spanning multiple application domains. In addition, the task design explicitly considered the impact of spatial visibility on user interaction experience [25]. Based on these considerations, we designed two experimental scenarios with distinct task themes: food retrieval (task 1, as shown in Fig. 1 A) and book placement (task 2, as shown in Fig. 1 B). In the task 1 scenario, participants were asked to issue a command with the general intent of "finding food outdoors." In the task 2 scenario, participants were asked to issue a command related to "placing books near the bookshelf on the second floor," with the specific wording of the command determined by the participant. Prior work has demonstrated that visibility can have potential impacts on users' task execution [14, 25]. To account for the impact of this factor, in terms of spatial visibility, the multi-robot collaboration process was not visible to participants in task 1, whereas in task 2, the collaboration process was fully visible throughout task execution.

2.4 Study Procedure

The study lasted approximately 70 minutes. Participants first completed informed consent and a demographic questionnaire. They then underwent a brief training session in VR, during which the researcher introduced the three heterogeneous robots (a quadrotor drone, a quadruped robot dog, and a robotic arm), their capabilities, and the voice-based interaction method. Participants were allowed to phrase commands freely within the task intent. Participants subsequently completed two tasks with different spatial visualization conditions. They were guided to familiarize themselves with the interface and instructed to carefully observe the task execution process. After a short break, participants performed the formal tasks in a counterbalanced order, issuing instructions to control the system for object retrieval and placement. The researcher monitored the sessions, recorded participants' real-time reactions and verbal comments, and provided operational assistance without offering guidance. Following task completion, participants took a short rest and then participated in a 30-minute semi-structured interview focusing on overall experience, prompt construction in spatial tasks, and perceived system challenges and design implications.

3 Results

We analyzed the semi-structured interview data using thematic analysis [1, 24] to address RQs. The first author familiar with the interview transcripts, generated and refined initial codes. All authors then organized codes into candidate themes, reviewed and revised them for coherence and distinctness, and collaboratively defined and named the final themes. Finally, all authors structured the reporting and writing of the qualitative findings.

3.1 (RQ1) How Should Prompts be Constructed for LLM-based Multi-Robot System

3.1.1 How to create prompt. Participants reported that effective prompt construction required an understanding of individual robot capabilities and inter-robot interactions. As P9 noted, *"Its inherent functionality, what they are capable of doing, provided me with ideas for constructing instruction." "Robot's action-related information helps me form expectations and make predictions"* (P11). When task allocation aligned with robot capabilities, system decisions were perceived as more natural and reasonable, forming the basis for prompt construction; for example, *"Using the drone made a lot of sense to me, because both of these tasks involve height"* (P10). Participants further described prompt refinement as a primary strategy for iteration and replanning, such as *"making the command more detailed"* (P5), explicitly assigning tasks to particular robots (e.g., *"Can the drone bring something to eat from outside?"* (P11)), *"Redesign the first prompt to be more structured and step by step"* (P3), and clearly defining expected outcomes. Without such clarification, vague prompts sometimes led to results that conflicted with participants' expectations (P9).

3.1.2 When users (re)prompt. Participants expressed a desire to intervene through prompting in several situations to alter system behavior. First, when the results did not meet expectations or when task execution was inefficient, participants wanted to re-run or adjust the system via prompts. As P5 noted, *"The results are slow, or the results are incorrect,"* while P7 stated, *"When the requested objective is not met or is met but with a visible lack of efficiency."* Second, when participants felt confused about the system's results or execution process, they hoped to use prompts to better understand the underlying reasons for the system's behavior. As P3 explained, *"Revising the first prompt is most valuable when I cannot observe or understand the full execution pipeline of the robot system."* Finally, when aiming for better overall system performance, participants wanted to refine prompts to improve outcomes. P9 mentioned this motivation occurring *"after several successful task executions."*

3.2 (RQ2) What Challenges do Users Face When Interacting with LLM-based Multi-Robot System

