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. Misalignment in LLM agents has thus 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, namely agent–environment, intra-agent, and inter-agent misalignment, and synthesize their causes, mitigation strategies, and evaluation practices into prescriptive takeaways for handling each category. We further analyze how misalignment propagates and compounds across categories, and 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; Fang et al., 2025a). This transition has broadened LLM applications from single-turn assistance to complex multi-step settings such as web automation (Zhou et al., 2024a), software engineering (Liu et al., 2024a), scientific discovery (Ren et al., 2025), and multi-agent collaboration (Du et al., 2024). With this growing autonomy, however, *agentic misalignment* emerges as a distinct and critical challenge (Naik et al., 2025; Wang et al., 2026b). Agentic misalignment refers to situations where the behavior of an LLM agent, or of multiple interacting 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). Therefore, agentic misalignment is not a matter of incorrect single outputs, but a systems-level challenge shaped by recurrent interaction and state.

Existing literature fails to systematically characterize misalignment in LLM agents as shown in Table 1. Surveys on LLM agents focus on capabilities such as architectures, planning, memory, tool use, and multi-agent collaboration (Wang et al., 2024a; Huang et al., 2024; Hu et al., 2025; Jiang et al., 2025; Jin et al., 2025), rather than the misalignment dynamics that govern long-horizon reliability. Other literature on stand-alone LLMs covers related phenomena such as knowledge conflicts (Xu et al., 2024), hallucination (Huang et al., 2025), and knowledge boundaries (Li et al., 2025b), but misses the iterative agent–environment interaction that makes the agentic case fundamentally harder. The most closely related to our scope are Lin et al. (2025) and Yu et al. (2025); both study the trustworthiness of agentic systems but organize the field around a single failure mode (hallucination or threat) rather than treating misalignment as a unified, closed-loop systems problem that explicitly connects the different categories. Consequently, the field still lacks a unified framework for understanding where misalignment arises in LLM agents,

Survey	Primary scope	Closed-loop framing	Agent–Env	Intra-Agent	Inter-Agent	Cross-category interactions
Wang et al. (2024a)	Agent construction modules	✗	✓	✓	✓	✗
Jin et al. (2025)	Multi-agent cooperation methods	✗	✓	✗	✓	✗
Xu et al. (2024)	LLM knowledge conflicts	✗	✗	✓	✗	✓
Lin et al. (2025)	Agent hallucination taxonomy	✗	✓	✓	✓	✗
Yu et al. (2025)	Agent trustworthiness threats	✗	✓	✗	✓	✓
Qu et al. (2025)	Misalignment in stand-alone LLMs	✗	✗	✓	✗	✗
Ours	**Misalignment in LLM agents**	✓	✓	✓	✓	✓

Table 1: Positioning of **Ours** relative to the most-related prior surveys. *Closed-loop framing*: whether the literature is organized under one unified closed-loop interaction dynamics. *Agent–Env / Intra-Agent / Inter-Agent*: dedicated coverage of the corresponding misalignment category. *Cross-category interactions*: whether the survey analyses how failures in one category propagate to or amplify failures in another. No prior survey jointly satisfies all criteria.

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 settings. Building on this taxonomy, we further examine **cross-category interactions** (Sec. 6), which concern how misalignment in one category propagates to or amplifies misalignment in another; **evaluation** (Sec. 7), which organizes benchmarks and metrics that probe misalignment across the three categories; and **challenges and future directions** (Sec. 8), which discuss open problems for building more robust and aligned agentic systems.

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 from three major categories: agent–environment misalignment, intra-agent misalignment, and inter-agent misalignment. ii) 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. iii) We highlight open problems and future research directions that are critical for building more robust and aligned 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, namely agent–environment, intra-agent, and inter-agent settings, alongside a unified view of 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. Formally, 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

