A Survey on Large Language Model based Human-Agent Systems

Anonymous ACL submission

Abstract

Recent advances in large language models (LLMs) have sparked growing interest in building fully autonomous agents. However, fully autonomous LLM-based agents still face significant challenges, including limited reliability due to hallucinations, difficulty in handling complex tasks, and substantial safety and ethical risks, all of which limit their feasibility and trustworthiness in real-world applications. To overcome these limitations, LLM-based human-agent systems (LLM-HAS) incorporate human-provided information, feedback, or control into the agent system to enhance system performance, reliability and safety. This paper provides the first comprehensive and structured survey of LLM-HAS. It clarifies fundamental concepts, systematically presents core components shaping these systems, including environment & profiling, human feedback, interaction types, orchestration and communication, explores emerging applications, and discusses unique challenges and opportunities. By consolidating current knowledge and offering a structured overview, we aim to foster further research and innovation in this rapidly evolving interdisciplinary field. Paper lists and resources are available at GitHub repository.

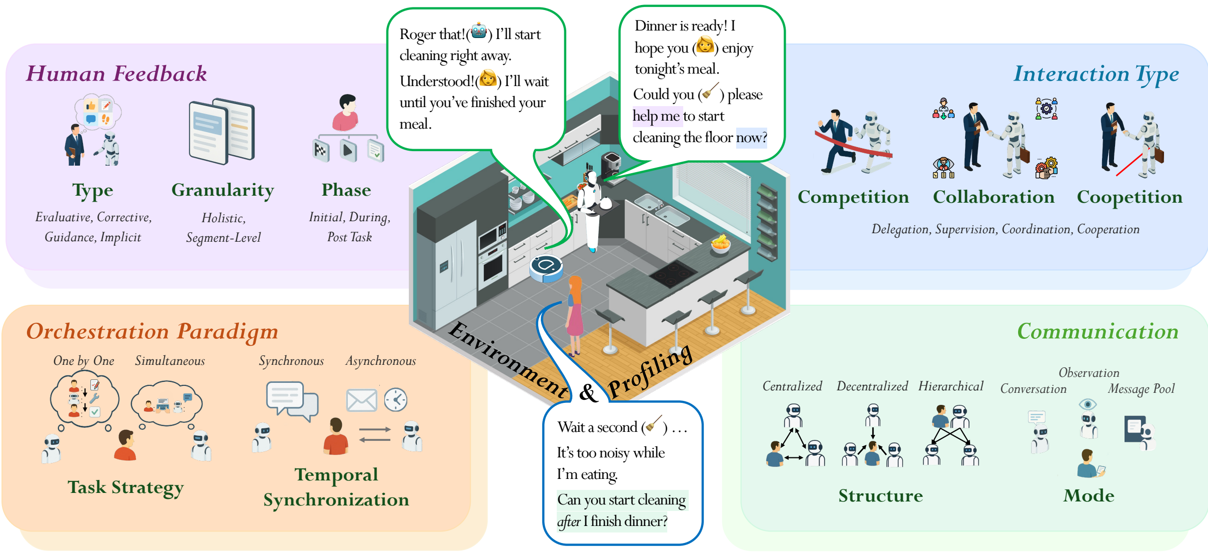
1 Introduction

Recent advances in Large Language Models (LLMs) have led to growing enthusiasm for building fully autonomous agent systems that use LLMs as a central engine to perceive environments, make decisions, and execute actions to achieve goals (Wang et al., 2024a; Li et al., 2024a). These agents are often equipped with modules for memory, planning, and tool use, aiming to automate complex workflows with minimal human involvement (Xie et al., 2024; Xi et al., 2025). However, the pursuit of *full autonomy* faces critical hurdles. ***(1) Reliability*** remains a major concern due to LLMs' propensity for hallucination, generating plausible but factually incorrect or nonsensical outputs, which undermines trust and can lead to significant errors, especially when actions are chained (Gosmar and Dahl, 2025; Xu et al., 2024; Glickman and Sharot, 2025). ***(2) Complexity*** often stalls autonomous agents; they struggle with very complicated tasks requiring deep domain expertise, long multi-step execution, nuanced reasoning, dynamic adaptation, or strict long-context consistency dependencies, as seen in scientific research (Feng et al., 2024; Yehudai et al., 2025). ***(3) Safety and Ethical Risks*** escalate with autonomy; agents can take unintended harmful actions, amplify societal biases present in training data, or create accountability gaps, particularly in critical decision-making scenarios involving finance, healthcare, or security (Mitchell et al., 2025; Deng et al., 2024; Wang et al., 2024c).

The persistence of these challenges suggests that full autonomy may be unsuitable for many real-world applications (Mitchell et al., 2025; Natarajan et al., 2025) and underscores a crucial insight often overlooked in the drive for pure automation: the indispensable role of human involvement. Humans are frequently needed to provide additional information, essential clarification, or domain knowledge, offer vital feedback and corrections, and exercise necessary oversight and control. These motivate a paradigm shift towards systems explicitly designed for human-agent collaboration: ***LLM-based Human-Agent Systems*** **(LLM-HAS)**.

While surveys on LLM-based autonomous agents (Wang et al., 2024a; Li et al., 2024a), multi-agent systems (Tran et al., 2025; Wu et al., 2025), and specific applications exist (Wang et al., 2025b; Peng et al., 2025), a dedicated synthesis focusing specifically on LLM-based human-agent systems is lacking. This survey fills the gap by providing a comprehensive and structured overview of

[*] Equal Contribution. [†] Corresponding Author.



LLM-based Human-Agent Systems (LLM-HAS)

Figure 1: Overview of LLM-based Human-Agent Systems (LAM-HAS). LLM-HAS are interactive frameworks where humans actively provide additional information, feedback, or control during interaction with an LLM-powered agent to enhance system performance, reliability, and safety. The system is composed of five core components: **Environment & Profiling** (including environment settings, and role definitions, goals, and agent capabilities such as planning and memory), **Human Feedback** (with varying types, timing, and granularity), **Interaction Types** (collaborative, competitive, cooperative, or mixed), **Orchestration** (task strategy and temporal synchronization), and **Communication** (information flow structure and mode).

the LLM-HAS. It clarifies the fundamental concepts (Section 2) and systematically presents its core components (Section 3), unique challenges and opportunities (Section 5), emerging applications (Section B), and implementation frameworks as well as datasets and benchmarks (C) within this specific niche. To the best of our knowledge, this is still the first survey on LLM-based human-agent systems. We aim to consolidate current knowledge and inspire further research and application in this rapidly evolving field. An open-source GitHub repository is maintained to provide a sustainable resource complementing our survey paper.

2 LLM-Based Human-Agent Systems

We define LLM-based human-agent systems as interactive frameworks where humans actively provide additional information, feedback, or control during interaction with an LLM-powered agent to enhance system performance, reliability, and safety (Feng et al., 2024; Shao et al., 2024; Mehta et al., 2024). The core idea is synergy: combining unique human strengths—like intuition, creativity, expertise, ethical judgment, and adaptability—with LLM agent capabilities such as vast knowledge recall, computational speed, and sophisticated language processing. LLM-HAS builds upon core LLM agent components but places critical emphasis on the human's interactive role and capabilities:

(1) ***Providing Information / Clarification:*** Humans provide additional information that agents lack or cannot reliably infer, such as login credentials, payment details, domain expertise, constraints, or resolve ambiguities, helping agents interpret situations more accurately (Naik et al., 2025; Kim et al., 2025).

(2) ***Providing Feedback / Error Correction:*** Humans evaluate agent outputs and provide feedback, ranging from simple ratings to complex critiques, demonstrations or corrections, effectively guiding agents' adjustment (Gao et al., 2024b; Dutta et al., 2024; Li et al., 2024b).

(3) ***Taking Control / Action:*** In high-stakes or sensitive scenarios (e.g., healthcare, privacy, or ethics), humans retain the authority to override, redirect, or halt agent actions, ensuring accountability, safety, and alignment with human values (Chen et al., 2025; Natarajan et al., 2025; Xiao and Wang, 2023).

Figure 1 provides a generalized overview of LLM-based human-agent systems. These systems operate within a defined **Environment** (e.g., physical world, simulation) that provides context and stimuli. **Human & Agent Profiling** characterizes the participants' roles and goals, and the agent's core LLM engine augmented with capabilities like planning, memory, and tool use. **Human Feedback** can occur during different phases in various types and granularities. Human-Agent **Interaction Types** may be collaborative (most common), competitive, cooperative, or mixed. The **Orchestration** layer governs high-level coordination, choosing a task strategy (e.g., sequential one-by-one versus parallel simultaneous execution) and a temporal synchronization mode (real-time synchronous exchanges versus delayed asynchronous workflows) so that each actor acts at the right moment. The **Communication** layer specifies how information flows, defining message structure (centralized, decentralized, hierarchical) and mode (conversation, observation signals, or shared message pools). The effective interplay and configuration of these components, along with various human feedback, are critical for tailoring the system to specific tasks and optimizing the overall system's performance. The taxonomy of LLM-based human-agent systems is outlined in Figure 3. A detailed and structured categorization of representative works is provided in the Table 5 and Table 6.

3 Core Components

In this section, we examine LLM-HAS through five core aspects: environment & profiling, human feedback, interaction type, orchestration paradigm, and communication. These dimensions provide a unified standard for analyzing existing work and guiding the design of future systems.

3.1 Environment and Profiling

Environment Setting. The environment in LLM-HAS defines a shared interaction space that can exist either in the physical world, such as offices (Sun et al., 2024b), or in fully simulated virtual environments where agents and humans engage under controlled conditions (Sun et al., 2024b; Zhang et al., 2024a; Guo et al., 2024b). These systems can be configured in various ways, including single-human single-agent, single-human multi-agent, multi-human single-agent, and multi-human multi-agent setups, each reflecting different collaboration dynamics and complexities.

Human & Agent Profiling. Human participants can be broadly categorized as *lazy* or *informative* users. Lazy users provide minimal guidance, typically offering evaluative feedback such as binary correctness or scalar rating. In contrast, informative users engage deeply by offering demonstrations, detailed guidance, refinements, or even taking over parts of the task (Wang et al., 2024b; Liu et al., 2024b; Han et al., 2025). On the other side, agents are profiled by their roles and capabilities, which range from versatile general assistants to specialized experts in mathematics, engineering, medicine, or robotic cleaning, each adapted to the particular demands of its operational context (Guo et al., 2024a; Samuel et al., 2024).

3.2 Human Feedback

Human Feedback Type. We categorize human feedback as *evaluative*, *corrective*, *guidance*, and *implicit* feedback. ***(1) Evaluative Feedback*** provides an assessment of the agent's output quality, typically as preference ranking, scalar rating, or binary assessment. A prime example is preference ranking, where users compare agent outputs, forming the basis of Reinforcement Learning from Human Feedback (RLHF) (Chaudhari et al., 2024). Alternatively, platforms like Uni-RLHF (Yuan et al., 2024) support scalar ratings or binary assessments. ***(2) Corrective Feedback*** offers direct edits or fixes to the agent's behavior. For instance, the PRELUDE (Gao et al., 2024a) framework learns latent preferences from user edits made to agent-generated text. ***(3) Guidance Feedback*** means the human proactively provides instructions, critiques, or demonstrations to shape the agent's behavior. Agents like InteractGen (Sun et al., 2024b), AutoManual (Chen et al., 2024a) can be bootstrapped using initial demonstrations, while methods like Self-Refine (Choudhury and Sodhi, 2025) employ iterative critiques and refinements to improve outputs. ***(4) Implicit Feedback*** is inferred by the agent observing user actions or control signals, rather than explicitly stated or direct output modifications. For example, an agent might learn user priorities by observing how a user adjusts control sliders in a system like VeriPlan (Lee et al., 2025a), or infer preferences by analyzing user behaviors like clicks and purchases in frameworks such as AgentA/B (Wang et al., 2025a). This contrasts with corrective feedback, where the user directly edits the output; here, the agent interprets the user's independent actions or control choices.