3.2.1 Lack of Decision Transparency and Task Execution Visibility Undermines User Understanding and Confidence. During system use, participants reported insufficient visual cues to support decision making and task execution. The system did not adequately convey robots' identities, spatial locations, or collaboration processes, leading to uncertainty and reduced trust. As P12 noted, *"I don't know exactly where the handoff between them happens."* Limited execution feedback and the absence of real-time progress indicators further hindered participants' understanding of system status. P4 stated, *"when I see the robots, I only get textual prompts, which doesn't allow me to intuitively grasp the current progress."* To improve decision transparency, participants suggested interactive confirmation and feedback mechanisms, such as restating planned actions before execution, for instance, *"there needs to be a feedback mechanism to confirm that what I wanted it to do has already been done"* (P2), as well as providing explanations of the system's reasoning process. P8 stated, *"make it more concrete, such as showing its thinking process."* To enhance execution visibility, participants proposed richer visualizations, including 3D overview thumbnails for global awareness, such as *"adding a top-down projection on the side so you can see how each robot moves"* (P8), and first-person views from individual robots to better convey ongoing actions, as P3 explained, *"the panel could show each robot's real-time viewpoint."*

3.2.2 Mismatch Between Textual Information and Spatial Context Increases Cognitive Load. At the interaction level, the current system design required participants to constantly divide their attention, thereby increasing cognitive load. First, purely text-based interaction elements resulted in high information density and cognitive overload. P3 noted that *"there are too many elements, and the information density is quite high."* Second, the placement and presentation of information panels demanded frequent attention shifts. As P12 observed, *"the panel blocks the robots, and I have to glance at it repeatedly, which makes it hard to see the overall robot operation."* As a result, participants had to distribute their attention between the panel and the spatial task environment, making it difficult to maintain focus. In response to this challenge, participants proposed several design improvements. They expressed a desire for spatial visualizations that embed system information, robot representations, and movement trajectories directly into the environment using highlighting or see-through effects. P4 suggested, *"if the route could be displayed directly in the scene rather than confined to a small panel, it would be more intuitive."* Participants also suggested anchoring task and status information directly to specific robots through spatial visualization; for example, *"the text could be directly anchored to the specific robot"* (P8).

3.2.3 Limited Flexibility in Task Allocation and Collaboration Logic. Participants identified several challenges related to task planning and inter-robot collaboration logic. One major issue was that task allocation often did not align with users' expectations of optimality. Specifically, participants felt that task assignments did not reflect what they perceived as the most efficient solution. In Task 2, P9 commented, *"the drone could probably do the whole thing—it could* *just pick it up from the table and deliver it directly."* Rather than prioritizing optimal paths or minimal completion time, the system appeared to favor showcasing multi-robot collaboration. As a result, many participants felt the collaboration strategy lacked flexibility. Specifically, the system failed to dynamically adjust the number of robots involved based on task characteristics and difficulty. P11 remarked, *"the steps waste resources—something that could clearly be done by one robot is unnecessarily split into three, which makes the system feel rigid."* To address this issue, participants proposed several design directions. First, task allocation should be grounded in robot capabilities and shortest path principles, with efficiency and effectiveness as primary goals. This includes dynamically adjusting the number of participating robots based on task complexity and each robot's strengths, leading to collaboration that is both reasonable and efficient. As P10 stated, *"task allocation should be designed around dynamic adjustment of the number of robots."* P4 similarly noted the importance of *"leveraging each robot's strengths and having them support each other within a task to achieve the goal."* Second, collaboration logic should be better aligned with robots' actual action capabilities. P4 emphasized, *"collaborative task assignment should be linked more closely to what robots can actually do in practice."* Similar sentiments were echoed by others: *"couldn't the drone just handle everything?"* (P10).

4 Discussion

4.1 Design Implications

Our study reveals key practices in prompt construction and interaction challenges for LLM-based MRS. Based on these insights, we propose three design implications to better support human-in-the-loop interaction in LLM-based MRS.