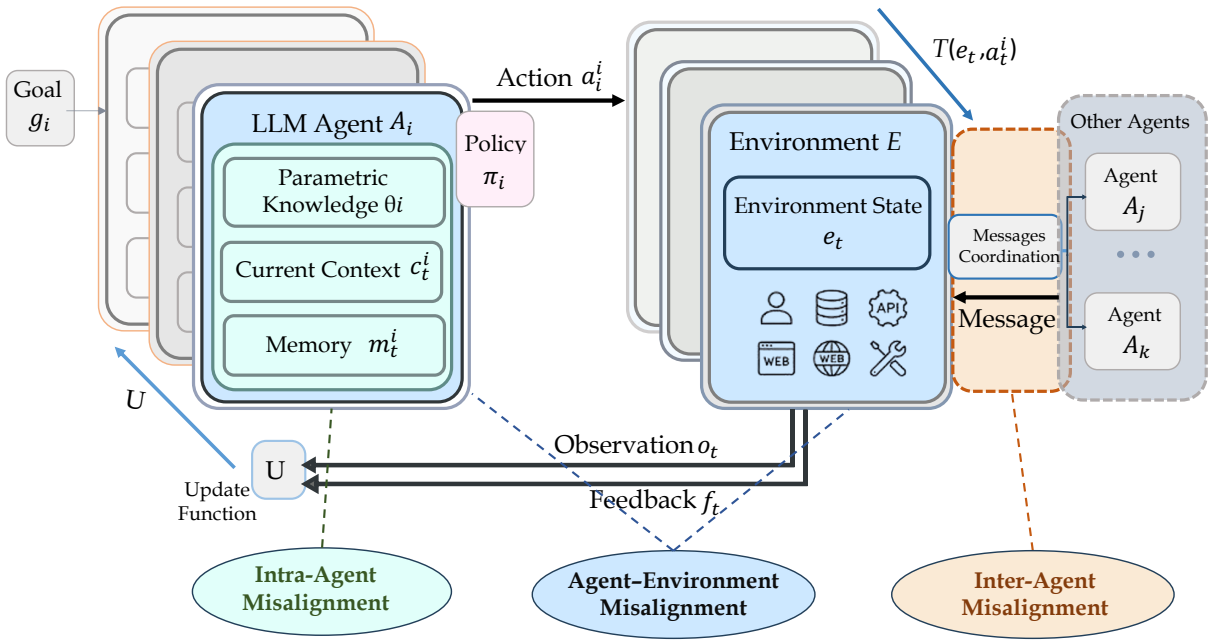

Figure 1: Three types of agentic misalignment between LLM agent, environment and other agents.

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. This formulation 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 $\mathcal{E}$.

2.2 Misalignment Types and Mitigation

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 3. 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.

2.3 Survey Scope and Selection

We curate the literature through keyword search on *arXiv* and the *ACL Anthology*, complemented by citation tracking from seed papers; full search terms and inclusion criteria are in Appendix A. The survey covers over 110 distinct works published from 2022 onward across NeurIPS, ICLR, ICML, *ACL, EMNLP, and arXiv. The mitigation-oriented core in Table 3 consists of 37 representative methods spanning the three loci: 14 in agent–environment, 9 in intra-agent, and 14 in inter-agent misalignment.

3 Agent-Environment Misalignment

Agent–environment misalignment arises at the closed-loop boundary in Fig. 1: either the agent acts on incorrect assumptions about the environ-

ment's transition T, or it misinterprets the observation o_t and feedback f_t that the update $\mathcal{U}$ then writes into context and memory.

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; 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 (**?**Wang et al., 2026b).

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., 2025; Wang et al., 2025a). 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., 2026b). 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., 2024b). In terms of Fig. 1, these methods improve the quality of o_t and f_t, reducing mismatch between environmental outcomes and agent updates.

Adapting the Agent through Trajectories and World Models. Agent-side methods adapt the policy π and update $\mathcal{U}$ from environment-specific interaction. Reinforcement learning, meta-learning, and exploration-based approaches tune them to reflect actual action consequences, failure patterns, and recovery behaviors (Qian and Liu, 2025; Zhang et al., 2025a; Mai et al., 2025; Luo et al., 2025b; Yang et al., 2026b); world-model methods go further by learning an explicit forward model of $T(e_t, a_t)$ and $\Phi(e_t, a_t)$: WALL-E induces symbolic rules from trajectories, RWML learns action-conditioned dynamics, and RAP uses the LLM as a predictive simulator for multi-step planning (Zhou et al., 2024b; Yu et al., 2026; Hao et al., 2023; Yang et al., 2026a). Both target the cross-boundary prediction problem in Fig. 1: 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 Take-Aways

The three families above are not alternatives. Each acts on a different point of the closed loop in §3. Signal exposure and protocol standardization are env-side: they sharpen o_t and f_t or remove interface ambiguity before the agent sees them. Agent-side adaptation tunes π and $\mathcal{U}$ from interaction or learns an explicit forward model of T. On ALFWorld, env-side methods reach competitive task-success and invalid-action numbers without touching the agent (Liu et al., 2025a; He et al., 2025), while agent-side methods target transition mismatch under shift at orders of magnitude higher cost in samples and offline trajectories (Yu et al., 2026; Qian and Liu, 2025).