Human Feedback Granularity. Human feedback also varies in granularity, from coarse-grained,

Dimension	Category	Definition Summary	Key Characteristics / Trade-offs	Example Work
Type	*Evaluative*	User provides an **assessment** of the agent's output quality, typically as **binary assessment**, **scalar rating**, or **preference ranking**.	① Easy to collect, scalable. ② Less specific signal for improvement.	*EmoAgent* (Qiu et al., 2025), *MINT* (Wang et al., 2024b), *SOTOPIA* (Zhou et al., 2024)
	Corrective	User **offers edits or fixes** to the agent's behavior.	① Highly informative, clear signal for improvement. ② Higher user effort, often fine-grained & interactive.	*SymbioticRAG* (Sun et al., 2025), *SWEET-RL* (Zhou et al., 2025), *AI Chains* (Wu et al., 2022)
	Guidance	User proactively provides **instructions**, **demonstrations**, or **critiques** to shape the agent's behavior.	① Bootstraps learning, conveys complex goals, proactive alignment. ② Requires clear specification from user.	*Drive As You Speack* (Cui et al., 2024), *Hierarchical Agent*(Liu et al., 2023b), *Ask Before Plan* (Zhang et al., 2024c)
	Implicit	**Inferred by the agent observing user actions or control signals**, rather than explicitly stated or direct output modifications.	① Natural, unobtrusive collection. ② Ambiguous, requires careful interpretation.	*MTOM*(Zhang et al., 2024b), *Attentive Supp.* (Tanneberg et al., 2024a), *MineWorld* (Guo et al., 2025)
Granularity	*Coarse-grained / Holistic*	Single assessment/signal for **an entire agent output**, **trajectory**, or **task outcome**.	① Simple for user, good for overall assessment ② Obscures specific errors, less precise learning signal.	*AssistantX* (Sun et al., 2024a), *Help Feedback* (Mehta et al., 2024), *AXIS* (Lu et al., 2024)
	Fine-grained / Segment-Level	Feedback targeting **specific parts of agent output**, **actions**, or **process**.	① Precise learning signal, crucial for debugging complex skills ② Potentially higher user effort/burden.	*Collaborative Gym* (Shao et al., 2024), *Prison Dilemm* (Jiang et al., 2025), *FineArena* (Xu et al., 2025)
Phase	*Initial Setup & Goal Definition*	Feedback provided **before** task execution, **configuring** the agent system and **defining** the **task**, **goals**, **constraints**, and **preference**.	① Initial and proactive alignment, prevents costly errors, sets constraints ② Requires upfront user input.	*AgentCoord* (Pan et al., 2024a), *GDfC* (Wang et al., 2025c), *SMALL* (Wang et al., 2024c)
	During Task Execution	Online, interactive feedback **while the agent is actively performing the task**, enabling **real-time adaptation**.	① Enables real-time adaptation, crucial for dynamic/collaborative tasks ② Requires timely notification and responsive interfaces.	*InteractGen* (Sun et al., 2024b), *CowPilot* (Huq et al., 2025), *EasyLAN* (Pan et al., 2024b)
	Post-Task Eval. & Refinement	Feedback provided **after task completion** to assess outcomes and **provide suggestions** for **future use**.	① Non-disruptive, good for aggregate data/offline learning ② No impact on completed task.	*HRT-ML* (Liu et al., 2024b), *M3HF* (Wang et al., 2025d), *MAIH* (Wang et al., 2024c)

Table 1: Dimensions of Human Feedback in LLM-based human–agent systems, including feedback type, granularity, and phase. For each dimension, a summary, key characteristics, and example works are provided for comparison. A detailed overview of human feedback types and their subtypes is provided in our appendix (Table 4).

holistic judgments to fine-grained, segment-level critiques. ***(1) Coarse-grained/Holistic feedback*** provides a single assessment for the entire agent output. Standard RLHF often relies on holistic preferences between complete responses, which simplifies feedback collection but struggles with credit assignment in complex tasks. ***(2) Fine-grained/Segment-Level Feedback*** by contrast, targets specific parts (e.g., sentences, paragraphs, code blocks). This is crucial in environments like ConvCodeWorld (Han et al., 2025), where feedback pertains to specific conversational turns or generated code segments, or in annotation tasks like PDFChatAnnotator (Tang et al., 2024), where feedback applies to specific annotations or parts of the document. This finer granularity provides more precise learning signals, crucial for debugging complex behaviors.

Human Feedback Phase. Human feedback can be incorporated at different phases of the LLM-agent pipeline (Wang et al., 2025d). ***(1) Initial Setup & Goal Definition*** occurs before task execution, configuring the agent system and defining goals, such as setting coordination strategies (AgentCoord (Pan et al., 2024a)) or critiquing plans before execution (Ask-before-Plan (Zhang et al., 2024c)). ***(2) During Task Execution*** involves online, interactive feedback while the agent is actively performing the task, enabling real-time adaptation. Examples include interactive instruction editing (InstructEdit (Wang et al., 2023)), mid-task refinements (Mutual Theory of Mind (Zhang et al., 2024b), Collaborative Gym (Shao et al., 2024)), or online interventions (HG-DAgger (Kelly et al., 2019)). ***(3) Post-Task Evaluation & Refinement*** happens after task completion to assess outcomes

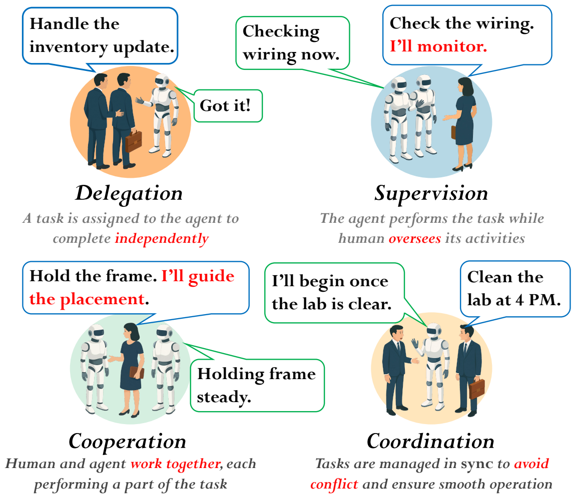


Figure 2: The subtype of the collaboration between humans and LLM-based agents.

and provide feedback for future use. Frameworks like MAIH (Wang et al., 2024c) and EmoAgent (Qiu et al., 2025) apply feedback loops after initial generation for benchmarking or offline learning, while AdaPlanner (Sun et al., 2023) archives successful plans post-task as skills for future use. Table 1 summarizes different dimensions of human feedback, key characteristics, and example work.

3.3 Human-Agent Interaction Types

Interaction types define how individuals communicate, exchange information, and take actions with one another. In LLM-HAS, interactions tend to be more dynamic and complex compared to multi-agent systems. This complexity arises from the various roles and responsibilities assigned to both human agents and those based on LLMs, necessitating a finer-grained framework to describe their collaborative behaviors. The following categorization highlights the three key interaction types: **Collaboration**, **Competition**, and **Coopetition**.

3.3.1 Collaboration

Collaborations are by far the most common interaction and foundational interaction, which involve humans and LLM-based agents working together to achieve a common goal. This partnership combines human creativity and contextual understanding with LLM-based agents to address challenges and improve the efficiency and quality of results (Vats et al., 2024; Du et al., 2024; Sun et al., 2025). Depending on the type of collaboration considered, it can be categorized into four main fine-grained subtypes (Figure 2): ***(1) Delegation & Direct Command*** (Kiewiet and McCubbins, 1991), ***(2) Supervision*** (Loganbill et al., 1982), ***(3) Cooperation*** (Rand and Nowak, 2013), and ***(4) Coordination*** (Turvey, 1990).

Delegation & Direct Command. In this interaction modality, a controlling party, usually a human, assigns explicit tasks to the LLM-based agent by providing clear and direct instructions. The agent is expected to execute these directives autonomously or on behalf of humans, ensuring that responsibilities are well-defined and actions align with the system's overarching objectives. Unlike supervision, where strategies can be dynamically adjusted in response to new situations, delegation involves providing instructions upfront. This means the agent follows a predetermined set of tasks rather than adapting to the situation. For instance, an investor specifies their risk preference to the agent executing the investment strategy like FineArena (Xu et al., 2025), or a driver utters the command to LLM-based agent like Drive as you Speak (Cui et al., 2024).

Supervision. Supervision is the process by which one party, usually a human operator, oversees, monitors, and guides the actions of an LLM-based agent. This involves real time evaluation and intervention to ensure the agent's output aligns with established goals and quality standards. Supervision also encompasses setting alert thresholds and providing corrective inputs when deviations occur. By maintaining a continuous feedback loop between the human and the agent, supervision helps calibrate agent behavior, catch and mitigate errors before they propagate and build confidence in the system. It also enables agents to handle routine tasks with increasing independence. For instance, agents notify humans to verify alignment (Liu et al., 2023b), and teleoperators monitor the LLM-generated motion plans (Liu et al., 2023a).

Cooperation. Cooperation refers to the voluntary and joint efforts of multiple parties to achieve agreed-upon goals. This collaboration type combines the various efforts and outcomes of different individuals and LLM-based agents toward a common objective. It emphasizes collective commitment, mutual assistance, and the pooling of resources to attain a shared result, thereby fostering a collaborative problem-solving environment. For instance, the human robot coordination in household activities (Chang et al., 2024), the cooperative embodied language agent (CoELA) (Zhang et al., 2024a), human designers collaborating with an LLM-based agent (Sharma et al., 2024).

Coordination. Coordination is the organized process of aligning and synchronizing the actions of humans and LLM-based agents to achieve a shared objective. Unlike cooperation, the key idea behind coordination is to avoid conflict and bias in both humans and LLM-based agents to reach the final goal. It involves clear communication, strategic planning, and the intentional division of tasks, ensuring that individual efforts are harmonized and effectively integrated to support common goals. For example, humans and agents work in a shared workspace to complete interdependent tasks (Zhang et al., 2024b), human–agent integration supports adaptive decision-making (Sun et al., 2024b), and the collaborative framework facilitates coordination between humans and agents (Pan et al., 2024a).

3.3.2 Competition

Competition is a form of interaction where participants aim to achieve their own goals, which often conflict with the objectives of others. In the LLM-HAS, competition emerges when agents or humans seek to enhance their personal performance or obtain resources, even if it negatively impacts collective results. In addition, competition also necessitates effective balancing mechanisms, like performance regulation or conflict resolution strategies, to prevent unproductive behaviors and ensure that the overall goals of the system remain intact. For instance, the SOTOPIA framework simulates social behaviors between humans and LLM-based agents (Zhou et al., 2024).

3.3.3 Coopetition

Coopetition is an interaction where cooperation and competition coexist at the same time. Within this interaction, participants collaborate on shared tasks or mutual goals while also seeking to outdo each other to improve their own performance or gain extra advantages. In terms of the LLM-HAS, this dual aspect implies that agents and human may join forces to address complex issues while competing in specific domains such as efficiency or precision. This approach not only combines the strengths of both collaboration and competition, but also fosters innovation driven by competitive incentives while also reaping the benefits of cooperative synergy. Successfully managing coopetition typically requires mechanisms for building trust and adaptable strategies that reconcile collective advantages with personal aspirations, which is a challenge for the LLM-HAS. For example, humans and agents play the prisoner's dilemma in the shared workspace (Jiang et al., 2025).