4.1.1 Designing Prompts by Integrating Task and Robot Information through Prompt Scaffolding. As shown in section 3.1.1, when vague instructions are problematic, users must tailor prompts to robot capabilities and execution details. Many users refine vague prompts into specific instructions. Prompt engineering spans features, knowledge, reasoning, planning, and reliability [19], but incorporating comprehensive information into one input is difficult. We suggest prompt scaffolding, which guides complex reasoning [29] by breaking tasks into a series of related prompts that build on each other. Instead of one-off prompts, the process begins with the overall goal, then decomposes it into subgoals as the task progresses, with prompt content revealed step by step according to the user's state and prompt completeness. Early stages focus on outcomes; later prompts include robot capabilities, their relationship to task requirements, and execution-level details like actions and collaboration logic. This structure helps users revise prompts and supports efficient, accurate instructions in real-time MRS. Prompt information is presented as system-feasible and system-recommended prompts [21]. As discussed in section 3.2.3, the timing at which prompts are presented and the content they convey remain critical within a prompt scaffolding framework. While step-by-step prompts are designed under the prompt scaffolding structure, it is also necessary to carefully determine when each prompt should appear and ensure that the content follows the principles of the shortest path and optimal robot selection to meet user expectations.

4.1.2 Make the system's current planning objective explicit and allow users to switch or trade off objectives. As shown in section 3.1.2, prompt input quality depends on user expertise and state, with experts and alert users producing better prompts [10, 23]. Therefore, prompt guidance should adapt dynamically to each user's skills and current condition. Inspired by proactive agent [20, 35], we propose a prompt-guidance agent (PGA) that detects user state and expertise, adjusting its support accordingly. When prompts are strong, the robot encourages independence; when quality is low or the user is idle, it offers detailed feedback or visual cues to re-engage the user. At the same time, informed by the insights from section 3.2.3, the PGA can generate multiple prompts hint-based on different goals and allow users to switch between or trade off among these goals, such as prioritizing the fastest completion, minimizing the number of robots used, selecting the safest path, enabling energy-saving modes, or choosing other alternative strategies. This adaptive approach ensures efficient, accurate control and improves system usability for all users.

4.1.3 Build spatial visual interfaces that integrate both global and detailed information to enhance decision transparency and execution visibility. As shown in sections 3.2.1 and 3.2.2, the qualitative study revealed key challenges: lack of transparency in decision-making, limited execution visibility, and mismatch between text and spatial information, which increase cognitive load and reduce user confidence. To address this, users suggested using inherently spatial visualizations—rather than traditional 2D methods—to improve transparency and visibility [2, 4, 13]. Our approach focuses on spatial and temporal visualization. Spatial visualization links information directly to specific robots and their tasks, letting users see robot capabilities and interactions within the environment. Path visualization and dialog boxes clarify coordination and show current and future movements. To reduce cognitive load, a 3D mini map summarizes room layout, robot positions, and movements, especially when information is spread out. Temporal visualization adds a timeline, allowing users to review system states and robot actions over time, with highlights for parallel execution. Together, these visualizations improve user understanding and system transparency.

5 Conclusion & Future Work

We conducted an empirical study using a VR-based simulated MRS to investigate how users prompt LLM-based multi-robot systems and the challenges they encounter. Findings show that users ground prompts in robot capabilities and execution details, with prompting serving as the primary mechanism for control and replanning, while current systems lack decision transparency and execution visibility. We derive design implications, including spatialized information visualization and integrated prompt guidance, to support more transparent and user-centered MRS interaction. Although limited by a small sample size, restricted tasks, and a simulated environment, this exploratory study provides actionable insights, which future work will extend to real-world scenarios and AR-based multi-robot interfaces.

References

[1] Sirwan Khalid Ahmed, Ribwar Arsalan Mohammed, Abdulqadir J Nashwan, Radhwan Hussein Ibrahim, Araz Qadir Abdalla, Barzan Mohammed M Ameen, and Renas Mohammed Khdhir. 2025. Using thematic analysis in qualitative research. *Journal of Medicine, Surgery, and Public Health* 6 (2025), 100198.

