Toward Understanding Misalignment in LLM Agents: A Survey of Taxonomy, Causes, Mitigation, and Evaluation

Anonymous ACL submission

Abstract

Large language models are increasingly deployed as autonomous agents that interact with external environments to solve long-horizon tasks. As these systems move toward real-world deployment, misalignment in LLM agents has emerged as a distinct challenge. Unlike misalignment in stand-alone LLMs, it unfolds through multi-step interaction and may arise from mismatches with the environment, failures within the agent's own decision state, or breakdowns in inter-agent coordination. However, existing literature remains fragmented across agent-focused surveys and stand-alone LLM studies, leaving no unified account of misalignment in agentic systems. In this survey, we present a closed-loop perspective that frames misalignment in LLM agents as a systems-level problem shaped by interaction among the agent, the environment, and other agents. We organize the literature into three categories—agent–environment, intra-agent, and inter-agent misalignment—and review their causes, mitigation strategies, and evaluation practices. Finally, we highlight open challenges and future directions for building more reliable, robust, and trustworthy LLM-agent systems.

1 Introduction

Large language models (LLMs) are rapidly evolving from passive text generators into LLM agents that maintain context, access memory, invoke tools, and interact with external environments to solve long-horizon tasks (Yao et al., 2022; Park et al., 2023; Wang et al., 2024; Su et al., 2024a; Fang et al., 2025). This transition has broadened LLM applications from single-turn assistance to complex multi-step settings such as web automation (Zhou et al., 2024b; Ning et al., 2025), software engineering (Liu et al., 2024a; Novikov et al., 2025), scientific discovery (Ren et al., 2025; Lyu et al., 2026), embodied decision making (Xie et al., 2024; Liu et al., 2025b), and multi-agent collaboration (Du et al., 2024; Jin et al., 2025). With this growing autonomy, ensuring reliable and consistent behavior over extended interactions has become an increasingly important challenge (Yu et al., 2025).

As these systems move toward real-world deployment, agentic misalignment emerges as a distinct and critical problem (Lynch et al., 2025; Naik et al., 2025; Wang et al., 2026b). Broadly, agentic misalignment refers to situations in which the behavior of an LLM agent, or multiple interacting LLM agents, deviates from task goals, human intentions, internal consistency requirements, or environmental constraints over the course of multi-step interaction. Such misalignment may arise even when each local decision appears reasonable, because at every step the agent must interpret observations, update its internal state, choose actions, and respond to feedback (Lin et al., 2025), and errors in any of these processes can steer subsequent behavior away from what is intended (Yu et al., 2025). For example, an agent may rely on stale context, form misleading internal state, invoke unsuitable tools, misinterpret environmental signals, or coordinate poorly with other agents (Zhang et al., 2025c). Agentic misalignment is therefore not simply a matter of incorrect single outputs, but a systems-level challenge shaped by recurrent interaction and evolving state.

Existing literature fails to systematically characterize misalignment in LLM agents. On one hand, surveys on LLM agents mainly focused on components like architectures (Wang et al., 2024), planning (Huang et al., 2024), memory (Hu et al., 2025b), tool use (Jiang et al., 2025) and multi-agent collaboration (Khan et al., 2024). However, these studies are largely organized around what agents can do and how well they achieve that, rather than the misalignment dynamics that determine whether they remain reliable over long-horizon scenarios. On the other hand, related work on stand-alone

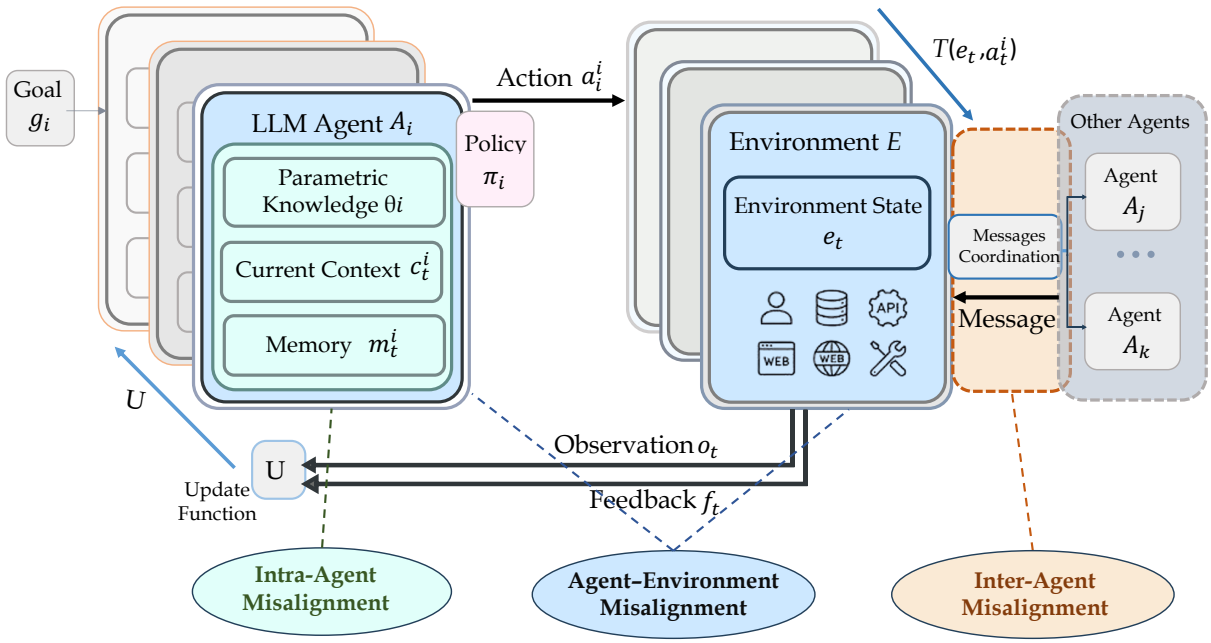

Figure 1: Three types of agentic misalignment between LLM agent, environment and other agents.

LLMs have examined misalignment-related phenomena such as knowledge conflicts (Xu et al., 2024), hallucination (Huang et al., 2025), and knowledge boundaries (Li et al., 2025a). However, misalignment in agentic systems is fundamentally more challenging because of its nature of iterative interaction with the environment. Consequently, the field still lacks a unified framework for understanding where misalignment arises in LLM agents, how it propagates across interaction trajectories, and how to mitigate such misalignment behavior.

In this survey, we address this gap by studying agentic misalignment through the closed-loop interaction among the agent, the environment, and other agents. We begin by formalizing LLM agents (Sec. 2) and defining key concepts. Then, we organize the literature into three categories: **agent-environment misalignment** (Sec. 3), which concerns failures at the interface between agents and the external world; **intra-agent misalignment** (Sec. 4), which concerns inconsistencies within the agent's own internal components; and **inter-agent misalignment** (Sec. 5), which concerns failures arising from communication, coordination, and heterogeneity in multi-agent setting. This taxonomy provides a unified lens for analyzing the origins, manifestations, mitigations or utilizations, and evaluations of agentic misalignment.