Orchestration Paradigm	**Description**
Task Strategy	**What order and grouping** of tasks do participants perform?
One-by-One	**Actors take turns** (e.g., human plans → agent executes → human reviews → agent refines).
Simultaneous	**Actors work in parallel** (e.g., agent streams partial suggestions while human types).
Temporal Synchronization	**When and how tightly** do actors' steps need to align in time?
Synchronous	(1) **Real-time interaction**: Humans and agents communicate simultaneously or in immediate sequence; (2) **Immediate response**: Participants expect or require prompt feedback. (e.g., live chat session, real-time voice assistant).
Asynchronous	(1) **Delayed interaction**: Communication occurs without the expectation of immediate responses; (2) **Flexible timing**: Participants can respond at their convenience. (e.g., email queues, human leaves comments, agent processes offline).

Table 2: Orchestration paradigms in LLM-based human–agent systems encompass two orthogonal dimensions: task strategy, which can be one-by-one or simultaneous, and temporal synchronization, which can be synchronous or asynchronous.

3.4 Orchestration Paradigm

The orchestration paradigm in LLM-HAS refers to *how* tasks and interactions are managed between humans and agents, covering two dimensions in our survey: **Task Strategy** (*ordering*) and **Temporal Synchronization** (*timing*). Table 2 summarizes the two dimensions of the orchestration paradigm.

3.4.1 Task Strategy

In LLM-HAS, the chosen task strategy, defined by the order and grouping of tasks performed by humans and agents, significantly impacts task execution effectiveness and efficiency. These strategies can typically be categorized into *one-by-one* and *simultaneous* paradigms.

One-by-One. The one-by-one strategy requires participants (humans and LLM-based agents) to perform tasks sequentially, taking clearly defined turns. For example, a human first outlines a plan, the agent then executes the task, the human subsequently reviews the output, and finally, the agent refines its work based on feedback (Liu et al., 2024a; Zhou et al., 2025). Such sequential interaction helps maintain a clear order of execution and reduces the complexity associated with concurrent task management. However, this rigidity may limit overall efficiency and flexibility, especially in dynamic scenarios requiring parallel processing or rapid interaction cycles (Bansal et al., 2024; Guo et al., 2024b).

Simultaneous. Simultaneous strategy describes an interaction pattern in which LLM-based agents and humans respond concurrently in real time. Compared to the one-by-one strategy, the simultaneous approach more closely mirrors real-world conditions encountered in many simulation tasks (Wang et al., 2025d; Zhang et al., 2025). However, this strategy demands sophisticated mechanisms to handle latency mitigation and seamless coordination between participants.

3.4.2 Temporal Synchronization

Temporal synchronization in LLM-HAS refers to the timing and coordination of interactions between humans and agents. It significantly influences system responsiveness, user experience, and task efficiency. It can be broadly categorized into two modes: *synchronous* and *asynchronous*.

Synchronous. Synchronous interaction involves real-time interactions where humans and agents engage simultaneously. Immediate response is expected, facilitating dynamic exchanges. Examples include live chat sessions, real-time voice assistants (e.g., Siri, Alexa), and collaborative decision-making scenarios (Zhang et al., 2024b; Liu et al., 2023b). This mode is advantageous for tasks requiring rapid responses, immediate clarification, or real-time collaboration (Mehta et al., 2024; Han et al., 2025).

Asynchronous. In contrast, asynchronous interaction occurs without the necessity for immediate responses. Participants can engage at their convenience, allowing for flexibility in communication. Examples include email exchanges, message queues, ticket-based support systems, and task assignments where agents process and report outcomes independently (Shao et al., 2024; Zhang et al., 2025). Asynchronous communication is beneficial for complex issues that require thoughtful analysis or when participants are in different time zones (Sun et al., 2024b,a).

3.5 Communication

The communication layer in LLM-HAS specifies how information flows, defining **communication structure** (*centralized*, *decentralized*, *hierarchical*) and **mode** (*conversation*, *observation signals*, or *shared message pools*). Due to space constraints, a detailed introduction is provided in Section A.

4 Applications and Resources

A diverse range of applications, implementation tools, and resources has emerged for LLM-HAS. We elaborate on the five most frequent application domains in Section B, summarize the corresponding datasets and benchmarks in Table 3, and provide a detailed introduction to representative open-source LLM-HAS frameworks in Section C.

5 Challenges and Opportunities

In this section, we highlight some existing challenges and opportunities for LLM-HAS.

Human Flexibility and Variability. Human feedback varies widely in terms of role, timing, and style across various LLM-HAS. Humans are often subjective, influenced by their personalities, which means different individuals interacting with an LLM-HAS may lead to different outcomes and conclusions. This highlights the need and opportunity for i) thorough investigations or benchmarks on how varied human feedback affects entire systems, and ii) flexible frameworks that can support and adapt to diverse human feedback. In addition, humans, regarded as a "special agent" in the LLM-HAS, are subject to fewer restrictions and evaluations than LLM-based agents. This limits how the LLM-HAS can be improved because the impedance may be on the human side instead of the agent. This concern remains and requires a refined strategy to define the strict, fine interaction rule and evaluation equally for both human and LLM-based agents. Also, many studies today substitute real human participants with LLM simulated human proxies, failing to capture human input's variety and unpredictability.

The performance gap between humans and the simulated human remains unknown, potentially making the comparison incomparable.

Mostly Agent-Centered Work. In most LLM-HAS studies, guidance flows in a single direction, with humans evaluating agent outputs and providing corrective or evaluative feedback. Namely, the current studies are mostly agent-centered. However, enabling agents to observe human actions, detect errors or inefficiencies, and offer timely suggestions can transform collaboration and reduce human effort by leveraging agent intelligence. When agents act as instructors by proposing alternative strategies, drawing attention to overlooked risks, and reinforcing effective practices as tasks unfold in real time, both humans and agents benefit. We believe that exploring human-centered LLM-HAS, or shifting toward an equalized LLM-HAS, will unlock the full promise of teamwork between humans and agents.

Inadequate Evaluation Methodologies. In existing evaluation frameworks for LLM-HAS, improvements focus primarily on agent accuracy and static benchmarks, which ignore the real burden placed on human collaborators (Ma et al., 2025). People dedicate varying amounts of time, attention and cognitive effort depending on the type and frequency of feedback they must provide, yet no standard metric captures this human workload or its impact on overall efficiency. Evaluation methods should measure factors such as time spent offering feedback, perceived mental workload and effort required to detect and correct errors, and they should cover every phase of the human agent collaboration from initial task assignment through post execution review. As human expertise and LLM-based agent capabilities merge to deliver unprecedented performance, both uncertainty and variability grow. A new evaluation approach or set of metrics that systematically and comprehensively quantifies contributions and costs for both humans and agents is essential to ensure truly efficient collaboration.

Unresolved Safety Vulnerabilities. Most LLM-HAS works emphasize improving agent performance and have left safety, robustness and privacy underexplored in the context of human interaction (Qiu et al., 2025). As people and LLM-based agents collaborate in dynamic workflows, the risk of misaligned behavior, unexpected failures, or unintended disclosure of sensitive information grows. Humans engaging with these systems need clear safeguards around data sharing, error recovery protocols when agents behave unpredictably and privacy protections that cover every stage of the interaction. Robustness measures must ensure agents handle ambiguous or adversarial inputs without passing harm on to their human partners (Glickman and Sharot, 2025). Without studies that emphasize human experience in safety and privacy design, real-world deployments will struggle to gain trust or meet acceptable risk thresholds. Rigorous investigation of how safety, robustness and privacy shape human agent workflows from design through deployment is essential to build collaborations that are both effective and respectful of human needs.

Applications and Beyond. The potential of LLM-HAS extends well beyond current applications. Many opportunities remain to be explored in challenging domains such as healthcare, finance, scientific research, education, and so on (Luo et al., 2025; Guo et al., 2024a). While fully autonomous LLM-based agent systems encounter difficulties in handling complex, long-term tasks and earning full trust in safety and reliability, the involvement of humans to provide additional information, feedback, and control allows LLM-HAS to greatly improve overall system performance and safety. This opens the door to impactful applications across a broad range of critical fields.

6 Conclusion

This paper presents a comprehensive review of LLM-based Human-Agent Systems. We introduce a structured taxonomy covering five core dimensions: environment and profiling, human feedback, interaction types, orchestration paradigms, and communication, and use it to classify and analyze existing research on LLM-HAS. We also summarize representative implementation frameworks, benchmark datasets, and evaluation metrics to support reproducibility and comparative analysis. Finally, we identify key challenges and unresolved issues in current LLM-HAS research. These issues remain major obstacles to the development of effective, adaptive, safe and trustworthy human-agent systems. We hope this review offers a comprehensive understanding of the LLM-HAS landscape and serves as a practical guide for future research.

Limitations

Although we strive to include a wide range of representative works (e.g., ACL, EMNLP, NAACL, EACL, COLM, NeurIPS, ICLR, ICML, etc.), some relevant research may not be included, especially recent preprints or interdisciplinary research in fields such as cognitive science. At the same time, although this review briefly discusses safety issues, it does not fully explore broader ethical and social impacts, including the allocation of responsibilities, long-term coexistence of humans and machines, and the consistency of values. These issues deserve further investigation in future work.

References

Adem Ait, Javier Luis Cánovas Izquierdo, and Jordi Cabot. 2024. Towards modeling human-agentic collaborative workflows: A bpmn extension. *arXiv preprint arXiv:2412.05958*.

Hao Bai, Yifei Zhou, Mert Cemri, Jiayi Pan, Alane Suhr, Sergey Levine, and Aviral Kumar. 2024. Digirl: Training in-the-wild device-control agents with autonomous reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 12461–12495. Curran Associates, Inc.

Gagan Bansal, Jennifer Wortman Vaughan, Saleema Amershi, Eric Horvitz, Adam Fourney, Hussein Mozannar, Victor Dibia, and Daniel S Weld. 2024. Challenges in human-agent communication. *arXiv preprint arXiv:2412.10380*.

Matthew Chang, Gunjan Chhablani, Alexander Clegg, Mikael Dallaire Cote, Ruta Desai, Michal Hlavac, Vladimir Karashchuk, Jacob Krantz, Roozbeh Mottaghi, Priyam Parashar, and 1 others. 2024. Partnr: A benchmark for planning and reasoning in embodied multi-agent tasks. *arXiv preprint arXiv:2411.00081*.

Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. 2024. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *arXiv preprint arXiv:2404.08555*.

Minghao Chen, Yihang Li, Yanting Yang, Shiyu Yu, Binbin Lin, and Xiaofei He. 2024a. Automanual: Generating instruction manuals by LLM agents via interactive environmental learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Ying-Jung Chen, Chi-Sheng Chen, and Ahmad Albarqawi. 2025. Reinforcing clinical decision support through multi-agent systems and ethical ai governance. *arXiv preprint arXiv:2504.03699*.

Yixin Chen, Guoxi Zhang, Yaowei Zhang, Hongming Xu, Peiyuan Zhi, Qing Li, and Siyuan Huang. 2024b. Synergai: Perception alignment for human-robot collaboration. *arXiv preprint arXiv:2409.15684*.

Sanjiban Choudhury and Paloma Sodhi. 2025. Better than your teacher: LLM agents that learn from privileged AI feedback. In *The Thirteenth International Conference on Learning Representations*.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. 2024. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 902–909.

Chengyuan Deng, Yiqun Duan, Xin Jin, Heng Chang, Yijun Tian, Han Liu, Yichen Wang, Kuofeng Gao, Henry Peng Zou, Yiqiao Jin, Yijia Xiao, Shenghao Wu, Zongxing Xie, Weimin Lyu, Sihong He, Lu Cheng, Haohan Wang, and Jun Zhuang. 2024. Deconstructing The Ethics of Large Language Models from Long-standing Issues to New-emerging Dilemmas: A Survey. *arXiv e-prints*, arXiv:2406.05392.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, and 3 others. 2023. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, and 21 others. 2024. LLMs assist NLP researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099, Miami, Florida, USA. Association for Computational Linguistics.