[2] Fereshteh Amini, Sébastien Rufiange, Zahid Hossain, Quentin Ventura, Pourang Irani, and Michael J McGuffin. 2014. The impact of interactivity on comprehending 2D and 3D visualizations of movement data. *IEEE transactions on visualization and computer graphics* 21, 1 (2014), 122–135.

[3] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. 2023. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on robot learning*. PMLR, 287–318.

[4] Carlos D Correa and Kwan-Liu Ma. 2010. Visibility histograms and visibility-driven transfer functions. *IEEE Transactions on Visualization and Computer Graphics* 17, 2 (2010), 192–204.

[5] Victor Nan Fernandez-Ayala, Xiao Tan, and Dimos V Dimarogonas. 2023. Distributed barrier function-enabled human-in-the-loop control for multi-robot systems. In *2023 IEEE International Conference on Robotics and Automation, ICRA 2023, London, United Kingdom of Great Britain and Northern Ireland, May 29 2023-Jun 2 2023*. Institute of Electrical and Electronics Engineers (IEEE), 7706–7712.

[6] Rafael Fierro, Luiz Chaimowicz, and Vijay Kumar. 2018. Multi-robot cooperation. In *Autonomous Mobile Robots*. CRC Press, 417–460.

[7] Roee M Francos and Alfred M Bruckstein. 2023. On the role and opportunities in teamwork design for advanced multi-robot search systems. *Frontiers in Robotics and AI* 10 (2023), 1089062.

[8] Avinash Gautam and Sudeept Mohan. 2012. A review of research in multi-robot systems. In *2012 IEEE 7th international conference on industrial and information systems (ICIIS)*. IEEE, 1–5.

[9] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2022. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608* (2022).

[10] Lucas Jasper Jacobsen and Kira Elena Weber. 2025. The promises and pitfalls of large language models as feedback providers: A study of prompt engineering and the quality of AI-driven feedback. *AI* 6, 2 (2025), 35.

[11] Shyam Sundar Kannan, Vishnunandan LN Venkatesh, and Byung-Cheol Min. 2024. Smart-llm: Smart multi-agent robot task planning using large language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 12140–12147.

[12] Alaa Khamis, Ahmed Hussein, and Ahmed Elmogy. 2015. Multi-robot task allocation: A review of the state-of-the-art. *Cooperative robots and sensor networks 2015* (2015), 31–51.

[13] Attila Kovari. 2024. AI for decision support: Balancing accuracy, transparency, and trust across sectors. *Information* 15, 11 (2024), 725.

[14] Brigitte Lapeyre, Sylvain Hourlier, Xavier Servantie, Bernard N'Kaoua, and Hélène Sauzéon. 2011. Using the Landmark–Route–Survey framework to evaluate spatial knowledge obtained from synthetic vision systems. *Human factors* 53, 6 (2011), 647–661.

[15] Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth* 1, 1 (2024), 9.

[16] Zhaoxing Li, Yue Wang, Wenbo Wu, Yanran Xu, and Sebastian Stein. 2025. HMCF: A human-in-the-loop multi-robot collaboration framework based on large language models. In *International Conference on Principles and Practice of Multi-Agent Systems*. Springer, 150–167.

[17] Pedro U. Lima and Luís M. M. Custódio. 2005. Multi-Robot Systems. In *Innovations in Robot Mobility and Control*. https://api.semanticscholar.org/CorpusID:46149598

[18] Kehui Liu, Zixin Tang, Dong Wang, Zhigang Wang, Xuelong Li, and Bin Zhao. 2025. Coherent: Collaboration of heterogeneous multi-robot system with large language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 10208–10214.

[19] Yao-Yang Liu, Zhen Zheng, Feng Zhang, Jin-Cheng Feng, Yi-Yang Fu, Ji-Dong Zhai, Bing-Sheng He, Xiao Zhang, and Xiao-Yong Du. 2026. A comprehensive taxonomy of prompt engineering techniques for large language models. *Frontiers of Computer Science* 20, 3 (2026), 2003601.