Our contributions are summarized as follows: i) We present a unified closed-loop perspective on misalignment in LLM agents, viewing it as a systems-level problem that emerges through iterative interaction among the agent, the environment, and other agents. ii) We propose a structured taxonomy that organizes the literature into three major categories: agent-environment misalignment, intra-agent misalignment, and inter-agent misalignment. iii) We systematically synthesize existing work within each category, covering its causes, mitigation and evaluation methods, thereby providing a comprehensive overview of the current research landscape. iv) We highlight open problems and future research directions that are critical for building more robust and trustworthy agentic systems.

2 Preliminaries

To ground the discussion in the remainder of this survey, we first define *LLM Agent*, the *environment* in which they operate, and the *interaction dynamics* between them. Based on this abstraction, we then introduce a unified notion of *misalignment* and later specialize it to three loci: agent-environment, intra-agent, and inter-agent settings with a unified view of the mitigation strategies.

2.1 Agent-Environment Formulation

LLM Agent. An *LLM agent* is a language-model-based system that pursues a goal through iterative, multi-step decision making. Unlike one-shot LLM usage, an agent continuously incorporates new information, takes actions that may affect the external world, and maintains state across steps. For-

mally, an agent $\mathcal{A}$ pursues a goal g, and at each step t selects an action a_t under policy π based on three sources: parametric knowledge θ (encoded in model weights), current context c_t (information available in the context window), and memory m_t (cross-step information such as prior observations, intermediate reasoning, or interaction history).

Environment. We define the environment ($\mathcal{E}$) as the external system with which the agent interacts throughout the decision process. At step t, the environment is in state e_t, receives an action a_t, and returns to the agent an observation o_t and feedback f_t reflecting the consequence of that action. The environment may include task interfaces, users, tools, databases, webpages, or other external information sources. Importantly, the agent typically does not observe the full environment state e_t directly, but instead reasons over partial, delayed, or otherwise mediated observations and feedback.

Interaction Dynamics. The interaction between an LLM agent and its environment can be viewed as a closed-loop sequential decision-making process. At each step t, the agent receives an observation o_t, combines it with its current context c_t and memory m_t, and selects an action a_t according to: $a_t \sim \pi(a \mid g, c_t, m_t; \theta)$. The executed action induces a new state and feedback from the environment in the formula: $e_{t+1} \sim \mathcal{T}(e_t, a_t)$ and $f_t \sim \Phi(e_t, a_t)$ where $\mathcal{T}$ denotes the environment transition function and Φ is the feedback function. The resulting observation and feedback are then incorporated into the agent's information for the next step: $(c_{t+1}, m_{t+1}) \sim \mathcal{U}(c_t, m_t, o_t, a_t, f_t)$, where $\mathcal{U}$ denotes the agent update function, which may append new context, revise memory or compress prior interaction history. It highlights that agent behavior is shaped by the interaction between internal information sources (e.g., θ, c_t, and m_t) and external signals from $\mathcal{E}$. T, misalignment may arise from mismatches among internal information, decision policy, and environmental signals.

2.2 Misalignment Types and Mitigation

We use *misalignment* to denote a mismatch among the goals, information, states, or decision processes that must remain compatible for effective agent behavior. We distinguish three loci based on where such mismatches arise: *(a) Agent-Environment Misalignment* arises when the agent fails to remain aligned with the external environment in which it operates. *(b) Intra-Agent Misalignment* arises when a single agent fails to maintain coherence among its internal information or decision processes. *(c) Inter-Agent Misalignment* arises when multiple agents fail to maintain compatible goals, beliefs or communication during joint or related tasks. Under this view, mitigation methods can be summarized in a unified form according to their primary intervention mechanism, including prompt-based, supervised-fine-tuning (SFT)-based, inference-based, Reinforcement Learning (RL)-based, and architecture-based approaches as demonstrated in Table **??**. This unified lens provides a common template for organizing prior work, while the later sections will examine how each misalignment type gives rise to its own specialized mitigation methods.

3 Agent-Environment Misalignment

As illustrated in Fig. 1, an LLM agent acts on the environment through its policy π, which maps goal g, parametric knowledge θ, current context c_t, and memory m_t to an action a_t. The environment then evolves according to its state e_t, transition function $T(e_t, a_t)$, and feedback function $\Phi(e_t, a_t)$, returning observation o_t and feedback f_t to the agent, which are incorporated through the update module $U(c_t, m_t, o_t, a_t, f_t)$. Agent–environment misalignment arises when these cross-boundary interactions do not match: the agent acts on incorrect assumptions about the environment, or it incorrectly interprets the signals the environment returns.

3.1 Underlying Causes

Action: Transition Mismatch Transition mismatch occurs when the agent's policy π predicts action outcomes that do not match the environment's true transition function $T(e_t, a_t)$. It arises from two sources: deployment shift, where latency, noise, or physical constraints alter the real effect of actions (Yamagata and Santos-Rodriguez, 2024; Lu et al., 2025b; Jonnarth et al., 2025); and partial observability, where the agent plans under incorrect assumptions about hidden state, delayed effects, or rare events. In agentic settings, this manifests as tool misuse, brittle plans, or optimization toward proxy behaviors (Chen et al., 2026a; Wang et al., 2026c).

Observation: Update Mismatch. Update mismatch occurs after action execution, when observations o_t and feedback f_t are too sparse or ambiguous for the update module $U(\cdot)$ to correctly

revise the agent's context and memory. It arises from two main sources. First, generic error signals prevent the agent from diagnosing whether a failure is caused by an invalid action or a missing precondition (Lumer et al., 2025; Lu et al., 2025a; Wang et al., 2025a; Lu et al., 2026). Second, unstated environmental dependencies—such as navigation requirements or tool configuration steps—are often not explicitly exposed to the agent (Udeshi et al., 2025; Zhou et al., 2026). In agentic settings, this leads the agent to update its internal state based on a distorted interpretation of the environment, causing it to persist with an invalid strategy.

3.2 Mitigation Strategies

Exposing Transition Rules and Failure Signals. One direct solution is to make the environment-facing interface more informative. Methods such as ALIGN augment the interaction loop by surfacing hidden constraints before acting and by converting vague outcomes into more diagnostic feedback after each step (Liu et al., 2025a). Related approaches similarly expose valid actions, state information, or execution traces through wrappers and auxiliary interfaces (Yang et al., 2023; Chen et al., 2024b; He et al., 2025; Sun et al., 2025; Zhou et al., 2024c). In terms of Fig. 1, these methods improve the quality of o_t and f_t, reducing mismatch between environmental outcomes and agent updates.

Grounding Policy and Update Through Interaction Trajectories. A second direction is to adapt the agent itself from environment-specific trajectories. Reinforcement learning, meta-learning, and exploration-based methods help the policy π and update process $U(\cdot)$ better reflect actual action consequences, failure patterns, and recovery behaviors observed during interaction (Qian and Liu, 2025; Zhang et al., 2025a; Mai et al., 2025; Luo et al., 2025b). Rather than generic adaptive training, the goal here is specifically to improve how agents choose actions and revise internal state from environment feedback.