Two take-aways follow for handling agent-environment misalignment. First, the families are

complementary rather than competing. Each addresses a distinct failure, so the right method is chosen based on the failure type rather than by any intrinsic ranking among families. Second, this gives a clear ordering for practice. Env-side interventions are the cheap default; agent-side adaptation is justified when the residual failure is transition mismatch under shift, with explicit world models paying off specifically when the bottleneck is a wrong forward model of T.

4 Intra-Agent Misalignment

Intra-agent misalignment refers to mismatches *within* the decision state of a single LLM agent, that is, inconsistencies among parametric knowledge θ, runtime context c_t, memory m_t, the update process $\mathcal{U}$, and the induced policy $\pi(a \mid g, c_t, m_t; \theta)$. We organize it into three forms: intrinsic, intrinsic–extrinsic, and extrinsic misalignment.

4.1 Underlying Causes

Intrinsic Misalignment. Intrinsic misalignment originates from 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., 2023), latent uncertainty among competing continuations (Xiong et al., 2024), reasoning drift over multi-step inference (Zhang et al., 2026c), and divergence between high-level planning and downstream action (Erdogan et al., 2025). Even when c_t, m_t, and $\mathcal{U}$ are held fixed, the induced policy can remain internally inconsistent: the mismatch lies within what the model encodes and how it reasons from that encoding.

Intrinsic-Extrinsic Misalignment. Intrinsic–extrinsic misalignment arises when intrinsic knowledge θ conflicts with runtime information in c_t or m_t: parametric knowledge disagrees with newly retrieved evidence, prior beliefs conflict with updated memory, or the agent fails to decide whether new context should override what θ already encodes (Xu et al., 2024; Kortukov et al., 2024; Huo et al., 2025; Li et al., 2025c). The conflict typically surfaces when $\mathcal{U}$ incorporates new observations without reconciling them against the prior parametric knowledge.

Extrinsic Misalignment. Extrinsic misalignment arises when the runtime substrate c_t or m_t becomes unreliable after incorporating external information, such as contradictory context snippets, stale retrieved passages, noisy user instructions, or memory write/revision errors (Liu et al., 2024b; Cheng et al., 2026; Hu et al., 2026). The failure is most directly tied to the update process $\mathcal{U}(c_t, m_t, o_t, a_t, f_t)$, where the resulting state becomes corrupted, stale, or poorly synchronized with later reasoning. The source is external, but the mismatch is expressed inside the agent because subsequent decisions condition on a distorted state.

4.2 Mitigation Strategies

Reasoning Stabilization. To stabilize how π is induced from the intrinsic substrate, three method families dominate. Self-verification (CoVe (Dhuliawala et al., 2024), CRITIC (Gou et al., 2024)) inserts explicit checking or critique steps to validate intermediate reasoning or action candidates before commitment. Reflective refinement (Reflexion (Shinn et al., 2023), Self-Refine (Madaan et al., 2023)) iteratively revises reasoning traces, memory, or outputs from self-generated feedback to reduce local drift. Plan-aware stabilization (Reflective Test-Time Planning (Hong et al., 2026), EvoAgent (Feng et al., 2025)) combines reflection with explicit planning or world-model guidance to keep action selection aligned with longer-horizon goals. Together they regularize the mapping from intrinsic state to action, yielding more coherent long-horizon policy trajectories.