Subhabrata Dutta, Timo Kaufmann, Goran Glavaš, Ivan Habernal, Kristian Kersting, Frauke Kreuter, Mira Mezini, Iryna Gurevych, Eyke Hüllermeier, and Hinrich Schuetze. 2024. Problem solving through human-ai preference-based cooperation. *arXiv preprint arXiv:2408.07461*.

Xueyang Feng, Zhi-Yuan Chen, Yujia Qin, Yankai Lin, Xu Chen, Zhiyuan Liu, and Ji-Rong Wen. 2024. Large language model-based human-agent collaboration for complex task solving. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1336–1357, Miami, Florida, USA. Association for Computational Linguistics.

Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. 2024a. Aligning LLM agents by learning latent preference from user edits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jie Gao, Simret Araya Gebreegziabher, Kenny Tsu Wei Choo, Toby Jia-Jun Li, Simon Tangi Perrault, and Thomas W Malone. 2024b. A taxonomy for human-llm interaction modes: An initial exploration. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–11.

Yiming Gao, Feiyu Liu, Liang Wang, Zhenjie Lian, Dehua Zheng, Weixuan Wang, Wenjin Yang, Siqin Li, Xianliang Wang, Wenhui Chen, and 1 others. 2024c. Enhancing human experience in human-agent collaboration: A human-centered modeling approach based on positive human gain. *arXiv preprint arXiv:2401.16444*.

Christos Gkournelos, Christos Konstantinou, and Sotiris Makris. 2024. An llm-based approach for enabling seamless human-robot collaboration in assembly. *CIRP Annals*, 73(1):9–12.

Moshe Glickman and Tali Sharot. 2025. How human–ai feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 9(2):345–359.

Ran Gong, Qiuyuan Huang, Xiaojian Ma, Hoi Vo, Zane Durante, Yusuke Noda, Zilong Zheng, Song-Chun Zhu, Demetri Terzopoulos, Li Fei-Fei, and 1 others. 2023. Mindagent: Emergent gaming interaction. *arXiv preprint arXiv:2309.09971*.

Diego Gosmar and Deborah A Dahl. 2025. Hallucination mitigation using agentic ai natural language-based frameworks. *arXiv preprint arXiv:2501.13946*.

Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. 2025. Mineworld: a real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024a. Large language model based multi-agents: a survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8048–8057.

Xudong Guo, Kaixuan Huang, Jiale Liu, Wenhui Fan, Natalia Vélez, Qingyun Wu, Huazheng Wang, Thomas L. Griffiths, and Mengdi Wang. 2024b. Embodied LLM agents learn to cooperate in organized teams. In *Language Gamification - NeurIPS 2024 Workshop*.

Hojae Han, Seung-won Hwang, Rajhans Samdani, and Yuxiong He. 2025. Convcodeworld: Benchmarking conversational code generation in reproducible feedback environments. *arXiv preprint arXiv:2502.19852*.

Allyson I Hauptman, Beau G Schelble, Nathan J McNeese, and Kapil Chalil Madathil. 2023. Adapt and overcome: Perceptions of adaptive autonomous agents for human-ai teaming. *Computers in Human Behavior*, 138:107451.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, and 1 others. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 3(4):6.

Faria Huq, Zora Zhiruo Wang, Frank F. Xu, Tianyue Ou, Shuyan Zhou, Jeffrey P. Bigham, and Graham Neubig. 2025. Cowpilot: A framework for autonomous and human-agent collaborative web navigation. *Preprint*, arXiv:2501.16609.

Guanxuan Jiang, Yuyang Wang, and Pan Hui. 2025. Experimental exploration: Investigating cooperative interaction behavior between humans and large language model agents. *Preprint*, arXiv:2503.07320.

Yucheng Jiang, Yijia Shao, Dekun Ma, Sina Semnani, and Monica Lam. 2024. Into the unknown unknowns: Engaged human learning through participation in language model agent conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9917–9955, Miami, Florida, USA. Association for Computational Linguistics.

Seth Karten, Mycal Tucker, Huao Li, Siva Kailas, Michael Lewis, and Katia Sycara. 2023. Interpretable learned emergent communication for human–agent teams. *IEEE Transactions on Cognitive and Developmental Systems*, 15(4):1801–1811.

Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J Kochenderfer. 2019. Hg-dagger: Interactive imitation learning with human experts. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8077–8083. IEEE.

D Roderick Kiewiet and Mathew D McCubbins. 1991. *The logic of delegation*. University of Chicago Press.

JiWoo Kim, Minsuk Chang, and JinYeong Bak. 2025. Beyond turn-taking: Introducing text-based overlap into human-llm interactions. *arXiv preprint arXiv:2501.18103*.

Christine Lee, David J. Porfirio, Xinyu Jessica Wang, Kevin Zhao, and Bilge Mutlu. 2025a. Veriplan: Integrating formal verification and llms into end-user planning. *ArXiv*, abs/2502.17898.

Dong Won Lee, Yubin Kim, Denison Guvenoz, Sooyeon Jeong, Parker Malachowsky, Louis-Philippe Morency, Cynthia Breazeal, and Hae Won Park. 2025b. The human robot social interaction (hsri) dataset: Benchmarking foundational models' social reasoning. *arXiv preprint arXiv:2504.13898*.

Ido Levy, Ben Wiesel, Sami Marreed, Alon Oved, Avi Yaeli, and Segev Shlomov. 2024. St-webagentbench: A benchmark for evaluating safety and trustworthiness in web agents. *arXiv preprint arXiv:2410.06703*.

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

Youquan Li, Miao Zheng, Fan Yang, Guosheng Dong, Bin Cui, Weipeng Chen, Zenan Zhou, and Wentao Zhang. 2024b. Fb-bench: A fine-grained multi-task benchmark for evaluating llms' responsiveness to human feedback. *arXiv preprint arXiv:2410.09412*.

Jonghan Lim, Sujani Patel, Alex Evans, John Pimley, Yifei Li, and Ilya Kovalenko. 2024. Enhancing human-robot collaborative assembly in manufacturing systems using large language models. In *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, pages 2581–2587.

Haokun Liu, Yaonan Zhu, Kenji Kato, Izumi Kondo, Tadayoshi Aoyama, and Yasuhisa Hasegawa. 2023a. Llm-based human-robot collaboration framework for manipulation tasks. *arXiv preprint arXiv:2308.14972*.

Haokun Liu, Yaonan Zhu, Kenji Kato, Atsushi Tsukahara, Izumi Kondo, Tadayoshi Aoyama, and Yasuhisa Hasegawa. 2024a. Enhancing the llm-based robot manipulation through human-robot collaboration. *IEEE Robotics and Automation Letters*.

Jijia Liu, Chao Yu, Jiaxuan Gao, Yuqing Xie, Qingmin Liao, Yi Wu, and Yu Wang. 2023b. Llm-powered hierarchical language agent for real-time human-ai coordination. *ArXiv*, abs/2312.15224.

Shipeng Liu, FNU Shrutika, Boshen Zhang, Zhehui Huang, and Feifei Qian. 2024b. Effect of adaptive communication support on human-ai collaboration. *arXiv preprint arXiv:2412.06808*.

Carol Loganbill, Emily Hardy, and Ursula Delworth. 1982. Supervision: A conceptual model. *The counseling psychologist*, 10(1):3–42.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30.

Junting Lu, Zhiyang Zhang, Fangkai Yang, Jue Zhang, Lu Wang, Chao Du, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. 2024. Turn every application into an agent: Towards efficient human-agent-computer interaction with api-first llm-based agents. *arXiv preprint arXiv:2409.17140*.

Xing Han Lù, Zdeněk Kasner, and Siva Reddy. 2024. Weblinx: Real-world website navigation with multi-turn dialogue. *arXiv preprint arXiv:2402.05930*.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, and 1 others. 2025. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*.

Qianou Ma, Dora Zhao, Xinran Zhao, Chenglei Si, Chenyang Yang, Ryan Louie, Ehud Reiter, Diyi Yang, and Tongshuang Wu. 2025. Sphere: An evaluation card for human-ai systems. *arXiv preprint arXiv:2504.07971*.

Nikhil Mehta, Milagro Teruel, Xin Deng, Sergio Figueroa Sanz, Ahmed Awadallah, and Julia Kiseleva. 2024. Improving grounded language understanding in a collaborative environment by interacting with agents through help feedback. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1306–1321, St. Julian's, Malta. Association for Computational Linguistics.

Yannick Metz, David Lindner, Raphaël Baur, and Mennatallah El-Assady. 2024. Mapping out the space of human feedback for reinforcement learning: A conceptual framework. *arXiv preprint arXiv:2411.11761*.

Margaret Mitchell, Avijit Ghosh, Alexandra Sasha Luccioni, and Giada Pistilli. 2025. Fully autonomous ai agents should not be developed. *arXiv preprint arXiv:2502.02649*.

Riya Naik, Ashwin Srinivasan, Estrid He, and Swati Agarwal. 2025. An empirical study of the role of incompleteness and ambiguity in interactions with large language models. *arXiv preprint arXiv:2503.17936*.

Sriraam Natarajan, Saurabh Mathur, Sahil Sidheekh, Wolfgang Stammer, and Kristian Kersting. 2025. Human-in-the-loop or ai-in-the-loop? automate or collaborate? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28594–28600.

Bo Pan, Jiaying Lu, Ke Wang, Li Zheng, Zhen Wen, Yingchaojie Feng, Minfeng Zhu, and Wei Chen. 2024a. Agentcoord: Visually exploring coordination strategy for llm-based multi-agent collaboration. *arXiv preprint arXiv:2404.11943*.

Lihang Pan, Yuxuan Li, Chun Yu, and Yuanchun Shi. 2024b. A human-computer collaborative tool for training a single large language model agent into a network through few examples. *arXiv preprint arXiv:2404.15974*.

Qiyao Peng, Hongtao Liu, Hua Huang, Qing Yang, and Minglai Shao. 2025. A survey on llm-powered agents for recommender systems. *arXiv preprint arXiv:2502.10050*.

Jiahao Qiu, Yinghui He, Xinzhe Juan, Yiming Wang, Yuhan Liu, Zixin Yao, Yue Wu, Xun Jiang, Ling

Yang, and Mengdi Wang. 2025. Emoagent: Assessing and safeguarding human-ai interaction for mental health safety. *arXiv preprint arXiv:2504.09689*.

David G Rand and Martin A Nowak. 2013. Human cooperation. *Trends in cognitive sciences*, 17(8):413–425.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*.

SeungWon Seo, SeongRae Noh, Junhyeok Lee, SooBin Lim, Won Hee Lee, and HyeongYeop Kang. 2025. Reveca: Adaptive planning and trajectory-based validation in cooperative language agents using information relevance and relative proximity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23295–23303.

Kathrin Seßler, Arne Bewersdorff, Claudia Nerdel, and Enkelejda Kasneci. 2025. Towards adaptive feedback with ai: Comparing the feedback quality of llms and teachers on experimentation protocols. *arXiv preprint arXiv:2502.12842*.

Yijia Shao, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. 2024. Collaborative gym: A framework for enabling and evaluating human-agent collaboration. *arXiv preprint arXiv:2412.15701*.

Ashish Sharma, Sudha Rao, Chris Brockett, Akanksha Malhotra, Nebojsa Jojic, and Bill Dolan. 2024. Investigating agency of llms in human-ai collaboration tasks. *Preprint*, arXiv:2305.12815.

Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. 2018. Learning when to communicate at scale in multiagent cooperative and competitive tasks. *arXiv preprint arXiv:1812.09755*.

Jonathan Spencer, Sanjiban Choudhury, Matthew Barnes, Matthew Schmittle, Mung Chiang, Peter Ramadge, and Siddhartha Srinivasa. 2020. Learning from interventions: Human-robot interaction as both explicit and implicit feedback. In *16th robotics: science and systems, RSS 2020*. MIT Press Journals.

Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. 2023. Adaplanner: Adaptive planning from feedback with language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nan Sun, Bo Mao, Yongchang Li, Lumeng Ma, Di Guo, and Huaping Liu. 2024a. Assistantx: An llm-powered proactive assistant in collaborative human-populated environment. *arXiv preprint arXiv:2409.17655*.

Nan Sun, Chengming Shi, and Yuwen Dong. 2024b. Interactgen: Enhancing human-involved embodied task reasoning through llm-based multi-agent collaboration. In *under review*.

Qiang Sun, Tingting Bi, Sirui Li, Eun-Jung Holden, Paul Duuring, Kai Niu, and Wei Liu. 2025. Symbioticrag: Enhancing document intelligence through human-llm symbiotic collaboration. *arXiv preprint arXiv:2505.02418*.

Yi Tang, Chia-Ming Chang, and Xi Yang. 2024. Pdfchatannotator: A human-llm collaborative multi-modal data annotation tool for pdf-format catalogs. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, page 419–430, New York, NY, USA. Association for Computing Machinery.

Daniel Tanneberg, Felix Ocker, Stephan Hasler, Joerg Deigmoeller, Anna Belardinelli, Chao Wang, Heiko Wersing, Bernhard Sendhoff, and Michael Gienger. 2024a. To help or not to help: Llm-based attentive support for human-robot group interactions. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9130–9137.

Daniel Tanneberg, Felix Ocker, Stephan Hasler, Joerg Deigmoeller, Anna Belardinelli, Chao Wang, Heiko Wersing, Bernhard Sendhoff, and Michael Gienger. 2024b. To help or not to help: Llm-based attentive support for human-robot group interactions. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9130–9137. IEEE.

Shahroz Tariq, Mohan Baruwal Chhetri, Surya Nepal, and Cecile Paris. 2025. A2c: A modular multi-stage collaborative decision framework for human-ai teams. *Expert Systems with Applications*, page 127318.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.

Michael T Turvey. 1990. Coordination. *American psychologist*, 45(8):938.

Vanshika Vats, Marzia Binta Nizam, Minghao Liu, Ziyuan Wang, Richard Ho, Mohnish Sai Prasad, Vincent Titterton, Sai Venkat Malreddy, Riya Aggarwal, Yanwen Xu, and 1 others. 2024. A survey on human-ai teaming with large pre-trained models. *arXiv preprint arXiv:2403.04931*.

Dakuo Wang, Ting-Yao Hsu, Yuxuan Lu, Limeng Cui, Yaochen Xie, William Headean, Bingsheng Yao, Akash Veeragouni, Jiapeng Liu, Sreyashi Nag, and 1 others. 2025a. Agenta/b: Automated and scalable web a/btesting with interactive llm agents. *arXiv preprint arXiv:2504.09723*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. 2023. Instructedit: Improving automatic masks for diffusion-based image editing with user instructions. *arXiv preprint arXiv:2305.18047*.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025b. A survey of llm-based agents in medicine: How far are we from baymax? *arXiv preprint arXiv:2502.11211*.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2024b. MINT: Evaluating LLMs in multi-turn interaction with tools and language feedback. In *The Twelfth International Conference on Learning Representations*.

Xingzhi Wang, Zhoumingju Jiang, Yi Xiong, and Ang Liu. 2025c. Human-llm collaboration in generative design for customization. *Journal of Manufacturing Systems*, 80:425–435.

Ziyan Wang, Meng Fang, Tristan Tomilin, Fei Fang, and Yali Du. 2024c. Safe multi-agent reinforcement learning with natural language constraints. *arXiv preprint arXiv:2405.20018*.

Ziyan Wang, Zhicheng Zhang, Fei Fang, and Yali Du. 2025d. M3hf: Multi-agent reinforcement learning from multi-phase human feedback of mixed quality. *arXiv preprint arXiv:2503.02077*.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Yaozu Wu, Dongyuan Li, Yankai Chen, Renhe Jiang, Henry Peng Zou, Liancheng Fang, Zhen Wang, and Philip S Yu. 2025. Multi-agent autonomous driving systems with large language models: A survey of recent advances. *arXiv preprint arXiv:2502.16804*.

Zhenyu Wu, Ziwei Wang, Xiuwei Xu, Jiwen Lu, and Haibin Yan. 2023. Embodied task planning with large language models. *arXiv preprint arXiv:2307.01848*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

Hengjia Xiao and Peng Wang. 2023. Llm a*: Human in the loop large language models enabled a* search for robotics. *arXiv preprint arXiv:2312.01797*.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, and 1 others. 2024. Can large language model agents simulate human trust behavior? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Congluo Xu, Zhaobin Liu, and Ziyang Li. 2025. Finarena: A human-agent collaboration framework for financial market analysis and forecasting. *arXiv preprint arXiv:2503.02692*.

Hongshen Xu, Zichen Zhu, Lei Pan, Zihan Wang, Su Zhu, Da Ma, Ruisheng Cao, Lu Chen, and Kai Yu. 2024. Reducing tool hallucination via reliability alignment. *arXiv preprint arXiv:2412.04141*.

Bingyu Yan, Xiaoming Zhang, Litian Zhang, Lian Zhang, Ziyi Zhou, Dezhuang Miao, and Chaozhuo Li. 2025. Beyond self-talk: A communication-centric survey of llm-based multi-agent systems. *arXiv preprint arXiv:2502.14321*.

Yang Ye, Hengxu You, and Jing Du. 2023. Improved trust in human-robot collaboration with chatgpt. *IEEE Access*, 11:55748–55754.

Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. 2025. Survey on evaluation of llm-based agents. *arXiv preprint arXiv:2503.16416*.

Yifu Yuan, Jianye HAO, Yi Ma, Zibin Dong, Hebin Liang, Jinyi Liu, Zhixin Feng, Kai Zhao, and YAN ZHENG. 2024. Uni-RLHF: Universal platform and benchmark suite for reinforcement learning with diverse human feedback. In *The Twelfth International Conference on Learning Representations*.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, and Chuang Gan. 2024a. Building cooperative embodied agents modularly with large language models. In *The Twelfth International Conference on Learning Representations*.

Shao Zhang, Xihuai Wang, Wenhao Zhang, Yongshan Chen, Landi Gao, Dakuo Wang, Weinan Zhang, Xinbing Wang, and Ying Wen. 2024b. Mutual theory of mind in human-ai collaboration: An empirical study with llm-driven ai agents in a real-time shared workspace task. *arXiv preprint arXiv:2409.08811*.

Shao Zhang, Xihuai Wang, Wenhao Zhang, Chaoran Li, Junru Song, Tingyu Li, Lin Qiu, Xuezhi Cao, Xunliang Cai, Wen Yao, and 1 others. 2025. Leveraging dual process theory in language agent framework for real-time simultaneous human-ai collaboration. *arXiv preprint arXiv:2502.11882*.

Xuan Zhang, Yang Deng, Zifeng Ren, See-Kiong Ng, and Tat-Seng Chua. 2024c. Ask-before-plan: Proactive language agents for real-world planning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10836–10863, Miami, Florida, USA. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. 2025. Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks. *Preprint*, arXiv:2503.15478.

A Communication

In LLM-HAS, communication serves as the fundamental mechanism defining the transmission, reception, and transformation of information between humans and LLM-based agents. It focuses specifically on how *information flows* across participants to support effective interaction and mutual understanding. Unlike LLM-based multi-agent systems (Yan et al., 2025), human-agent systems introduce a unique dimension (i.e., flexible, and cognitively diverse human participation). This leads to a broader and more complex communication landscape, encompassing both human-to-agent and agent-to-agent exchanges, each influenced by human interpretability, feedback style, and interaction latency.

To systematically analyze communication behavior in such systems, we propose a two-dimensional taxonomy that captures the communication behavior characteristics of humans and agents from macro-structures to micro-interaction rules. Specifically, we divide this section into the following parts: **Communication Structure**, which describes the macro-level organization of information channels, and **Communication Mode**, which characterizes the micro-level methods of message exchange.

A.1 Communication Structure

Communication structure refers to the organizational structure of agents, including both humans and agents, in LLM-HAS. It determines how information flows at the macro level and shapes the rules of interaction at the micro level. While originally developed for LLM-based multi-agent environments (Guo et al., 2024a), these structures have been effectively adapted to human-agent scenarios by treating humans as specialized agents. In such systems, the communication structure not only governs the efficiency of information exchange but also significantly impacts the system's adaptability, scalability, and robustness to human variability. We categorize the representative structures into three types: **Centralized**, **Decentralized**, and **Hierarchical**.

In **Centralized** structure, one primary agent or a group of core agents acts as a central node to coordinate all communications within the system. This central agent manages interactions among other agents, simplifying coordination and minimizing conflicts (Cui et al., 2024). **Decentralized** structure employs peer-to-peer communication, enabling direct interactions among agents without centralized control. Agents autonomously manage their communications based on systemic information, enhancing system flexibility, adaptability, and robustness (Shao et al., 2024; Driess et al., 2023). In addition, **Hierarchical** structure organizes agents into clearly defined levels, assigning distinct roles and responsibilities according to their position within the hierarchy (Liu et al., 2023b; Pan et al., 2024b). High-level agents typically fulfill managerial or strategic roles, providing overarching guidance and supervision, while lower-level agents perform specialized tasks and execute detailed operations.

A.2 Communication Mode

Communication mode defines the manner through which humans and agents exchange information within LLM-HAS. Specifically, communication mode describes the methods employed by participants to transmit, acquire, and utilize information, critically shaping interaction efficiency and the overall performance of the system. Broadly, communication modes can be categorized into three primary approaches: **Conversation**, **Observation**, and **Shared Message Pool**.

Conversation. The conversation-based mode is currently the most prevalent and intuitive approach in LLM-HAS, wherein agents and humans directly engage through natural language dialogues. This interaction format typically utilizes conversational interfaces that allow iterative exchanges, questions, clarifications, and dynamic responses, facilitating efficient collaboration and mutual understanding (Shao et al., 2024). For instance, conversational LLM agents can assist users by answering queries, explaining complex concepts, or collaboratively solving reasoning tasks through iterative dialogues (Wang et al., 2024b). While intuitive and flexible, conversational interactions rely significantly on the communicative clarity and dialogue management capabilities of LLM agents.