Learning Explicit Environment Models for Planning. Another line of work reduces mismatch by making the agent's internal model of $T(e_t, a_t)$ and $\Phi(e_t, a_t)$ more explicit. WALL-E learns symbolic rules from observed trajectories, RWML learns action-conditioned world models, and RAP uses LLMs as predictive simulators for multi-step planning (Zhou et al., 2024c; Yu et al., 2026; Hao et al., 2023). These approaches directly target the cross-boundary prediction problem in Fig. 1: the agent should anticipate how the environment will transition before committing to an action.

Standardizing Agent-Environment Protocols. A complementary solution is to reduce ambiguity at the environment side. Formal languages such as RLang make preconditions and effects explicit, while Gym-style APIs and newer systems such as OpenEnv standardize action–observation protocols across tasks (Rodriguez-Sanchez et al., 2023; Towers et al., 2025; Brockman et al., 2016). Shared state representations, such as temporal knowledge graphs in Graphiti MCP, further help agents maintain a consistent view of environmental state across interactions (Rasmussen et al., 2025). These efforts reduce the need for the agent to infer basic environment semantics from weak signals alone.

3.3 Evaluation Frameworks

Existing relevant evaluation approaches can be broadly grouped into three categories: distributional generalization, interaction quality, and execution correctness. For distributional generalization, many works compare in-distribution (ID) and out-of-distribution (OOD) performance, where large gaps indicate poor robustness under changing environment dynamics. For example, RWML (Yu et al., 2026) is evaluated on both splits of ALFWorld (Shridhar et al., 2021), showing improved transfer through self-supervised world-model learning. For interaction quality, metrics such as invalid-action rates and interface reliance (IR) (Zhou et al., 2024a) capture whether agents depend on interface shortcuts rather than genuine understanding (Gu et al., 2026). For execution correctness in tool-using agents, benchmarks (Barres et al., 2025) show that failures often stem from malformed arguments or incorrect call ordering rather than wrong tool choice, making tool-selection accuracy, argument validity, and execution correctness key evaluation dimensions.

4 Intra-Agent Misalignment

Intra-agent misalignment refers to mismatches within the decision state of a single LLM agent. Concretely, it arises when the agent's parametric knowledge θ, current context c_t, memory m_t, the update process $\mathcal{U}$, and the resulting policy $\pi(a \mid g, c_t, m_t; \theta)$ fail to remain mutually compatible during inference and action selection. Unlike agent–environment misalignment, which is defined

at the boundary between the agent and the external world, intra-agent misalignment concerns inconsistencies that are represented and must be resolved inside the agent itself. This paper organizes intra-agent misalignment into three forms: intrinsic misalignment, intrinsic-extrinsic misalignment, and extrinsic misalignment.

4.1 Underlying Causes

Intrinsic Misalignment. Intrinsic misalignment is generated primarily by the agent's intrinsic substrate, namely the parametric knowledge θ and the internal inference dynamics that instantiate the policy π. Typical cases include unstable outputs under the same input (Wang et al., 2022; **?**; Chen et al., 2026b), latent uncertainty (Xiong et al., 2024) among competing continuations, reasoning drift over multi-step inference (Zhang et al., 2026c), and divergence between high-level planning and downstream action (Erdogan et al., 2025). In these cases, even when c_t, m_t, and the update process $\mathcal{U}$ are held fixed, the induced policy may remain unstable or internally inconsistent. The core issue is therefore misalignment within what the model encodes and how it reasons from that encoding.

Intrinsic-Extrinsic Misalignment. Intrinsic-extrinsic misalignment arises when the agent's intrinsic knowledge θ conflicts with runtime information maintained in c_t or m_t. Typical cases include disagreement between parametric knowledge and newly retrieved evidence, conflicts between prior beliefs and updated memory, or failures to decide whether information newly written into c_t or m_t should override what is already encoded in θ (Xu et al., 2024; Kortukov et al., 2024; Huo et al., 2025; Li et al., 2025c). This type of misalignment is closely tied to the update process $\mathcal{U}$, since the conflict often emerges when new observations or feedback are incorporated into the run-time state but are not reconciled properly with prior parametric knowledge.

Extrinsic Misalignment. Extrinsic misalignment arises when the runtime substrate itself, namely c_t and m_t, becomes unreliable after incorporating external information. This could include scenarios like contradictory context snippets, stale retrieved passages, noisy user instructions and memory write or revision errors (Liu et al., 2024b; Cheng et al., 2025; Hu et al., 2025a). This type of failure is most directly tied to the update process $\mathcal{U}(c_t, m_t, o_t, a_t, f_t)$, where the resulting context c_t or memory m_t may become corrupted, inconsistent, stale, or poorly synchronized with later reasoning. Here the source of the problem is external, but the mismatch is expressed inside the agent because subsequent decisions are conditioned on distorted context or memory.

4.2 Mitigation Strategies

Reasoning Stabilization. For intrinsic misalignment, the main mitigation objective is to stabilize how the policy π is induced from the agent's intrinsic substrate. Existing methods mainly follow three directions. Self-verification methods, such as CoVe (Dhuliawala et al., 2024) and CRITIC (Gou et al., 2023), insert explicit checking or critique steps to validate intermediate reasoning or action candidates before commitment. Reflective refinement methods like Reflexion (Shinn et al., 2023) and Self-Refine (Madaan et al., 2023) iteratively use self-generated feedback to revise reasoning traces, memory, or outputs, reducing local drift. Plan-aware stabilization methods, such as Reflective Test-Time Planning (Hong et al., 2026) and EvoAgent (Feng et al., 2025), combine reflection with explicit planning or world-model guidance to keep action selection aligned with longer-horizon goals. Overall, these methods aim to regularize the mapping from intrinsic knowledge and memory to action, so that repeated or long-horizon inference yields a more coherent policy trajectory.

Knowledge Reconciliation. For intrinsic-extrinsic misalignment, the key mitigation challenge is to reconcile parametric knowledge θ with the run-time state (c_t, m_t). A common strategy is to explicitly detect and arbitrate conflicts between internal knowledge and newly incorporated evidence. Self-RAG (Asai et al., 2023), CRAG (Yan et al., 2024), and CoV-RAG (He et al., 2024) improve this process by combining retrieval with critique and corrective retrieval. More conflict-centered methods such as FaithfulRAG (Zhang et al., 2025b) and Micro-Act (Huo et al., 2025) go further by modeling disagreements between parametric and retrieved knowledge and resolving them through structured self-reasoning. Overall, these methods improve how update function $\mathcal{U}$ incorporates new observations and feedback into context and memory.

Run-Time State Governance. For extrinsic misalignment, the main mitigation objective is to govern the construction and evolution of the run-time

state itself, namely c_t and m_t, through the update process $\mathcal{U}$. Existing methods can be broadly viewed as memory architecture, memory management, and memory admission/control mechanisms. MemGPT (Packer et al., 2023), MemoryOS (Kang et al., 2025), Mem0 (Chhikara et al., 2025), and MIRIX (Wang and Chen, 2025) provide more structured memory architectures for long-horizon agents. A-MEM (Xu et al., 2025) and Reflective Memory Management (Tan et al., 2025) improve how memories are linked, summarized, and revised over time. TierMem (Zhu et al., 2026a) and A-MAC (Zhang et al., 2026a) further target specific governance failures such as lossy summarization and unreliable memory admission.