[20] Yaxi Lu, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, et al. 2024. Proactive agent: Shifting llm agents from reactive responses to active assistance. *arXiv preprint arXiv:2410.12361* (2024).

[21] Stephen MacNeil, Andrew Tran, Joanne Kim, Ziheng Huang, Seth Bernstein, and Dan Mogil. 2023. Prompt middleware: Mapping prompts for large language models to UI affordances. *arXiv preprint arXiv:2307.01142* (2023).

[22] Zhao Mandi, Shreeya Jain, and Shuran Song. 2024. Roco: Dialectic multi-robot collaboration with large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 286–299.

[23] Cristina Martinez Montes and Ranim Khojah. 2025. Emotional Strain and Frustration in LLM Interactions in Software Engineering. *arXiv preprint arXiv:2504.10050* (2025).

[24] Michael Muller. 2014. Curiosity, creativity, and surprise as analytic tools: Grounded theory method. In *Ways of Knowing in HCI*. Springer, 25–48.

[25] Jayam Patel, Tyagaraja Ramaswamy, Zhi Li, and Carlo Pinciroli. 2021. Transparency in multi-human multi-robot interaction. *arXiv preprint arXiv:2101.10495* (2021).

[26] Jorge Pena Queralta, Jussi Taipalmaa, Bilge Can Pullinen, Victor Kathan Sarker, Tuan Nguyen Gia, Hannu Tenhunen, Moncef Gabbouj, Jenni Raitoharju, and Tomi Westerlund. 2020. Collaborative multi-robot search and rescue: Planning, coordination, perception, and active vision. *Ieee Access* 8 (2020), 191617–191643.

[27] Yara Rizk, Mariette Awad, and Edward W Tunstel. 2019. Cooperative heterogeneous multi-robot systems: A survey. *ACM Computing Surveys (CSUR)* 52, 2 (2019), 1–31.

[28] Daniel Set ´o-Rey. 2021. Multi-Robot Systems: A Brief Technological Review. https://api.semanticscholar.org/CorpusID:251310786

[29] Ana Teixeira de Melo, Letícia Renault, Leo SD Caves, Philip Garnett, Paula Duarte Lopes, Raquel Ribeiro, and Filipe Santos. 2025. An AI tool for scaffolding complex thinking: challenges and solutions in developing an LLM prompt protocol suite. *Cognition, Technology & Work* 27, 3 (2025), 651–693.

[30] Shu Wang, Muzhi Han, Ziyuan Jiao, Zeyu Zhang, Ying Nian Wu, Song-Chun Zhu, and Hangxin Liu. 2024. Llm^ 3: Large language model-based task and motion planning with motion failure reasoning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 12086–12092.

[31] Yongdong Wang, Runze Xiao, Jun Younes Louhi Kasahara, Ryosuke Yajima, Keiji Nagatani, Atsushi Yamashita, and Hajime Asama. 2024. Dart-llm: Dependency-aware multi-robot task decomposition and execution using large language models. *arXiv preprint arXiv:2411.09022* (2024).

[32] Sean Wilson, Paul Glotfelter, Li Wang, Siddharth Mayya, Gennaro Notomista, Mark Mote, and Magnus Egerstedt. 2020. The robotarium: Globally impactful opportunities, challenges, and lessons learned in remote-access, distributed control of multirobot systems. *IEEE Control Systems Magazine* 40, 1 (2020), 26–44.

[33] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. 2023. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots* 47, 8 (2023), 1087–1102.

[34] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

[35] Bufang Yang, Lilin Xu, Liekang Zeng, Kaiwei Liu, Siyang Jiang, Wenrui Lu, Hongkai Chen, Xiaofan Jiang, Guoliang Xing, and Zhenyu Yan. 2025. ContextAgent: Context-Aware Proactive LLM Agents with Open-World Sensory Perceptions. *arXiv preprint arXiv:2505.14668* (2025).

[36] Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S Yu. 2023. Large language models for robotics: A survey. *arXiv preprint arXiv:2311.07226* (2023).

[37] Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. 2023. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485* (2023).