Knowledge Reconciliation. To reconcile θ with the run-time state (c_t, m_t), mitigation focuses on explicitly detecting and arbitrating conflicts between internal knowledge and incorporated evidence. Retrieval-with-critique methods, including Self-RAG (Asai et al., 2023), CRAG (Yan et al., 2024), and CoV-RAG (He et al., 2024), combine retrieval with critique and corrective retrieval. Conflict-centered methods like FaithfulRAG (Zhang et al., 2025b) and MicroAct (Huo et al., 2025) go further by modeling parametric–retrieved disagreements and resolving them through structured self-reasoning, improving how $\mathcal{U}$ 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 manage-

ment, 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 Take-Aways

The three families above map directly onto the three sub-types of intra-agent misalignment in §4.1. Reasoning stabilization addresses intrinsic misalignment by regularizing how the policy π is induced from θ alone. Knowledge reconciliation addresses intrinsic-extrinsic misalignment by explicitly arbitrating conflicts between θ and the run-time state (c_t, m_t). Run-time state governance addresses extrinsic misalignment by controlling the construction and revision of c_t and m_t themselves.

Two take-aways follow for handling intra-agent misalignment. First, the correct family is determined by which substrate is broken, not by the surface symptoms. Apparent reasoning instability may stem from a corrupted run-time state rather than from intrinsic drift, and apparent retrieval errors may persist after the evidence is clean if θ itself is inconsistent. Second, the families compose by substrate order. Governance ensures the run-time state is reliable; reconciliation aligns θ with that state; stabilization regularizes how π is induced from the result. Skipping a layer leaves the failure unaddressed even when the local fix succeeds.

5 Inter-Agent Misalignment

Inter-agent misalignment arises when multiple agents collaborating on shared goals maintain incompatible informational states, reasoning trajectories, or decisions. We trace it to three structural asymmetries: differences in parametric knowledge θ, in contextual information c_t, and in the interaction protocol that governs how each agent updates its state from others' outputs.

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 encode distinct priors and 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. Even when heterogeneity is built into the system intentionally (e.g., Mixture-of-Agents (Wang et al., 2025c)), the resulting complementary priors become a structural source of representational misalignment across agents.

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., 2025a; 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 Strategies

Reconciling Heterogeneous Beliefs. For misalignment driven by model heterogeneity and role conditioning, mitigation targets divergence in priors and reasoning stance. Optimization frameworks such as MASS jointly tune agent roles and communication topology to reduce persistent belief-level divergence (Zhou et al., 2026a). Structured reconciliation protocols stabilize heterogeneous groups by requiring explicit justification and cross-agent comparison before aggregation (Chen et al., 2024a;

Wu and Ito, 2025). Reversible reasoning architectures complement these by letting agents roll back intermediate commitments when downstream evidence contradicts them (Zhao et al., 2025a).

Resolving Information Asymmetry. When misalignment is driven by divergent evidence pools, mitigation operates at the epistemic layer. Contradiction-aware retrieval and provenance-sensitive reasoning reduce the chance that inconsistent or stale documents propagate through collaborative reasoning chains (Wang et al., 2025b; Wu et al., 2024). Voting and aggregation protocols that surface evidence disagreement explicitly, rather than averaging it into a single consensus, keep contradictions visible to downstream resolution steps (Kaesberg et al., 2025).

Protocol-Level Stabilization. At the protocol layer, mitigation reduces divergence introduced by aggregation rules. Debate-based protocols with calibrated confidence communication keep deliberation from collapsing into majority voting without meaningful exchange (Du et al., 2024; Liang et al., 2024; Zhu et al., 2026b). Dynamic debate architectures stabilize deliberation further through specialized roles that prevent both premature consensus and runaway disagreement (Li et al., 2026). Confidence-weighted voting and minority-preserving aggregation protect correct outliers from being suppressed by conformity pressure and sycophancy (Choi et al., 2025a,b; Pitre et al., 2025).

5.3 Take-Aways

The three families above map directly onto the three causes from §5.1: reconciling heterogeneous beliefs for heterogeneity in θ and role conditioning, resolving information asymmetry for asymmetric access to evidence in c_t, and protocol-level stabilization for aggregation rules. Two take-aways follow for handling inter-agent misalignment. First, the right intervention is determined by which asymmetry dominates. Heterogeneous priors are not removed by stronger aggregation rules; protocol fixes do not resolve evidence conflicts; and evidence-layer mitigation does not realign agents whose priors fundamentally disagree. Second, the three asymmetries often compound in practice, so robust systems combine all three families rather than relying on a single layer.

6 Cross-Category Interactions

The three loci above are operationally coupled: misalignment in one category rarely remains there, and the systems-level loop promised in §2 closes through these interactions.