4.3 Evaluation Frameworks

Evaluation of intra-agent misalignment remains fragmented across benchmark families rather than unified under a single framework. Intrinsic misalignment is typically assessed through long-horizon evaluation of policy stability, with benchmarks such as DeepPlanning (Zhang et al., 2026b) and UltraHorizon (Luo et al., 2025a) exposing failures in planning consistency, long-range coherence, and trajectory stability. Intrinsic-extrinsic misalignment is best evaluated by testing whether the agent can reconcile parametric knowledge θ with runtime information in c_t and m_t; representative benchmarks include ConflictBank (Su et al., 2024b) and MemoryAgentBench (Hu et al., 2025a), which focus on conflicts between encoded knowledge, retrieved knowledge, and their resolution. Extrinsic misalignment is best captured by benchmarks that stress the quality of the runtime state and the robustness of the update process $\mathcal{U}$ under noisy, stale, or evolving inputs, such as RealMem (Bian et al., 2026) and AgentNoiseBench (Wang et al., 2026a).

5 Inter-Agent Misalignment

Inter-agent misalignment refers to situations in which multiple agents, operating under shared or partially aligned goals, maintain incompatible informational states, reasoning trajectories, or final decisions during collaboration. Such misalignments originate from interaction among agents and arise from three structural asymmetries: differences in parametric knowledge θ induced by heterogeneous model backbones or role-conditioned prompts, differences in contextual information c_t from asymmetric access to external evidence and differences in interaction structure that influence how agents update their context and memory m_t during collaboration.

5.1 Underlying Causes

Model Heterogeneity and Role Conditioning. Differences in backbone architecture, pretraining, instruction tuning, or role-specific prompting constitute a fundamental source of inter-agent misalignment. Heterogeneous agents may encode distinct priors or privilege different decision criteria, producing persistent belief-level divergence that communication alone cannot resolve (Xiong et al., 2023; Wu and Ito, 2025; Malfa et al., 2025). Role-conditioned prompts further amplify this effect by selectively biasing attention and argumentative stance. Recent architectures such as Mixture-of-Agents explicitly exploit such heterogeneity to improve performance (Wang et al., 2025c), but complementary diversity also introduces representational misalignment as a structural byproduct.

Information Asymmetry. In retrieval augmented systems, agents accessing different evidence pools develop divergent beliefs even when model parameters are identical. Benchmarks such as ClashEval and Conflicting-RAG demonstrate that unresolved document contradictions systematically mislead collaborative reasoning (Wu et al., 2024; Wang et al., 2025b). Evidence selection thus becomes a structural driver of epistemic divergence, particularly when retrieval operates independently for each agent.

Interaction Protocol Design. Beyond agent and information differences, interaction protocol independently shapes collective outcomes. Debate-based, voting-based, or consensus-seeking designs impose different constraints on information exchange, causing identical agents to converge or diverge solely due to protocol structure (Choi et al., 2025; Chen et al., 2023). Aggregation rules determine whether minority-correct signals are preserved or suppressed, and systems lacking mechanisms for viewpoint diversity often collapse into majority voting dynamics (Zhu et al., 2026b; Li et al., 2026). Procedural design thus acts as a direct amplifier or dampener of inter-agent disagreement.

5.2 Mitigation and Utilization Strategies

Because inter-agent misalignment arises from structural asymmetries in collaborative systems, practical system design focuses not on eliminating it

entirely but on shaping how it influences collaborative reasoning. Recent research therefore explores two complementary directions: regulating harmful misalignment that destabilizes coordination or propagates incorrect assumptions, and leveraging the reasoning diversity it introduces to improve collective problem solving (Wu and Ito, 2025; Cemri et al., 2025; Malfa et al., 2025).

Mitigate Harmful Misalignment. A first line of mitigation targets the structural and epistemic sources of misalignment. At the architectural level, optimization frameworks such as MASS treat agent roles and communication topology as jointly tunable variables, demonstrating that co-optimizing these elements reduces persistent divergence across reasoning processes (Zhou et al., 2025). Structured reconciliation protocols further stabilize heterogeneous agent groups by encouraging explicit justification and cross-agent comparison (Chen et al., 2024a; Wu and Ito, 2025). At the epistemic level, contradiction-aware retrieval pipelines and provenance-sensitive reasoning protocols reduce the likelihood that inconsistent documents propagate through collaborative reasoning chains (Wang et al., 2025b; Wu et al., 2024). Hybrid mechanisms combine debate-style reasoning exchange with confidence-weighted voting and protocols that explicitly preserve minority signals (Choi et al., 2025; Kaesberg et al., 2025).

Utilize Misalignment for Exploration. Rather than suppressing all divergence, a complementary line of work treats reasoning heterogeneity as a resource. Mixture-of-Agents architectures organize heterogeneous LLMs into layered structures in which agents iteratively refine each other's outputs, exploiting complementary capabilities while requiring coordination mechanisms to prevent inconsistent beliefs from propagating (Wang et al., 2025c). Debate-based protocols regulate this process by encouraging agents to critique intermediate steps and refine hypotheses through iterative interaction, though effective collaboration requires both viewpoint diversity and calibrated confidence communication to avoid collapsing into majority voting (Du et al., 2024; Liang et al., 2024; Zhu et al., 2026b). Dynamic debate architectures address this by explicitly generating diverse reasoning paths and specialized agent roles to sustain exploration during deliberation (Li et al., 2026). Reversible reasoning further complements these approaches by introducing rollback mechanisms that allow agents to backtrack when intermediate reasoning becomes inconsistent (Zhao et al., 2025).

Adaptive Optimization Across Episodes. Beyond single-episode coordination, misalignment can serve as a diagnostic signal for long-term system improvement. Evaluation-in-the-loop frameworks treat coordination structures themselves as optimization targets, refining prompts, agent roles, and communication protocols based on observed failures. MASS exemplifies structural co-optimization across episodes (Zhou et al., 2025), while evolutionary frameworks such as AlphaEvolve iteratively refine solution strategies through generation–evaluation–selection cycles (Novikov et al., 2025). By systematically analyzing recurring failure modes, such approaches transform misalignment from a problem to eliminate into a driver of sustained coordination improvement (Cemri et al., 2025; Malfa et al., 2025).

5.3 Evaluation of Inter-Agent Misalignment

Evaluating inter-agent misalignment requires going beyond task accuracy to capture interaction dynamics. Recent work suggests that effective evaluation should consider coordination stability, reasoning consistency under conflicting evidence. Empirical studies show that different coordination protocols can yield substantially different outcomes even with identical agents, highlighting the importance of evaluating aggregation behavior and convergence dynamics (Radeva et al., 2025). Comparative analyses further reveal that multi-agent collaboration may introduce performance trade-offs due to coordination overhead and inconsistent information flow (Zhu et al., 2025). At the reasoning level, debate-based evaluation frameworks such as ChatEval explicitly model disagreement and reconciliation to better align with human judgments (Chan et al., 2024), while recent benchmarks extend evaluation to social dynamics such as consensus formation and role-induced bias (Reza, 2026).