Observation. In the observation-based communication mode, agents acquire information implicitly by observing participants behaviors, decisions, or interactions within their environment, rather than through explicit verbal communication. This mode leverages indirect signals, including user actions, feedback cues, or behavioral traces, to infer intentions, preferences, or states (Lu

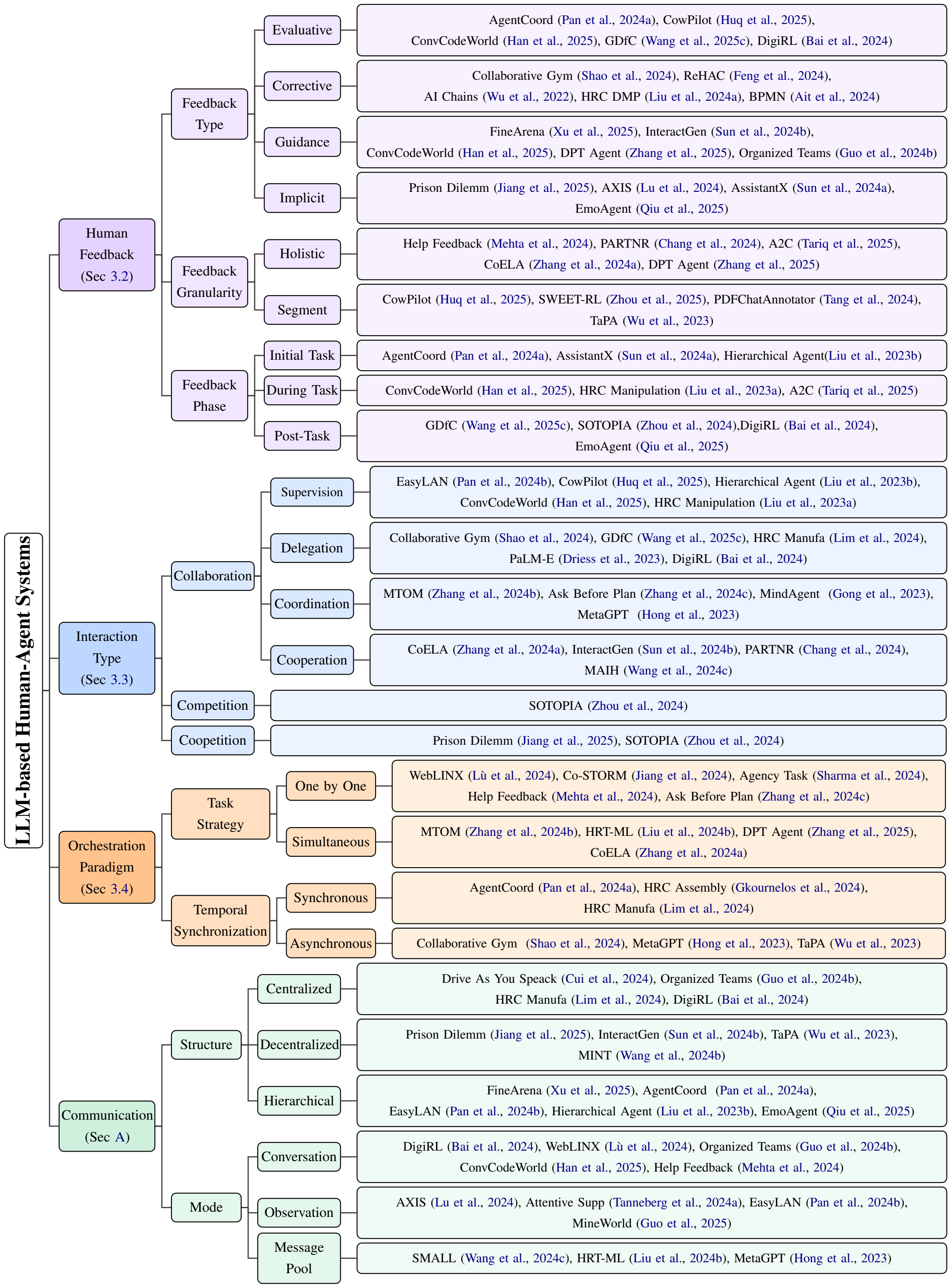


Figure 3: Taxonomy of LLM-based Human-Agent Systems. A more detailed and structured categorization of representative works is provided in the appendix (Table 5 and 6).

et al., 2024). For example, an LLM-driven tutoring system may adaptively provide targeted instructions by continuously observing student problem-solving behaviors without explicit verbal queries (Pan et al., 2024b). However, relying solely on observational signals can introduce ambiguity, potentially impacting inference accuracy unless complemented by robust inferential mechanisms.

Message Pool. The shared message pool mode involves agents and humans exchanging information through a common information repository. Participants publish messages or data into a message pool, subscribing and retrieving relevant messages based on specific interests or tasks (Sun et al., 2024a). This approach significantly simplifies direct agent-to-agent or human-to-agent interactions, reduces communication complexity, and enhances information management efficiency. A prominent example includes the MetaGPT framework (Hong et al., 2023), where LLM-based agents collaboratively retrieve information dynamically from a shared message pool, streamlining cooperation and information dissemination. Despite these advantages, shared message pools must carefully manage access control to avoid information conflicts or inefficient retrieval.

B Applications

A diverse range of applications has emerged for LLM-HAS. We elaborate on the five most frequent domains below and summarize corresponding datasets and benchmarks in Table 3. With new applications appearing almost weekly in this fast-growing field, we maintain an open-source GitHub repository to track recent developments.

Embodied AI. Applications in Embodied AI involve various aspects of dynamic and complex real-world tasks, benefiting from valuable human feedback and interactions in LLM-HAS. Ye et al. (2023) explores incorporating LLMs in human-robotic collaboration assembly tasks, allowing seamless communication between robots and humans and increasing trust in human operators. To address the challenges of false planning due to suboptimal environment changes, Seo et al. (2025) proposes REVECA to enable efficient memory management and optimal planning. Additionally, Tanneberg et al. (2024b) extends the agents' collaboration with a group of humans via Attentive Support, enabling agents' ability to remain silent to not disturb the group if desired.

Software Development. The inherently collaborative nature of software development makes human-agent collaboration vital to improve development efficiency (Lu et al., 2024; Han et al., 2025; Zhou et al., 2025). Feng et al. (2024) introduces ReHAC framework, wherein agents are trained to determine the optimal stages for human intervention within the problem-solving process, offering improved generalizability over the traditional heuristic-based approaches. Building on this direction, Zhou et al. (2025); Han et al. (2025); Wang et al. (2024b) investigate a broader spectrum of human feedback types via multi-turn human-agent interactions. These approaches incorporate carefully designed optimization objectives to effectively capture more diverse and nuanced interactions between humans and agents.

Conversational Systems. In conversational systems, due to the frequent presence of ambiguity and lack of necessary information that agents cannot reliably infer, such as login credentials and payment details, effective human-agent collaboration constitutes a critical component of the system. Zhang et al. (2024c) introduces Proactive Agent Planning, wherein agents are trained to predict classification needs based on the user-agent conversational interactions and current environment, thereby leading to improved reasoning efficacy. Wu et al. (2022) introduces Chaining the LLM to improve the quality of task outcomes and enhance the transparency and controllability of the conversational systems.

Gaming. LLM-HAS are naturally well-suited to simulated gaming environments due to their dynamicity and sophistication. Proper human-agent interactions have been shown to enhance humans' experience, satisfaction and understanding of both the environment and agents (Gong et al., 2023; Gao et al., 2024c). Collaborative interactions also contribute to improved agents' task performance and decision-making capabilities. For instance, MindAgent framework (Gong et al., 2023) illustrates the efficacy of human-agent collaboration through measurable improvements in task outcomes when humans and agents work together. Mehta et al. (2024) demonstrates agents achieve improved outcomes when interacting with humans via autonomous confusion detection and clarification questions and inquiries. Ait et al. (2024) introduces Meta-Command Communication-based framework to enable effective human-agent collaboration. To address challenges related to execution latency while maintaining strong reasoning capabilities, Liu et al. (2023a) proposes Hierarchical Language Agent that promotes faster responses, stronger cooperation, and more consistent language communications.

Finance. Given the complexity of stock markets and financial systems, where investors' strategies and risk preferences are critical determinants of out-

Domain	Datasets & Benchmarks	Proposed or Used by	Data Link
Embodied AI	TaPA	TaPA (Wu et al., 2023)	Link
	EmboInteract	InteractGen (Sun et al., 2024b)	–
	AssistantX	AssistantX (Sun et al., 2024a)	–
	IGLU Multi-Turn	Help Feedback (Mehta et al., 2024)	Link
	PARTNR	PARTNR (Chang et al., 2024)	Link
	MINT	MINT (Wang et al., 2024b)	Link
	C-WAH	REVECA (Seo et al., 2025)	Link
	HSRI	HSRI (Lee et al., 2025b)	–
Conversational Systems	WEBLINX	WebLINX (Lù et al., 2024)	–
	Ask-before-Plan	Ask Before Plan (Zhang et al., 2024c)	Link
	Agency Dialogue	Agency Task (Sharma et al., 2024)	–
	WildSeek	Co-STORM (Jiang et al., 2024)	Link
	MINT	MINT (Wang et al., 2024b)	Link
	HOTPOTQA	ReHAC (Feng et al., 2024)	Link
	StrategyQA	ReHAC (Feng et al., 2024)	Link
Software Development	MINT	MINT (Wang et al., 2024b)	Link
	InterCode	ReHAC (Feng et al., 2024)	Link
	ColBench	SWEET-RL (Zhou et al., 2025)	Link
	ConvCodeWorld	ConvCodeWorld (Han et al., 2025)	Link
	ConvCodeBench	ConvCodeWorld (Han et al., 2025)	Link
Gaming	CuisineWorld	MindAgent (Gong et al., 2023)	Link
	MineWorld	MineWorld (Guo et al., 2025)	Link
Finance	FinArena-Low-Cost	FineArena (Xu et al., 2025)	Link

Table 3: Datasets and Benchmarks across various domains.

comes, human-agent collaboration is increasingly recognized as a valuable paradigm. FinArena (Xu et al., 2025) demonstrates the potential of integrating experienced investors with advanced AI agents to support stock prediction tasks. This collaborative framework has been shown to improve investment performance, yielding competitive annualized returns and Sharpe ratios (Xu et al., 2025).

C Implementation Tools and Resources

C.1 Human-Agent Framework

This section provides a detailed introduction to three representative open-source LLM-HAS frameworks: Collaborative Gym (Shao et al., 2024), COWPILOT (Huq et al., 2025), and DPT-Agent (Zhang et al., 2025). They differ in key configuration aspects, including environment settings, interaction types, orchestration paradigms, and communication strategies. Specifically, **Collaborative Gym** (Shao et al., 2024) facilitates asynchronous interactions among humans, agents, and task environments, supporting various simulated and real-world tasks such as travel planning, data analysis, and academic writing. It emphasizes flexible, real-time collaboration and evaluates both outcomes and interaction quality, making it a robust tool for studying human-agent dynamics. **COWPILOT** (Huq et al., 2025) provides a framework for human-agent collaborative web navigation through a Chrome extension. It employs a "Suggest-then-Execute" model under human supervision, allowing dynamic interventions to enhance task completion rates and reduce human workload. It effectively demonstrates how human intervention can significantly improve agent performance. **DPT-Agent** (Zhang et al., 2025) applies Dual Process Theory (DPT) to enable real-time simultaneous human-agent interactions. It features intuitive, fast decision-making and deliberative reasoning components, employing Theory of Mind and asynchronous reflection to manage latency and adapt dynamically to human actions. This approach excels in environments requiring immediate and adaptive responses.

Other frameworks, such as **A2C** (Tariq et al., 2025), **FinArena** (Xu et al., 2025), and **human-robot collaboration framework** (Liu et al., 2023a), also contribute significantly to specific domains like cybersecurity, financial forecasting, and robotic manipulation, respectively. These frameworks further demonstrate the diverse potential and adaptability of LLM-HAS.

C.2 Datasets and Benchmarks

We summarize the commonly used datasets and benchmarks for Large Language Model-based Human-Agent Systems in Table 3. Diverse domains employ distinct methodologies for evaluating these systems, aligned closely with their unique application contexts. Within the domain of embodied AI, the primary approach involves simulated environments (Sun et al., 2024b,a; Mehta et al., 2024), designed to assess how effectively agents cooperate and execute tasks in dynamic, interactive scenarios. Another significant domain, Conversational Systems, encompasses applications such as question answering (Feng et al., 2024), website navigation (Lù et al., 2024; Levy et al., 2024), design decision assistance (Sharma et al., 2024), and travel planning (Zhang et al., 2024c), adopting benchmarks that evaluate the ability of language models to function as user-aligned conversational assistants, ensuring interactions meet user expectations. Despite the extensive application coverage of current benchmarks, there remains a clear necessity for the development of more comprehensive and standardized benchmarking frameworks.