Propagation and Compounding. Local misalignment cascades: an ambiguous observation–feedback pair (agent–environment update mismatch) is written by $\mathcal{U}$ as distorted context or memory and resurfaces as intra-agent extrinsic misalignment, then as inter-agent disagreement when shared in communication (Cemri et al., 2025; Lin et al., 2025). Two-locus combinations compound super-linearly: a weak world model paired with noisy memory updates makes long-horizon planning catastrophically brittle, and model heterogeneity over agents with parametric–context conflicts yields debates that cannot converge because each agent's "evidence" already reflects an unresolved internal contradiction (Wang et al., 2024b; Zhu et al., 2026b).

Joint Mitigation through the Update Function $\mathcal{U}$. Single-locus fixes do not stack: aggressive memory consolidation that stabilizes intra-agent reasoning can erase the diversity inter-agent debate exploits (Wu and Ito, 2025). The update function $\mathcal{U}$ sits at the intersection of all three loci, mediating how environmental feedback enters context and memory and what each agent then exposes to others. Failures in $\mathcal{U}$ (provenance loss, over-compression, conflicting writes, stale snapshots) therefore manifest at all three loci simultaneously and make $\mathcal{U}$ a natural unified target for cross-locus mitigation.

7 Evaluation

Evaluation of agentic misalignment must address three shared questions across the three loci: *what* is being measured (outcome correctness vs. process quality), *under what distribution* (in-distribution, out-of-distribution, or perturbed), and *at what granularity* (final task, trajectory, or single step). Below we organize representative benchmarks per locus along these axes (summarized in Appendix B, Table 2), and end with cross-cutting gaps.

Agent–Environment Evaluation. Two metric families dominate. *Distributional generalization* probes *transition mismatch* by comparing ID and OOD success rates—e.g., RWML on ALFWorld (Yu et al., 2026; Shridhar et al., 2021) to show

that world-model learning improves transfer. *Interface and execution* metrics probe *update mismatch*: invalid-action rates and interface reliance (Zhou et al., 2025) reveal whether agents lean on surface cues, while tool-use benchmarks such as τ^2-Bench (Barres et al., 2025) show that failures often come from malformed arguments or wrong call ordering rather than wrong tool choice. AgentBench (Liu et al., 2024c) and OSWorld (Xie et al., 2024) extend evaluation to realistic open-ended environments spanning web, OS, and tool-use settings.

Intra-Agent Evaluation. Each form maps to a distinct benchmark family. *Intrinsic* misalignment is probed by long-horizon planning-stability tests like DeepPlanning (Zhang et al., 2026b) and UltraHorizon (Luo et al., 2025a). *Intrinsic–extrinsic* misalignment is tested by benchmarks pitting θ against runtime evidence—ConflictBank (Su et al., 2024) and MemoryAgentBench (Hu et al., 2026). *Extrinsic* misalignment is captured by benchmarks stressing c_t, m_t, and the update $\mathcal{U}$ under evolving or noisy input: RealMem (Bian et al., 2026) and AgentNoiseBench (Wang et al., 2026a).

Inter-Agent Evaluation. Three metric families address the causes in §5.1. *Coordination stability* measures whether identical agents converge under different protocols: MultiAgentBench (Zhu et al., 2025) reports that topology (star, chain, tree, graph) materially shifts collaboration quality even with identical agents. *Conflict resolution* probes information asymmetry: ClashEval (Wu et al., 2024) and Conflicting-RAG (Wang et al., 2025b) quantify how unresolved evidence contradictions distort downstream reasoning. *Social and aggregation dynamics* are evaluated via debate-based judging frameworks such as ChatEval (Chan et al., 2024) and benchmarks targeting consensus formation and role-induced bias (Reza, 2026).

Cross-Cutting Gaps. Four gaps cut across all three loci. *(i) Scalability of LLM-as-judge*: many process-level metrics rely on LLM judges or human annotation, both of which inherit known biases and scale poorly (Zheng et al., 2023; Gu et al., 2024), so reproducible low-cost graders for trajectory and coordination quality remain an open need. *(ii) No cross-locus benchmarks*: existing suites probe one locus at a time, leaving compound failures untested—a representative case is an intra-agent knowledge conflict that triggers an agent–environment update mismatch and then propagates through inter-agent communication, where no standardized test covers the full chain. *(iii) Attribution probes*: empirical failure-mode taxonomies (Cemri et al., 2025) document *what* goes wrong, while controlled tests that distinguish whether a mitigation removes a target cause or merely masks it remain sparse. *(iv) Process vs. outcome coverage*: most benchmarks still emphasize final task success, while metrics for intermediate stability, recovery, and faithful state update remain under-developed. Closing these gaps is a prerequisite for measuring real progress on agentic misalignment.