6 Challenges and Future Directions

Attribution and Localization of Misalignment. A central open challenge is attribution: determining where misalignment originates and how it propagates through the agent loop. Recent work has begun to study failure attribution, error diagnosis, and modular debugging, but these efforts are still mostly post hoc and benchmark-specific. They rarely provide a unified way to localize misalign-

ment across perception, state update, policy formation, and coordination. Future work should therefore develop finer-grained attribution methods that can identify not only where failure appears, but also how local errors cascade into later reasoning and action.

Realistic and Process-Level Evaluation. Another major challenge is benchmark design. Existing evaluation studies repeatedly show that current agent benchmarks often overemphasize final success while under-measuring process quality, robustness, and recovery. Although recent benchmarks have started to cover long-horizon planning, memory evolution, and noisy interaction, each still captures only part of the problem. Future benchmarks should jointly evaluate long-horizon reasoning, partial observability, memory updates, perturbation robustness, and process-level quality so that misalignment can be studied as a dynamic problem.

Balancing Autonomy and Controllability. As LLM agents become more autonomous, a key challenge is to preserve controllability without overly constraining useful agent behavior. Existing work often improves autonomy through stronger planning, memory, or reflection, but these same mechanisms can also make agent behavior harder to predict, supervise, or redirect. This tension becomes especially important when agents update their own state, revise plans, or act over long horizons with limited human intervention. Future research should therefore study how to balance agent autonomy with explicit mechanisms for intervention, correction, and controllable behavior under realistic deployment settings.

Cross-Modal Misalignment in Multimodal Agents. An emerging challenge is cross-modal misalignment in multimodal agents. As agents increasingly combine language with visual, auditory, spatial, or embodied inputs, failures may arise when information across modalities is inconsistent, incomplete, or interpreted unevenly (Li et al., 2025b). Existing multimodal agent work has mostly focused on improving capability and task performance, while the question of how conflicts across modalities affect memory, reasoning, and action remains less explored. Future work should therefore investigate how multimodal evidence is fused, updated, and verified over time, and how modality-specific errors propagate into broader agent misalignment.

Scaling Misalignment Mitigation to Multi-Agent Systems. Another open problem is how to scale mitigation strategies from single-agent settings to multi-agent systems. Many current methods for reflection, memory management, and conflict resolution are designed for one agent with one decision state, but collaborative systems introduce additional challenges such as role mismatch, protocol ambiguity, and divergent local beliefs. As the number of interacting agents grows, local misalignment can also spread more quickly across communication and coordination channels. Future work should therefore study how mitigation mechanisms can be extended from individual reasoning and memory control to distributed settings with shared tasks, evolving roles, and heterogeneous agents.

Scalable Coordination under Human and Resource Constraints. A final challenge is to make alignment mechanisms effective under practical deployment constraints. Existing multi-agent and value-alignment work highlights that coordination quality depends not only on reasoning accuracy, but also on communication efficiency cost and latency, and social compatibility. At the same time, current evaluation still tends to emphasize technical performance more than human-centered, economic, or safety-related considerations. Future research should therefore move beyond asking whether an alignment method improves task success, and instead ask whether it remains controllable, cost-effective and socially compatible in realistic human-in-the-loop and resource-constrained deployments.

7 Conclusion

In this survey, we present a unified view of misalignment in LLM agents across agent–environment, intra-agent, and inter-agent settings. Grounded in a common formulation of LLM agents as sequential decision-making systems with parametric knowledge, run-time context, memory, policy, and state update, we organize the literature around how misalignment arises, propagates, is mitigated, and is evaluated. By bringing these previously fragmented lines of work under a shared framework, we highlight misalignment as a systems-level challenge for increasingly autonomous, interactive, and collaborative agents. We hope this survey provides a useful foundation for future research on building more reliable, coherent, and controllable LLM-agent systems.

8 Limitations

This survey has several limitations. First, given page constraints and the rapid growth of the literature, we necessarily make selective choices about coverage, emphasis, and granularity. Some related topics, benchmarks, and methodological variants are therefore discussed only briefly. Second, while we organize the field into agent–environment, intra-agent, and inter-agent misalignment, these categories are intended as an analytical framework rather than a rigid partition, and some works naturally cut across multiple loci. Third, LLM-agent research is evolving quickly, so any survey can provide only a partial snapshot of a moving landscape. New benchmarks, memory systems, and multi-agent frameworks may extend or reshape the taxonomy over time. Finally, our emphasis is on conceptual synthesis and organization of the field, rather than exhaustive empirical comparison. We hope this survey provides a useful starting point for future refinement as the study of misalignment in LLM agents continues to mature.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. 2025. τ^2-bench: Evaluating conversational agents in a dual-control environment. *Preprint*, arXiv:2506.07982.

Haonan Bian, Zhiyuan Yao, Sen Hu, Zishan Xu, Shaolei Zhang, Yifu Guo, Ziliang Yang, Xueran Han, Huacan Wang, and Ronghao Chen. 2026. Realmem: Benchmarking llms in real-world memory-driven interaction. *arXiv preprint arXiv:2601.06966*.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. Openai gym. *Preprint*, arXiv:1606.01540.

Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, and 1 others. 2025. Why do multi-agent llm systems fail? *arXiv preprint arXiv:2503.13657*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better llm-based evaluators through multi-agent debate. In *International Conference on Learning Representations (ICLR)*.

Chen Chen, Kim Young Il, Yuan Yang, Wenhao Su, Yilin Zhang, Xueluan Gong, Qian Wang, Yongsen Zheng, Ziyao Liu, and Kwok-Yan Lam. 2026a. The shadow self: Intrinsic value misalignment in large language model agents. *Preprint*, arXiv:2601.17344.

Chen Chen, Kim Young Il, Yuan Yang, Wenhao Su, Yilin Zhang, Xueluan Gong, Qian Wang, Yongsen Zheng, Ziyao Liu, and Kwok-Yan Lam. 2026b. The shadow self: Intrinsic value misalignment in large language model agents. *arXiv preprint arXiv:2601.17344*.

Huaben Chen, Wenkang Ji, Lufeng Xu, and Shiyu Zhao. 2023. Multi-agent consensus seeking via large language models. *arXiv preprint arXiv:2310.20151*.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2024a. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Minghao Chen, Yihang Li, Yanting Yang, Shiyu Yu, Binbin Lin, and Xiaofei He. 2024b. Automanual: Constructing instruction manuals by llm agents via interactive environmental learning. *Preprint*, arXiv:2405.16247.

Yize Cheng, Arshia Soltani Moakhar, Chenrui Fan, Parsa Hosseini, Kazem Faghih, Zahra Sodagar, Wenxiao Wang, and Soheil Feizi. 2025. Your llm agents are temporally blind: The misalignment between tool use decisions and human time perception. *arXiv preprint arXiv:2510.23853*.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*.

Hyeong Kyu Choi, Jerry Zhu, and Sharon Li. 2025. Debate or vote: Which yields better decisions in multi-agent large language models? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 3563–3578.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Learning Representations (ICLR)*.

Lutfi Eren Erdogan, Nicholas Lee, Sehoon Kim, Suhong Moon, Hiroki Furuta, Gopala Anumanchipalli, Kurt Keutzer, and Amir Gholami. 2025. Plan-and-act: Improving planning of agents for long-horizon tasks. *arXiv preprint arXiv:2503.09572*.