D Evaluation Metrics

In this section, we introduce evaluation metrics specifically designed for human-agent systems across four key aspects: feedback mechanisms, adaptability, trust and safety, and communication methods. To evaluate feedback mechanisms, (Liu et al., 2024b) assesses a human-robot teaming framework using multi-modal language feedback at varying frequency levels (inactive, passive, active, superactive). (Metz et al., 2024) proposes seven metrics, expressiveness, ease, definiteness, context independence, precision, unbiasedness, and informativeness, to evaluate feedback quality. In the education domain, (Seßler et al., 2025) adopts six dimensions based on educational feedback theory. (Spencer et al., 2020) evaluates the Expert Intervention Learning (EIL) method by comparing robot performance with and without expert intervention. For adaptability, (Hauptman et al., 2023) examines how human-LLM agents respond to cyber incidents under different levels of autonomy across five NIST-defined phases. For trust and safety, (Levy et al., 2024) introduces a benchmark that evaluates web agents on their ability to comply with policies, avoid unsafe behavior, respect security constraints, and handle errors gracefully, including seeking user input when needed. Finally, (Karten et al., 2023) assess four categories of communication methods in human-agent teaming, focusing on effectiveness and interpretability within simulated environments of Predator-Prey(Lowe et al., 2017) and Traffic Junction(Singh et al., 2018).

In addition to these aspects, AXIS (Lu et al., 2024) and SYNERGAI (Chen et al., 2024b) evaluate the effectiveness and robustness of human–LLM agent systems in the domains of operating systems and embedded AI, respectively. These studies highlight how evaluation criteria can vary significantly depending on the specific task or application context, reflecting differences in system constraints, performance expectations, and interaction complexity.

E Human Feedback Type and Subtype

In this appendix, we present a detailed overview of human feedback types and their subtypes, as summarized in Table 4. This table provides concise definitions and illustrates how humans provide feedback to LLM-based agents in LLM-HAS. While the main paper introduced the broad categories of evaluative, corrective, guidance, and implicit feedback, here we expand each category into more granular subtypes, ranging from scalar ratings and preference rankings to direct edits, demonstrations, and inferred behavioral signals. Recognizing these subtypes clarifies the ways in which humans interact with LLM agents, by offering precise instructions and well-defined tasks, to enhance the accuracy and quality of generated outputs. This deeper understanding empowers users to optimize their interactions with LLM-based agents. Additionally, the systematic breakdown of human feedback provides a foundation for cross-study comparisons. It underscores the diverse strategies through which human users can guide, correct, or collaborate with LLM-based agents in a more detailed way.

F Difference with Multi-Agent Systems

While both LLM-HAS and MAS involve collaboration among multiple entities, the key distinction lies in the nature and role of the collaborating parties (Feng et al., 2024; Shao et al., 2024). Multi-agent systems are typically composed exclusively of autonomous agents—each designed to make decisions, communicate, and coordinate tasks with one another. In these MAS, each agent operates based on its own set of objectives and algorithms, and the overall behavior emerges from their interactions (Tran et al., 2025; Guo et al., 2024a).

In contrast, LLM-based human–agent systems explicitly incorporate humans as active participants within the decision-making loop (Feng et al., 2024). Rather than letting the system run purely on the combined strategies of several LLM-

Human Feedback Type	Description	How it Helps Agents
Evaluative Feedback	User provides an assessment of the agent's output quality.	Signals overall correctness or preference, guiding general alignment.
Preference Ranking	User compares two or more agent outputs and selects the preferred one.	Helps the agent learn relative quality and subjective nuances.
Scalar Rating	User assigns a numerical score (e.g., 1–5) to the agent's output.	Provides a quantitative measure of satisfaction or quality.
Binary Assessment	User indicates simple correctness (e.g., yes/no, thumbs up/down).	Offers a basic signal of success or failure.
Corrective Feedback	User modifies or directly improves the agent's output.	Provides explicit examples of desired output, enabling direct learning from errors.
Direct Edits / Refinements	User manually changes the agent's generated text or code.	Shows the agent the precise correction needed.
Guidance Feedback	User provides instructions or explanations to steer the agent.	Offers deeper context, reasoning, or demonstrations for learning complex behaviors.
Demonstrations	User shows the agent how to perform a task correctly.	Teaches specific procedures or desired interaction patterns.
Instructions / Critiques	User provides natural language explanations, critiques, or step-by-step guidance.	Helps the agent understand why an output is wrong and how to improve.
Implicit Feedback	Agent infers user preference from their behavior.	Reveals preferences and usability issues without explicit feedback requests.
Human Action / Control	Human directly takes actions and control.	Collaborates with humans to effectively finish tasks or learns from human actions.

Table 4: Human Feedback Types and Subtypes. The subtypes of evaluative feedback includes preference ranking, scalar rating, and binary assessment. The subtypes of corrective feedback includes the direct edits or refinement. The subtypes of guidance feedback includes the demonstration and instructions or critiques. The subtypes of implicit feedback include the human action or control.

powered agents, these systems are engineered with mechanisms to allow human supervision, intervention, and feedback (Mehta et al., 2024). This human-in-the-loop design is critical when balancing the strengths of LLMs—such as processing vast amounts of knowledge and performing rapid reasoning—with the need for contextual, ethical, and domain-specific judgments that humans uniquely provide (Vats et al., 2024).

Furthermore, multi-agent systems often assume that the collaboration among agents can lead to a form of "collective intelligence" where agents work toward shared objectives (Sun et al., 2024b). In many such frameworks, the communication protocols, coordination strategies, and role dynamics are all defined among non-human entities. In contrast, in human–agent systems, the interaction protocols are designed to enhance transparency and provide control for human decision-makers (Shao et al., 2024). The system can selectively escalate issues for human review, enable corrective actions when the automated decision may be off-mark, and integrate human feedback to iteratively improve the agent's performance over time (Mehta et al., 2024).

G Tables

Table 5 catalogs the environmental configuration and human feedback type, and Table 6 categorizes the interaction, orchestration, and communication of the current works, respectively.

Table 5: ① Environment Configuration and ② Human Feedback to LLM-based agents in human–agent systems. Environment Configuration specifies whether a single or multiple humans collaborate with one or more LLM-based agents, while Human Feedback characterizes the type, subtype, granularity, and interaction phase of the human feedback to the LLM-based agents.

			Environment Configuration		**Human Feedback to LLM-based Agent**			
Paper	***Venue***	**Code/ Data**	**Human**	**LLM Agent**	**Type**	**Subtype**	**Granularity**	**Phase**
Collaborative Gym (Shao et al., 2024)	*Arxiv'24*	Link	Single	Single	Corrective, Guidance	Refinements, Instructions	Segment	During Task
MTOM (Zhang et al., 2024b)	*Arxiv'24*	–	Single	Single	Implicit	Human Action	Segment	During Task
FineArena (Xu et al., 2025)	*Arxiv'25*	–	Single	Multiple	Guidance	Demonstrations	Segment, Holistic	Initial Setup, During Task
Prison Dilemm (Jiang et al., 2025)	*Arxiv'25*	–	Single	Single	Implicit	Human Action	Segment	During Task
InteractGen (Sun et al., 2024b)	*THU'24*	–	Multiple	Multiple	Guidance	Demonstrations	Segment	Initial Setup, During Task
AI Chains (Wu et al., 2022)	*CHI'24*	–	Single	Single	Corrective	Refinements	Segment	During Task
Drive As You Speak (Cui et al., 2024)	*WACV'24*	–	Single	Single	Guidance	Demonstrations	Holistic	Initial Setup
AgentCoord (Pan et al., 2024a)	*Arxiv'24*	Link	Single	Multiple	Evaluative, Corrective	Preference Ranking, Refinements	Segment, Holistic	Initial Setup, During Task
CowPilot (Huq et al., 2025)	*Arxiv'25*	Link	Single	Single	Corrective, Evaluative	Binary Assessment, Refinements	Segment	During Task
EasyLAN (Pan et al., 2024b)	*Arxiv'24*	–	Single	Multiple	Corrective, Guidance	Demonstrations, Refinements	Segment, Holistic	During Task
Hierarchical Agent (Liu et al., 2023b)	*AAMAS'24*	–	Single	Multiple	Guidance	Demonstrations	Segment	During Task
SWEET-RL (Zhou et al., 2025)	*Arxiv'25*	Link	Single	Single	Corrective, Implicit	Refinements, Human Action	Segment	Initial Setup, During Task
HRC Assembly (Gkournelos et al., 2024)	*CIRP'24*	–	Single	Multiple	Guidance	Demonstrations	Segment	During Task
REVECA (Seo et al., 2025)	*Arxiv'24*	–	Single	Multiple	Guidance	Demonstrations	Holistic	Initial Setup
AssistantX (Sun et al., 2024a)	*Arxiv'24*	Link	Multiple	Multiple	Implicit, Guidance	Human Action, Demonstrations	Holistic, Segment	Initial Setup, During Task
MINT (Wang et al., 2024b)	*ICLR'24*	Link	Multiple	Single	Evaluative, Corrective, Guidance	Binary Assessment, Refinements, Instructions	Holistic	During Task
Help Feedback (Mehta et al., 2024)	*EACL'24*	–	Single	Single	Evaluative, Guidance	Demonstrations, Instructions, Binary Assessment	Holistic, Segment	During Task
ConvCodeWorld (Han et al., 2025)	*ICLR'25*	Link	Single	Single	Guidance, Evaluative	Demonstrations, Instructions, Binary Assessment	Segment, Holistic	During Task
ReHAC (Feng et al., 2024)	*ACL'24*	Link	Single	Single	Corrective	Refinements	Segment	During Task
DPT Agent (Zhang et al., 2025)	*Arxiv'25*	Link	Single	Single	Guidance	Instructions	Holistic	During Task
HRC Manipulation (Liu et al., 2023a)	*IEEE'23*	–	Single	Single	Corrective, Guidance	Demonstrations, Refinements	Segment	During Task
HRC DMP (Liu et al., 2024a)	*IEEE'24*	–	Single	Single	Corrective, Guidance	Refinements, Demonstrations	Segment	During Task
PARTNR (Chang et al., 2024)	*ICLR'25*	Link	Single	Single	Guidance	Demonstrations	Holistic	Inital Setup
Organized Teams (Guo et al., 2024b)	*Arxiv'24*	Link	Single	Multiple	Guidance	Demonstrations	Holistic, Segment	Initial Setup, During Task
CoELA (Zhang et al., 2024a)	*ICLR'23*	–	Single	Multiple	Guidance	Demonstrations	Holistic, Segment	Initial Setup, During Task
Agency Task (Sharma et al., 2024)	*EACL'24*	Link	Single	Single	Guidance	Demonstrations	Segment	During Task
GDfC (Wang et al., 2025c)	*SME'25*	–	Single	Multiple	Guidance, Evaluative	Demonstrations, Binary Assessment, Preference Ranking	Holistic, Segment	Initial Setup, During Task, Post Task
PDFChatAnnotator (Tang et al., 2024)	*IUI'24*	–	Single	Single	Corrective, Guidance	Demonstrations, Refinements	Segment	During Task
Attentive Supp. (Tanneberg et al., 2024a)	*IEEE'24*	Link	Multiple	Single	Implicit, Guidance	Demonstrations, Human Action	Segment	During Task
HRC Trust (Ye et al., 2023)	*IEEE'23*	–	Single	Single	Guidance	Demonstrations, Instructions	Segment	During Task
BPMN (Ait et al., 2024)	*Arxiv'24*	Link	Multiple	Multiple	Guidance, Corrective	Instructions, Refinements	Segment	During Task, Post Task
Co-STORM (Jiang et al., 2024)	*EMNLP'24*	Link	Single	Multiple	Guidance	Demonstrations	Segment	During Task
HRC Manufa. (Lim et al., 2024)	*IEEE'24*	–	Single	Single	Corrective, Guidance	Demonstrations, Refinements, Instructions	Segment	Initial Setup, During Task
A2C (Tariq et al., 2025)	*Arxiv'24*	Link	Multiple	Multiple	Guidance, Evaluative	Binary Assessment, Instructions	Holistic, Segment	During Task
MindAgent (Gong et al., 2023)	*NAACL'24*	Link	Single	Multiple	Guidance	Demonstrations	Segment	During Task
Ask Before Plan (Zhang et al., 2024c)	*EMNLP'24*	Link	Single	Multiple	Guidance	Demonstrations	Segment	Initial Setup, During Task
SOTOPIA (Zhou et al., 2024)	*ICLR'24*	–	Multiple	Multiple	Evaluative, Implicit	Scaler Rating, Human Action	Holistic, Segment	During Task, Post Task
PaLM-E (Driess et al., 2023)	*ICML'23*	Link	Single	Single	Guidance, Implicit	Demonstrations, Human Action	Holistic, Segment	Initial Setup, During Task
TaPA (Wu et al., 2023)	*Arxiv'23*	Link	Single	Single	Guidance	Demonstrations	Holistic, Segment	Initial Setup
MetaGPT (Hong et al., 2023)	*ICLR'24*	Link	Single	Multiple	Guidance	Demonstrations	Holistic	Initial Setup
DigiRL (Bai et al., 2024)	*NeurIPS'24*	Link	Single	Single	Evaluative, Guidance	Binary Assessment, Demonstrations	Holistic	During Task, Post Task
WebLINX (Lù et al., 2024)	*Arxiv'24*	Link	Single	Multiple	Guidance	Demonstrations	Holistic, Segment	Initial Setup, During Task
MineWorld (Guo et al., 2025)	*Arxiv'25*	Link	Multiple	Single	Implicit	Human Action	Segment	During Task
M3HF (Wang et al., 2025d)	*ICML'25*	–	Multiple	Multiple	Evaluative, Guidance	Binary Assessment, Instructions	Segment, Holistic	During Task, Post Task
SMALL (Wang et al., 2024c)	*Arxiv'24*	–	Single	Multiple	Guidance	Instructions	Segment	Initial Setup
MAIH (Wang et al., 2024c)	*Arxiv'24*	–	Single	Single	Implicit	Human Action	Segment	During Task, Post Task
HRT-ML (Liu et al., 2024b)	*Arxiv'24*	–	Single	Multiple	Corrective, Guidance, Implicit	Refinements, Instructions, Human Action	Holistic	Initial Setup, During Task, Post Task
AXIS (Lu et al., 2024)	*Arxiv'25*	–	Single	Multiple	Evaluative, Implicit, Corrective	Human Action, Refinements, Binary Assessment	Holistic	Initial Setup, During Task, Post Task
EmoAgent (Qiu et al., 2025)	*Arxiv'25*	–	Single	Multiple	Corrective, Implicit, Guidance	Human Action, Instructions, Binary Assessment	Segment, Holistic	During Task, Post Task
SymbioticRAG (Sun et al., 2025)	*Arxiv'25*	–	Single	Single	Corrective, Implicit, Evaluative	Binary Assessment, Refinements, Demonstrations, Instructions, Human Action	Segment	Initial Setup, During Task, Post Task