8 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 on failure attribution, error diagnosis, and modular debugging remains mostly post hoc and benchmark-specific, rarely providing a unified way to localize misalignment across perception, state update, policy formation, and coordination. Future work should develop finer-grained attribution methods that identify not only where failure appears but also how local errors cascade into later reasoning.

Balancing Autonomy and Controllability. As LLM agents become more autonomous, a key challenge is to preserve controllability without overly constraining useful behavior. Stronger planning, memory, and reflection improve autonomy yet make behavior harder to predict, supervise, or redirect—especially when agents update their own state, revise plans, or act over long horizons with limited human intervention. Future research should study how to combine agent autonomy with explicit mechanisms for intervention, correction, and oversight under realistic deployment.

Additional directions on process-level evaluation and on scaling mitigation under practical deployment constraints are discussed in Appendix D.

9 Conclusion

This survey presents a unified framework for misalignment in LLM agents across agent–environment, intra-agent, and inter-agent settings. By organizing prior work on its causes, propagation, mitigation, and evaluation, we frame misalignment as a systems-level challenge for building reliable and controllable LLM-agent systems.

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? In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*.

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)*.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yize Cheng, Arshia Soltani Moakhar, Chenrui Fan, Parsa Hosseini, Kazem Faghih, Zahra Sodagar, Wenxiao Wang, and Soheil Feizi. 2026. Your llm agents are temporally blind: The misalignment between tool use decisions and human time perception. In *Findings of the Association for Computational Linguistics: ACL 2026*.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI)*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pages 2993–3000.

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

Min Choi, Keonwoo Kim, Sungwon Chae, and Sangyeob Baek. 2025b. An empirical study of group conformity in multi-agent llms. In *Findings of the Association for Computational Linguistics: ACL*.

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. In *Proceedings of the International Conference on Machine Learning (ICML)*.

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

Wei Fang, Yang Zhang, Kaizhi Qian, James Glass, and Yada Zhu. 2025b. Play2prompt: Zero-shot tool instruction optimization for llm agents via tool play. In *Findings of the Association for Computational Linguistics: ACL 2025*.

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. 2024. Critic: Large language models can self-correct with tool-interactive critiquing. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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. In *Findings of the Association for Computational Linguistics: ACL 2025*.

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. 2026. Evaluating memory in llm agents via incremental multi-turn interactions. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, and 1 others. 2025. 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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

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.

Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. 2024. Studying large language model behaviors under context-memory conflicts with real documents. In *Conference on Language Modeling (COLM)*.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. 2025a. Taming knowledge conflicts in language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See Kiong Ng, Tat-Seng Chua, and Yang Deng. 2025b. 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, 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.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024c. AgentBench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. 2025b. Breaking mental set to improve reasoning through diverse multi-agent debate. In *The Thirteenth International Conference on Learning Representations*.

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. 2025. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

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.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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*.

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

Priya Pitre, Naren Ramakrishnan, and Xuan Wang. 2025. Consensagent: Towards efficient and effective consensus in multi-agent llm interactions through sycophancy mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22112–22133.

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

Yubin Qu, Song Huang, Long Li, Peng Nie, and Yongming Yao. 2025. Beyond intentions: A critical survey of misalignment in llms. *Computers, Materials and Continua*, 85(1):249–300.

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. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

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. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. 2025. Measuring and enhancing trustworthiness of LLMs in RAG through grounded attributions and learning to refuse. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*.

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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

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. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*.

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. In *Conference on Language Modeling (COLM)*.

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. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024b. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6106–6131.

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. *Preprint*, arXiv:2602.04197.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

Jiayi Wu, Hengyi Cai, Lingyong Yan, Hao Sun, Xiang Li, Shuaiqiang Wang, Dawei Yin, and Ming Gao. 2025. PA-RAG: RAG alignment via multi-perspective preference optimization. In *Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*.

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. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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.