Jinyuan Fang, Yanwen Peng, Xi Zhang, Yingxu Wang, Xinhao Yi, Guibin Zhang, Yi Xu, Bin Wu, Siwei Liu, Zihao Li, and 1 others. 2025. A comprehensive survey of self-evolving ai agents: A new paradigm bridging foundation models and lifelong agentic systems. *arXiv preprint arXiv:2508.07407*.

Tongtong Feng, Xin Wang, Zekai Zhou, Ren Wang, Yuwei Zhan, Guangyao Li, Qing Li, and Wenwu Zhu. 2025. Evoagent: Self-evolving agent with continual world model for long-horizon tasks. *arXiv preprint arXiv:2502.05907*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

Weizheng Gu, Chengze Li, Zhuohao Yu, Mengyuan Sun, Zhibang Yang, Wei Wang, Hongrui Jia, Shikun Zhang, and Wei Ye. 2026. What do agents learn from trajectory-sft: Semantics or interfaces? *Preprint*, arXiv:2602.01611.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *Preprint*, arXiv:2305.14992.

Bolei He, Nuo Chen, Xinran He, Lingyong Yan, Zhenkai Wei, Jinchang Luo, and Zhen-Hua Ling. 2024. Retrieving, rethinking and revising: The chain-of-verification can improve retrieval augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10371–10393.

Kaiyu He, Mian Zhang, Shuo Yan, Peilin Wu, and Zhiyu Zoey Chen. 2025. Idea: Enhancing the rule learning ability of large language model agent through induction, deduction, and abduction. *Preprint*, arXiv:2408.10455.

Yining Hong, Huang Huang, Manling Li, Li Fei-Fei, Jiajun Wu, and Yejin Choi. 2026. Learning from trials and errors: Reflective test-time planning for embodied llms. *arXiv preprint arXiv:2602.21198*.

Yuanzhe Hu, Yu Wang, and Julian McAuley. 2025a. Evaluating memory in llm agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257*.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, and 1 others. 2025b. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*.

Nan Huo, Jinyang Li, Bowen Qin, Ge Qu, Xiaolong Li, Xiaodong Li, Chenhao Ma, and Reynold Cheng. 2025. Micro-act: Mitigating knowledge conflict in llm-based rag via actionable self-reasoning. *arXiv preprint arXiv:2506.05278*.

Pengcheng Jiang, Jiacheng Lin, Zhiyi Shi, Zifeng Wang, Luxi He, Yichen Wu, Ming Zhong, Peiyang Song, Qizheng Zhang, Heng Wang, and 1 others. 2025. Adaptation of agentic ai. *arXiv preprint arXiv:2512.16301*.

Weiqiang Jin, Hongyang Du, Biao Zhao, Xingwu Tian, Bohang Shi, and Guang Yang. 2025. A comprehensive survey on multi-agent cooperative decision-making: Scenarios, approaches, challenges and perspectives. *arXiv preprint arXiv:2503.13415*.

Arvi Jonnarth, Ola Johansson, Jie Zhao, and Michael Felsberg. 2025. Sim-to-real transfer of deep reinforcement learning agents for online coverage path planning. *IEEE Access*, 13:106883–106905.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. Voting or consensus? decision-making in multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11640–11671, Vienna, Austria. Association for Computational Linguistics.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. 2025. Memory os of ai agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25972–25981.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*.

Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. 2024. Studying large language model behaviors under context-memory conflicts with real documents. *arXiv preprint arXiv:2404.16032*.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See Kiong Ng, Tat-Seng Chua, and Yang Deng. 2025a. Knowledge boundary of large language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5131–5157.

Yanshu Li, Yi Cao, Hongyang He, Qisen Cheng, Xiang Fu, Xi Xiao, Tianyang Wang, and Ruixiang Tang. 2025b. M^{2} iv: Towards efficient and fine-grained multimodal in-context learning via representation engineering. *arXiv preprint arXiv:2504.04633*.

Yanshu Li, Jianjiang Yang, Tian Yun, Pinyuan Feng, Jinfa Huang, and Ruixiang Tang. 2025c. TACO: Enhancing multimodal in-context learning via task mapping-guided sequence configuration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 736–763, Suzhou, China. Association for Computational Linguistics.

Zhenghao Li, Zhi Zheng, Wei Chen, Jielun Zhao, Yong Chen, Tong Xu, and Enhong Chen. 2026. Dynadebate: Breaking homogeneity in multi-agent debate with dynamic path generation. *Preprint*, arXiv:2601.05746.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Xixun Lin, Yucheng Ning, Jingwen Zhang, Yan Dong, Yilong Liu, Yongxuan Wu, Xiaohua Qi, Nan Sun, Yanmin Shang, Kun Wang, and 1 others. 2025. Llm-based agents suffer from hallucinations: A survey of taxonomy, methods, and directions. *arXiv preprint arXiv:2509.18970*.

Junwei Liu, Kaixin Wang, Yixuan Chen, Xin Peng, Zhenpeng Chen, Lingming Zhang, and Yiling Lou. 2024a. Large language model-based agents for software engineering: A survey. *ACM Transactions on Software Engineering and Methodology*.

Kaiming Liu, Xuanyu Lei, Ziyue Wang, Peng Li, and Yang Liu. 2025a. Agent-environment alignment via automated interface generation. *Preprint*, arXiv:2505.21055.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2025b. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE/ASME Transactions on Mechatronics*.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Felix Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. 2025a. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities. *Preprint*, arXiv:2408.04682.

Shuo Lu, Yanyin Chen, Wei Feng, Jiahao Fan, Fengheng Li, Zheng Zhang, Jingjing Lv, Junjie Shen, Ching Law, and Jian Liang. 2025b. Uni-layout: Integrating human feedback in unified layout generation and evaluation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 7709–7718.

Shuo Lu, Haohan Wang, Wei Feng, Weizhen Wang, Shen Zhang, Yaoyu Li, Ao Ma, Zheng Zhang, Jingjing Lv, Junjie Shen, and 1 others. 2026. One size, many fits: Aligning diverse group-wise click preferences in large-scale advertising image generation. *arXiv preprint arXiv:2602.02033*.

Elias Lumer, Pradeep Honaganahalli Basavaraju, Myles Mason, James A. Burke, and Vamse Kumar Subbiah. 2025. Graph rag-tool fusion. *Preprint*, arXiv:2502.07223.

Haotian Luo, Huaisong Zhang, Xuelin Zhang, Haoyu Wang, Zeyu Qin, Wenjie Lu, Guozheng Ma, Haiying He, Yingsha Xie, Qiyang Zhou, and 1 others. 2025a. Ultrahorizon: Benchmarking agent capabilities in ultra long-horizon scenarios. *arXiv preprint arXiv:2509.21766*.

Xufang Luo, Yuge Zhang, Zhiyuan He, Zilong Wang, Siyun Zhao, Dongsheng Li, Luna K. Qiu, and Yuqing Yang. 2025b. Agent lightning: Train any ai agents with reinforcement learning. *Preprint*, arXiv:2508.03680.

Aengus Lynch, Benjamin Wright, Caleb Larson, Kevin K. Troy, Stuart J. Ritchie, Sören Mindermann, Ethan Perez, and Evan Hubinger. 2025. Agentic misalignment: How llms could be an insider threat. *Anthropic Research*. Https://www.anthropic.com/research/agentic-misalignment.