Table 6: ① Interaction ② Orchestration ③ Communication in LLM-based human–agent systems. Interaction types capture the human and agent collaboration type; Orchestration covers task strategy and temporal synchronization; Communication describes how messages are structured and delivered in the system.

			Interaction		**Orchestration**		**Communication**	
Paper	***Venue***	**Code/ Data**	**Types**	**Variant**	**Strategy**	**Synchronization**	**Structure**	**Mode**
Collaborative Gym (Shao et al., 2024)	*Arxiv'24*	Link	Collaboration	Cooperation, Delegation	One-by-One	Asynchronous	Decentralized	Conversation
MTOM (Zhang et al., 2024b)	*Arxiv'24*	–	Collaboration	Coordination, Cooperation	Simultaneous	Synchronous	Decentralized	Conversation
FineArena (Xu et al., 2025)	*Arxiv'25*	–	Collaboration	Delegation, Cooperation	One-by-One	Synchronous	Hierarchical	Conversation
Prison Dilemm (Jiang et al., 2025)	*Arxiv'25*	–	Coopetition	–	One-by-One	Asynchronous	Decentralized	Conversation
InteractGen (Sun et al., 2024b)	*THU'24*	–	Collaboration	Cooperation, Delegation, Coordination	One-by-One	Asynchronous	Decentralized	Message Pool
AI Chains (Wu et al., 2022)	*CHI'24*	–	Collaboration	Cooperation	One-by-One	Synchronous	Decentralized	Conversation
Drive As You Speak (Cui et al., 2024)	*WACV'24*	–	Collaboration	Delegation	One-by-One	Synchronous	Centralized	Conversation
AgentCoord (Pan et al., 2024a)	*Arxiv'24*	Link	Collaboration	Coordination	One-by-One	Synchronous	Hierarchical	Conversation
CowPilot (Huq et al., 2025)	*Arxiv'25*	Link	Collaboration	Supervision, Delegation, Coordination	One-by-One	Synchronous	Decentralized	Conversation
EasyLAN (Pan et al., 2024b)	*Arxiv'24*	–	Collaboration	Delegation, Supervision	One-by-One	Synchronous	Hierarchical	Observation
Hierarchical Agent (Liu et al., 2023b)	*AAMAS'24*	–	Collaboration	Supervision, Delegation, Cooperation	One-by-One	Synchronous	Hierarchical	Conversation
SWEET-RL (Zhou et al., 2025)	*Arxiv'25*	Link	Collaboration	Delegation	One-by-One	Synchronous	Centralized	Conversation
HRC Assembly (Gkournelos et al., 2024)	*CIRP'24*	–	Collaboration	Delegation, Cooperation	One-by-One	Synchronous	Decentralized	Conversation
REVECA (Seo et al., 2025)	*Arxiv'24*	–	Collaboration	Cooperation	One-by-One	Synchronous	Decentralized	Conversation
AssistantX (Sun et al., 2024a)	*Arxiv'24*	Link	Collaboration	Delegation, Cooperation	One-by-One	Asynchronous	Decentralized	Message Pool
MINT (Wang et al., 2024b)	*ICLR'24*	Link	Collaboration	Delegation, Cooperation	One-by-One	Synchronous	Decentralized	Conversation
Help Feedback (Mehta et al., 2024)	*EACL'24*	–	Collaboration	Supervision, Delegation, Cooperation	One-by-One	Asynchronous	Decentralized	Conversation
ConvCodeWorld (Han et al., 2025)	*ICLR'25*	Link	Collaboration	Supervision, Delegation	One-by-One	Asynchronous	Decentralized	Conversation
ReHAC (Feng et al., 2024)	*ACL'24*	Link	Collaboration	Coordination, Supervision	One-by-One	Synchronous	Decentralized	Conversation
DPT Agent (Zhang et al., 2025)	*Arxiv'25*	Link	Collaboration	Coordination	Simultaneous	Asynchronous	Decentralized	Observation
HRC Manipulation (Liu et al., 2023a)	*IEEE'23*	–	Collaboration	Supervision, Delegation	One-by-One	Synchronous	Decentralized	Conversation
HRC DMP (Liu et al., 2024a)	*IEEE'24*	–	Collaboration	Delegation, Supervision	One-by-One	Synchronous	Decentralized	Conversation
PARTNR (Chang et al., 2024)	*ICLR'25*	Link	Collaboration	Coordination, Cooperation	Simultaneous	Synchronous	Decentralized, Centralized	Observation
Organized Teams (Guo et al., 2024b)	*Arxiv'24*	Link	Collaboration	Cooperation, Coordination	One-by-One	Synchronous	Decentralized, Centralized, Hierarchical	Conversation
CoELA (Zhang et al., 2024a)	*ICLR'23*	–	Collaboration	Cooperation, Coordination	Simultaneous	Synchronous	Decentralized	Conversation
Agency Task (Sharma et al., 2024)	*EACL'24*	Link	Collaboration	Cooperation, Delegation	One-by-One	Synchronous	Decentralized	Conversation
GDfC (Wang et al., 2025c)	*SME'25*	–	Collaboration	Delegation	One-by-One	Synchronous	Decentralized	Conversation
PDFChatAnnotator (Tang et al., 2024)	*IUI'24*	–	Collaboration	Delegation	One-by-One	Synchronous	Decentralized	Conversation
Attentive Supp. (Tanneberg et al., 2024a)	*IEEE'24*	Link	Collaboration	Coordination	One-by-One	Synchronous	Decentralized	Observation
HRC Trust (Ye et al., 2023)	*IEEE'23*	–	Collaboration	Delegation	One-by-One	Synchronous	Decentralized	Conversation
BPMN (Ait et al., 2024)	*Arxiv'24*	Link	Collaboration	Coordination	Simultaneous	Asynchronous	Decentralized	Message Pool
Co-STORM (Jiang et al., 2024)	*EMNLP'24*	Link	Collaboration	Coordination	One-by-One	Synchronous	Centralized	Conversation
HRC Manufa. (Lim et al., 2024)	*IEEE'24*	–	Collaboration	Delegation, Cooperation	One-by-One	Synchronous	Centralized	Conversation
A2C (Tariq et al., 2025)	*Arxiv'24*	Link	Collaboration	Cooperation	One-by-One	Asynchronous	Hierarchical	Conversation
MindAgent (Gong et al., 2023)	*NAACL'24*	Link	Collaboration	Coordination	Simultaneous	Synchronous	Centralized	Conversation
Ask Before Plan (Zhang et al., 2024c)	*EMNLP'24*	Link	Collaboration	Coordination, Delegation	One-by-One	Synchronous	Hierarchical	Conversation
SOTOPIA (Zhou et al., 2024)	*ICLR'24*	–	Collaboration, Competition, Coopetition	Coordination, Cooperation	One-by-One	Synchronous	Decentralized	Conversation
PaLM-E (Driess et al., 2023)	*ICML'23*	Link	Collaboration	Delegation	One-by-One	Synchronous	Decentralized	Conversation
TaPA (Wu et al., 2023)	*Arxiv'23*	Link	Collaboration	Delegation	One-by-One	Asynchronous	Decentralized	Conversation
MetaGPT (Hong et al., 2023)	*ICLR'24*	Link	Collaboration	Coordination	One-by-One	Asynchronous	Decentralized	Message Pool
DigiRL (Bai et al., 2024)	*NeurIPS'24*	Link	Collaboration	Delegation	One-by-One	Synchronous	Centralized	Conversation
WebLINX (Lù et al., 2024)	*Arxiv'24*	Link	Collaboration	Delegation	One-by-One	Synchronous	Hierarchical	Conversation
MineWorld (Guo et al., 2025)	*Arxiv'25*	Link	Collaboration	Delegation	One-by-One	Synchronous	Decentralized	Observation
M3HF (Wang et al., 2025d)	*ICML'25*	–	Collaboration	Cooperation	One-by-One, Simultaneous	Synchronous	Centralized	Message Pool
SMALL (Wang et al., 2024c)	*Arxiv'24*	–	Collaboration	Delegation	One-by-One	Asynchronous	Hierarchical	Message Pool
MAIH (Wang et al., 2024c)	*Arxiv'24*	–	Collaboration	Delegation, Cooperation, Coordination	One-by-One, Simultaneous	Asynchronous	Decentralized, Hierarchical	Message Pool
HRT-ML (Liu et al., 2024b)	*Arxiv'24*	–	Collaboration	Coordination, Cooperation	One-by-One, Simultaneous	Asynchronous, Synchronous	Hierarchical, Centralized	Message Pool, Conversation
AXIS (Lu et al., 2024)	*Arxiv'25*	–	Collaboration	Delegation	One-by-One	Synchronous	Centralized	Conversation, Observation
EmoAgent (Qiu et al., 2025)	*Arxiv'25*	–	Collaboration	Supervision, Coordination, Cooperation	One-by-One	Synchronous	Hierarchical, Centralized	Conversation, Observation
SymbioticRAG (Sun et al., 2025)	*Arxiv'25*	–	Collaboration	Cooperation, Supervision, Delegation	One-by-One	Synchronous	Centralized	Conversation