Shuo Yang, Caren Han, Yihao Ding, Shuhe Wang, and Eduard Hovy. 2026a. Tooltree: Efficient llm tool planning via dual-feedback monte carlo tree search and bidirectional pruning. In *The Fourteenth International Conference on Learning Representations*, volume 3.

Shuo Yang, Soyeon Caren Han, Xueqi Ma, Yan Li, Mohammad Reza Ghasemi Madani, and Eduard Hovy. 2026b. Evotool: Self-evolving tool-use policy optimization in llm agents via blame-aware mutation and diversity-aware selection. *arXiv preprint arXiv:2603.04900*.

Zeyuan Yang, Peng Li, and Yang Liu. 2023. Failures pave the way: Enhancing large language models through tuning-free rule accumulation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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.

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. 2025a. Reagent: Reversible multi-agent reasoning for knowledge-enhanced multi-hop qa. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2025b. Steering knowledge selection behaviours in LLMs via SAE-based representation engineering. In *Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Han Zhou, Xingchen Wan, Ruoxi Sun, Hamid Palangi, Shariq Iqbal, Ivan Vulić, Anna Korhonen, and Sercan Ö Arık. 2026a. Multi-agent design: Optimizing agents with better prompts and topologies. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. 2025. Rel-a.i.: An interaction-centered approach to measuring human-lm reliance. In *Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*.

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. 2024a. Webarena: A realistic web environment for building autonomous agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Siyu Zhou, Tianyi Zhou, Yijun Yang, Guodong Long, Deheng Ye, Jing Jiang, and Chengqi Zhang. 2024b. 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. 2026b. Wise-flow: 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 Survey Scope and Selection Details

Search keywords by locus. Agent–environment: *world model alignment*, *tool use failure*, *interface grounding*, *action–effect mismatch*, *sim-to-real gap*. Intra-agent: *knowledge conflict*, *memory inconsistency*, *planning–execution gap*, *parametric–context conflict*, *self-consistency*. Inter-agent: *multi-agent debate*, *agent coordination*, *consensus formation*, *model heterogeneity*, *information asymmetry*.

Inclusion and exclusion. We include papers that (i) target LLM agents operating through multi-step interaction with an environment, memory, or other agents, and (ii) introduce a cause analysis, mitigation method, or evaluation framework for misalignment. We exclude stand-alone LLM hallucination work without an agentic component, isolated tool-learning work without an alignment focus, and position papers without methodological contribution.

Mitigation classification. Each method in Table 3 is assigned to a single primary mitigation class (prompt-based, supervised fine-tuning, inference-time, reinforcement learning, or architectural) based on its main intervention mechanism. Multi-class methods are assigned to their dominant intervention to keep the table comparable across categories.

B Evaluation Benchmarks

Table 2 groups representative benchmarks by locus and metric family, complementing the discussion in §7.

Locus	Metric family	Representative benchmarks
Agent–Env	ID/OOD generalization	ALFWorld, τ^2-Bench
	Interaction & execution	AgentBench, OSWorld
Intra-Agent	Long-horizon stability	DeepPlanning, UltraHorizon
	Knowledge reconciliation	ConflictBank, MemoryAgentBench
	State robustness	RealMem, AgentNoiseBench
Inter-Agent	Coordination stability	MultiAgentBench
	Conflict resolution	ClashEval, Conflicting-RAG
	Social & aggregation	ChatEval, Reza et al. (2026)

Table 2: Representative benchmarks for evaluating misalignment, organized by locus and metric family.

C Main Paper Collections

As shown in Table 3, the papers can be grouped by category, subcategory, mitigation strategy, evaluation setting, and domain.

Table 3: 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.