Yougang Lyu, Xi Zhang, Xinhao Yi, Yuyue Zhao, Shuyu Guo, Wenxiang Hu, Jan Piotrowski, Jakub Kaliski, Jacopo Urbani, Zaiqiao Meng, and 1 others. 2026. Evoscientist: Towards multi-agent evolving ai scientists for end-to-end scientific discovery. *arXiv preprint arXiv:2603.08127*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594.

Xinji Mai, Haotian Xu, Zhong-Zhi Li, Xing W, Weinong Wang, Jian Hu, Yingying Zhang, and Wenqiang Zhang. 2025. Agent rl scaling law: Agent rl with spontaneous code execution for mathematical problem solving. *Preprint*, arXiv:2505.07773.

Emanuele La Malfa, Gabriele La Malfa, Samuele Marro, Jie M. Zhang, Elizabeth Black, Michael Luck, Philip Torr, and Michael J. Wooldridge. 2025. Large language models miss the multi-agent mark. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*.

Akshat Naik, Patrick Quinn, Guillermo Bosch, Emma Gouné, Francisco Javier Campos Zabala, Jason Ross Brown, and Edward James Young. 2025. Agentmisalignment: Measuring the propensity for misaligned behaviour in llm-based agents. *arXiv preprint arXiv:2506.04018*.

Liangbo Ning, Ziran Liang, Zhuohang Jiang, Haohao Qu, Yujuan Ding, Wenqi Fan, Xiao-yong Wei, Shanru Lin, Hui Liu, Philip S Yu, and 1 others. 2025. A survey of webagents: Towards next-generation ai agents for web automation with large foundation models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6140–6150.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. 2025. Alphaevolve: A coding agent for scientific and algorithmic discovery. *Preprint*, arXiv:2506.13131.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. 2023. Memgpt: towards llms as operating systems.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Hongjin Qian and Zheng Liu. 2025. Metaagent: Toward self-evolving agent via tool meta-learning. *Preprint*, arXiv:2508.00271.

Irina Radeva, Ivan Popchev, Lyubka Doukovska, and Miroslava Dimitrova. 2025. Multi-agent coordination strategies vs. retrieval-augmented generation in llms: A comparative evaluation. *Electronics*, 14(24):4883.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. 2025. Zep: A temporal knowledge graph architecture for agent memory. *Preprint*, arXiv:2501.13956.

Shuo Ren, Can Xie, Pu Jian, Zhenjiang Ren, Chunlin Leng, and Jiajun Zhang. 2025. Towards scientific intelligence: A survey of llm-based scientific agents. *arXiv preprint arXiv:2503.24047*.

Zarreen Reza. 2026. The social laboratory: A psychometric framework for multi-agent LLM evaluation. In *Women in Machine Learning Workshop @ NeurIPS 2025*.

Rafael Rodriguez-Sanchez, Benjamin A. Spiegel, Jennifer Wang, Roma Patel, Stefanie Tellex, and George Konidaris. 2023. Rlang: A declarative language for describing partial world knowledge to reinforcement learning agents. *Preprint*, arXiv:2208.06448.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. Alfworld: Aligning text and embodied environments for interactive learning. *Preprint*, arXiv:2010.03768.

Yu Su, Diyi Yang, Shunyu Yao, and Tao Yu. 2024a. Language agents: Foundations, prospects, and risks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 17–24, Miami, Florida, USA. Association for Computational Linguistics.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024b. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. *arXiv preprint arXiv:2408.12076*.

Qiushi Sun, Kanzhi Cheng, Zichen Ding, Chuanyang Jin, Yian Wang, Fangzhi Xu, Zhenyu Wu, Chengyou Jia, Liheng Chen, Zhoumianze Liu, Ben Kao, Guohao Li, Junxian He, Yu Qiao, and Zhiyong Wu. 2025. Os-genesis: Automating gui agent trajectory construction via reverse task synthesis. *Preprint*, arXiv:2412.19723.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, and 1 others. 2025. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8416–8439.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. 2025. Gymnasium: A standard interface for reinforcement learning environments. *Preprint*, arXiv:2407.17032.

Meet Udeshi, Venkata Sai Charan Putrevu, Prashanth Krishnamurthy, Ramesh Karri, and Farshad Khorrami. 2025. Samosa: Sandbox for malware orchestration and side-channel analysis. *Preprint*, arXiv:2508.14261.

Haizhou Wang, Nanqing Luo, Xusheng Li, and Peng LIu. 2025a. Unmasking the shadows: Pinpoint the implementations of anti-dynamic analysis techniques in malware using llm. *Preprint*, arXiv:2411.05982.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025b. Retrieval-augmented generation with conflicting evidence. *Preprint*, arXiv:2504.13079.

Junlin Wang, Jue WANG, Ben Athiwaratkun, Ce Zhang, and James Zou. 2025c. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations*.

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

Ruipeng Wang, Yuxin Chen, Yukai Wang, Chang Wu, Junfeng Fang, Xiaodong Cai, Qi Gu, Hui Su, An Zhang, Xiang Wang, and 1 others. 2026a. Agentnoisebench: Benchmarking robustness of tool-using llm agents under noisy condition. *arXiv preprint arXiv:2602.11348*.

Xinyue Wang, Yuanhe Zhang, Zhengshuo Gong, Haoran Gao, Fanyu Meng, Zhenhong Zhou, Li Sun, Yang Liu, and Sen Su. 2026b. From helpfulness to toxic proactivity: Diagnosing behavioral misalignment in llm agents. *arXiv preprint arXiv:2602.04197*.

Xinyue Wang, Yuanhe Zhang, Zhengshuo Gong, Haoran Gao, Fanyu Meng, Zhenhong Zhou, Li Sun, Yang Liu, and Sen Su. 2026c. From helpfulness to toxic proactivity: Diagnosing behavioral misalignment in llm agents. *Preprint*, arXiv:2602.04197.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yu Wang and Xi Chen. 2025. Mirix: Multi-agent memory system for llm-based agents. *arXiv preprint arXiv:2507.07957*.

Kevin Wu, Eric Wu, and James Zou. 2024. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. *Advances in neural information processing systems*, 37:33402–33422.

Zengqing Wu and Takayuki Ito. 2025. The hidden strength of disagreement: Unraveling the consensus-diversity tradeoff in adaptive multi-agent systems. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15288–15308.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Preprint*, arXiv:2404.07972.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7572–7590.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.

Taku Yamagata and Raul Santos-Rodriguez. 2024. Safe and robust reinforcement learning: Principles and practice. *arXiv preprint arXiv:2403.18539*.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation.

Zeyuan Yang, Peng Li, and Yang Liu. 2023. Failures pave the way: Enhancing large language models through tuning-free rule accumulation. *Preprint*, arXiv:2310.15746.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Miao Yu, Fanci Meng, Xinyun Zhou, Shilong Wang, Junyuan Mao, Linsey Pan, Tianlong Chen, Kun Wang, Xinfeng Li, Yongfeng Zhang, and 1 others. 2025. A survey on trustworthy llm agents: Threats and countermeasures. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6216–6226.

Xiao Yu, Baolin Peng, Ruize Xu, Yelong Shen, Pengcheng He, Suman Nath, Nikhil Singh, Jiangfeng Gao, and Zhou Yu. 2026. Reinforcement world model learning for llm-based agents. *Preprint*, arXiv:2602.05842.

Guilin Zhang, Wei Jiang, Xiejiashan Wang, Aisha Behr, Kai Zhao, Jeffrey Friedman, Xu Chu, and Amine Anoun. 2026a. Adaptive memory admission control for llm agents. *arXiv preprint arXiv:2603.04549*.

Hanchen Zhang, Xiao Liu, Bowen Lv, Xueqiao Sun, Bohao Jing, Iat Long Iong, Zhenyu Hou, Zehan Qi, Hanyu Lai, Yifan Xu, Rui Lu, Hongning Wang, Jie Tang, and Yuxiao Dong. 2025a. Agentrl: Scaling agentic reinforcement learning with a multi-turn, multi-task framework. *Preprint*, arXiv:2510.04206.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025b. Faithfulrag: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21863–21882.

Weichen Zhang, Yiyou Sun, Pohao Huang, Jiayue Pu, Heyue Lin, and Dawn Song. 2025c. Mirage-bench: Llm agent is hallucinating and where to find them. *arXiv preprint arXiv:2507.21017*.

Yinger Zhang, Shutong Jiang, Renhao Li, Jianhong Tu, Yang Su, Lianghao Deng, Xudong Guo, Chenxu Lv, and Junyang Lin. 2026b. Deepplanning: Benchmarking long-horizon agentic planning with verifiable constraints. *arXiv preprint arXiv:2601.18137*.

Yufeng Zhang, Xuepeng Wang, Lingxiang Wu, and Jinqiao Wang. 2026c. Characterizing and mitigating reasoning drift in large language models. In *The Fourteenth International Conference on Learning Representations*.

Xinjie Zhao, Fan Gao, Rui Yang, Yingjian Chen, Yuyang Wang, Ying Zhu, Jiacheng Tang, and Irene Li. 2025. Reagent: Reversible multi-agent reasoning for knowledge-enhanced multi-hop qa. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Han Zhou, Xingchen Wan, Ruoxi Sun, Hamid Palangi, Shariq Iqbal, Ivan Vulić, Anna Korhonen, and Sercan Ö Arık. 2025. Multi-agent design: Optimizing agents with better prompts and topologies. *arXiv preprint arXiv:2502.02533*.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. 2024a. Rel-a.i.: An interaction-centered approach to measuring human-lm reliance. *Preprint*, arXiv:2407.07950.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024b. Webarena: A realistic web environment for building autonomous agents. *Preprint*, arXiv:2307.13854.

Siyu Zhou, Tianyi Zhou, Yijun Yang, Guodong Long, Deheng Ye, Jing Jiang, and Chengqi Zhang. 2024c. Wall-e: World alignment by rule learning improves world model-based llm agents. *Preprint*, arXiv:2410.07484.

Yuqing Zhou, Zhuoer Wang, Jie Yuan, Hong Wang, Samson Koelle, Ziwei Zhu, and Wei Niu. 2026. Wiseflow: Workflow-induced structured experience for self-evolving conversational service agents. *Preprint*, arXiv:2601.08158.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. 2025. MultiAgentBench : Evaluating the collaboration and competition of LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8580–8622, Vienna, Austria. Association for Computational Linguistics.

Qiming Zhu, Shunian Chen, Rui Yu, Zhehao Wu, and Benyou Wang. 2026a. From lossy to verified: A provenance-aware tiered memory for agents. *arXiv preprint arXiv:2602.17913*.

Xiaochen Zhu, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier, and Andreas Vlachos. 2026b. Demystifying multi-agent debate: The role of confidence and diversity. *Preprint*, arXiv:2601.19921.

A Main Paper Collections

As shown in Table **??**, the papers can be grouped by category, subcategory, mitigation strategy, evaluation setting, and domain.

Method / Paper	Pub.	Primary Mitigation	Application Domain
Agent–Environment			
WISE-Flow	arXiv'26	Prompt	Conversational agents
ALIGN	arXiv'25	Architecture	Tool use / interface-grounded agents
AutoManual	NeurIPS'24	Prompt	Environment adaptation
Failures Pave the Way	EMNLP'23	Prompt	Rule accumulation / environment interaction
Graph RAG-Tool Fusion	arXiv'25	SFT	Tool use / graph-enhanced reasoning
OS-Genesis	ACL'25	RL	GUI agents / synthetic trajectory generation
WALL-E	NeurIPS'25	SFT	World-model planning
MetaAgent	arXiv'25	RL	Self-evolving tool-use agents
RAP	EMNLP'23	Inference	World-model planning
RLang	ICML'23	Architecture	Environment specification / planning
AgentRL	arXiv'25	RL	General agent tasks
Agent RL Scaling Law	arXiv'25	RL	Mathematical reasoning / tool use
Agent Lightning	arXiv'25	RL	General agent training
RWML	arXiv'26	RL	World-model planning
Intra-Agent			
MemGPT	arXiv'23	Architecture	Long-horizon memory agents
Mem0	ACL'25	Architecture	Long-term memory agents
Self-Refine	NeurIPS'23	Inference	General reasoning / self-revision
CoVe	ICLR'24	Inference	Verification-based generation / hallucination reduction
SpARE	NAACL'25	Inference	Knowledge-conflict QA / context-memory conflict
JUICE	EMNLP'24	Inference	Knowledge-conflict reasoning
Trust-Align	ICLR'25	SFT	RAG / knowledge-intensive QA
RAG with Conflicting Evidence	NAACL'25	Inference	RAG / conflicting evidence resolution
PA-RAG	arXiv'25	SFT	RAG / knowledge-intensive QA
Inter-Agent			
PLAY2PROMPT	arXiv'25	Prompt	Tool-augmented agents
MASS	arXiv'25	Architecture	MAS coordination
RECONCILE	ACL'24	Inference	Multi-model consensus reasoning
Debate or Vote?	NeurIPS'25	Inference	Decision protocol analysis
Voting or Consensus?	Findings ACL'25	Inference	Aggregation dynamics
CONSENSAGENT	Findings ACL'25	Prompt	Consensus stability
Multiagent Debate	ICLR'24	Inference	Factual QA / reasoning
Divergent Thinking	EMNLP'24	Prompt	Logical reasoning
Breaking Mental Set	ICLR'25	Prompt	Mathematical / logical reasoning
ReAgent	EMNLP'25	Inference	Multi-hop QA
Hidden Strength of Disagreement	EMNLP'25	Inference	Adaptive MAS / diversity tradeoff
Demystifying Multi-Agent Debate	arXiv'26	Inference	Debate dynamics
Mixture-of-Agents	ICLR'25	Architecture	Multi-model collaboration
DynaDebate	arXiv'26	Inference	Debate-based reasoning

Table 1: Representative mitigation-oriented papers organized by misalignment category and primary mitigation mechanism. We omit benchmark, diagnostic, resource, and position papers, and assign each retained paper to a single primary mitigation class for comparability.