Method / Paper	Pub.	Primary Mitigation	Application Domain
Agent–Environment			
WISE-Flow (Zhou et al., 2026b)	arXiv'26	Prompt	Conversational agents
ALIGN (Liu et al., 2025a)	arXiv'25	Architecture	Tool use / interface-grounded agents
AutoManual (Chen et al., 2024b)	NeurIPS'24	Prompt	Environment adaptation
Failures Pave the Way (Yang et al., 2023)	EMNLP'23	Prompt	Rule accumulation / environment interaction
Graph RAG-Tool Fusion (Lumer et al., 2025)	arXiv'25	SFT	Tool use / graph-enhanced reasoning
OS-Genesis (Sun et al., 2025)	ACL'25	RL	GUI agents / synthetic trajectory generation
WALL-E (Zhou et al., 2024b)	arXiv'24	SFT	World-model planning
MetaAgent (Qian and Liu, 2025)	arXiv'25	RL	Self-evolving tool-use agents
RAP (Hao et al., 2023)	EMNLP'23	Inference	World-model planning
RLang (Rodriguez-Sanchez et al., 2023)	ICML'23	Architecture	Environment specification / planning
AgentRL (Zhang et al., 2025a)	arXiv'25	RL	General agent tasks
Agent RL Scaling Law (Mai et al., 2025)	arXiv'25	RL	Mathematical reasoning / tool use
Agent Lightning (Luo et al., 2025b)	arXiv'25	RL	General agent training
RWML (Yu et al., 2026)	arXiv'26	RL	World-model planning
Intra-Agent			
MemGPT (Packer et al., 2023)	arXiv'23	Architecture	Long-horizon memory agents
Mem0 (Chhikara et al., 2025)	ECAI'25	Architecture	Long-term memory agents
Self-Refine (Madaan et al., 2023)	NeurIPS'23	Inference	General reasoning / self-revision
CoVe (Dhuliawala et al., 2024)	Findings ACL'24	Inference	Verification-based generation / hallucination reduction
SpARE (Zhao et al., 2025b)	NAACL'25	Inference	Knowledge-conflict QA / context-memory conflict
JUICE (Li et al., 2025a)	ICML'25	Inference	Knowledge-conflict reasoning
Trust-Align (Song et al., 2025)	ICLR'25	SFT	RAG / knowledge-intensive QA
RAG with Conflicting Evidence (Wang et al., 2025b)	NAACL'25	Inference	RAG / conflicting evidence resolution
PA-RAG (Wu et al., 2025)	NAACL'25	SFT	RAG / knowledge-intensive QA
Inter-Agent			
PLAY2PROMPT (Fang et al., 2025b)	arXiv'25	Prompt	Tool-augmented agents
MASS (Zhou et al., 2026a)	arXiv'25	Architecture	MAS coordination
RECONCILE (Chen et al., 2024a)	ACL'24	Inference	Multi-model consensus reasoning
Debate or Vote? (Choi et al., 2025a)	NeurIPS'25	Inference	Decision protocol analysis
Voting or Consensus? (Kaesberg et al., 2025)	Findings ACL'25	Inference	Aggregation dynamics

continued on next page

Table 3 – continued from previous page

Method / Paper	Pub.	Primary Mitigation	Application Domain
CONSENSAGENT (Pitre et al., 2025)	Findings ACL'25	Prompt	Consensus stability
Multiagent Debate (Du et al., 2024)	ICLR'24	Inference	Factual QA / reasoning
Divergent Thinking (Liang et al., 2024)	EMNLP'24	Prompt	Logical reasoning
Breaking Mental Set (Liu et al., 2025b)	ICLR'25	Prompt	Mathematical / logical reasoning
ReAgent (Zhao et al., 2025a)	EMNLP'25	Inference	Multi-hop QA
Hidden Strength of Disagreement (Wu and Ito, 2025)	EMNLP'25	Inference	Adaptive MAS / diversity tradeoff
Demystifying Multi-Agent Debate (Zhu et al., 2026b)	arXiv'26	Inference	Debate dynamics
Mixture-of-Agents (Wang et al., 2025c)	ICLR'25	Architecture	Multi-model collaboration
DynaDebate (Li et al., 2026)	arXiv'26	Inference	Debate-based reasoning

D Additional Future Directions

We list two further directions complementing those in §8.

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.

Scaling Mitigation under Practical Deployment Constraints. Scaling mitigation from single-agent to multi-agent settings introduces challenges that go beyond technical accuracy: role mismatch, protocol ambiguity, and divergent local beliefs let local misalignment spread quickly through communication channels, while practical deployment further demands attention to communication cost, latency, controllability, and social compatibility. Future research should therefore extend single-agent mitigation primitives (reflection, memory management, conflict resolution) to distributed settings with shared tasks, evolving roles, and heterogeneous agents, and evaluate them not only on task success but also on oversight burden, cost-efficiency, and alignment with human values under realistic resource-constrained, human-in-the-loop deployments.