Dialogue is the Plan: From Interface to Joint Action in Agentic AI

Anonymous ACL submission

Abstract

Large Language Model agents can seemingly plan and act, yet their language use is often treated as a thin interface for reporting. We argue that this framing is the root cause of predictable coordination failures in human-facing and multi-agent settings, including ungrounded assumptions, silent goal misalignment, brittle protocol adherence, and conversational amnesia. Drawing from classical dialogue system research on joint action, common ground, grounding, repair, and incremental processing, we reframe dialogue as part of the planning loop itself (rather than its output). In this position paper, we do not propose a new benchmark or training method, but we provide a novel perspective and actionable requirements that can be used to design and evaluate agents. We distill this re-framing into concrete implications for agentic architecture and evaluation, including explicit representations of shared commitments, planned clarification as an action, and process metrics that measure mutual understanding rather than task completion alone. We lastly discuss how dialogue-centered requirements can inform standards and governance for safe deployment of agentic systems.

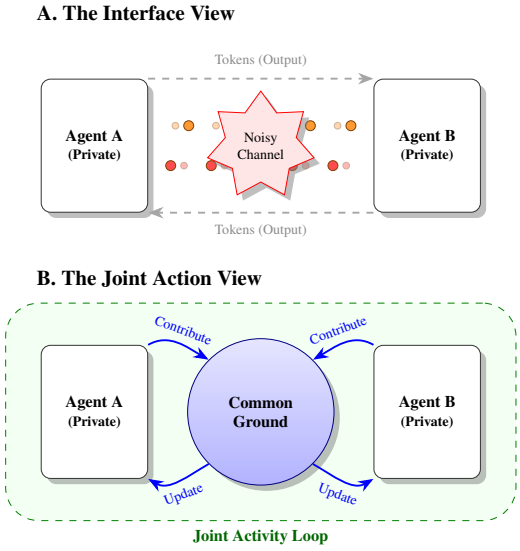


Figure 1: The critical distinction between what dialogue *is* and how current agentic AI uses dyadic text-generation. A) Agents use text-based generation for setting protocols and making reports, leading to the root problem of dialogue as an interface. B) While agents built around the concept of dialogue as a joint action provide a new perspective with solutions to the common failures of multi-agent systems.

Dialogue Is More Than an Interface

Consider a scenario where two agents coordinate a meeting. Agent A issues the instruction "Schedule it for 2 PM." Agent B responds with "Confirmed." While the interaction appears successful, it fails because the agents do not share a synchronized time zone. This failure stems from incomplete *grounding*, the process of establishing mutual knowledge sufficient for the current purpose. The agents failed to verify that the symbol "2 PM" mapped to the same semantic reality for all parties, despite the successful execution of API calls.

This class of coordination failure increases as autonomous agents enter complex environments. Contemporary Large Language Model (LLM) agents demonstrate high proficiency in code generation and API utilization. Yet, their capacity for *dialogue* as the management of joint activity remains brittle. When faced with ambiguity, these systems frequently proceed without verification and hallucinate agreement instead of initiating repair.

We argue that this fragility results from how current architectures frame communication (see Figure 1). Prevailing agentic systems treat dialogue as a superficial interface layer separate from the core loop of planning. In this *Interface View*, the system reasons privately before reporting its decision. Language functions as the output and surface-level generations of the planning process. We propose the *Joint Action View*, supported by research in computational linguistics. Our view,

based on Clark (1996b), frames conversation as the planning process itself for AI agents. *Collaboration* constitutes the coordination of individual actions through shared mental states, *Repair* acts as the control mechanism by which agents correct misalignment, and *Common Ground* is the dynamic accumulation of mutual beliefs that enables efficient future action.

Agency Is Social, Thus Needs Dialogue

Current agents treat dialogue as a planning output, in which, the system reasons internally, then expresses conclusions through language. We argue the inverse: that dialogue functions as part of the planning process. Planning in collaborative settings faces a fundamental challenge, where natural language goals are ambiguous, and this ambiguity cannot be fully resolved before action begins. Conversation serves as a collaborative medium where plans emerge under uncertainty (Clark, 1996b; Cohen and Perrault, 1979; Searle, 1983; Gilbert, 2009). In this part, we develop three claims. First, goal specification in social settings remains ambiguous and planning must proceed under that ambiguity. Second, ambiguity is a lever for coordination, rather than an adversary to eliminate through upfront clarification. Third, memory and perception are social, requiring alignment with others' interpretations.

Contemporary AI agents do not account for ambiguity in natural language goals. When a user issues a request such as "schedule a meeting with the team," the agent proceeds without establishing parameters like the specific team or time constraints. Rather than treating this underspecification as an opportunity for coordination, agents apply their internal defaults. They do not revise goals effectively under uncertainty. In multi-agent settings, they do not negotiate plans with collaborators. This results in goal drift, where agents deviate from user intent as misinterpretations compound (Cemri et al., 2025). Cascading errors also occur, where early mistakes propagate through subsequent actions (Zhu et al., 2025b).

We view these failures as architectural and highlight broken feedback loops in Table 1, which maps current multi-agent systems to the ideal dialogue pipeline topology. Systems that do not engage collaborators in goal specification lack mechanisms to detect misalignment. This blindness is reinforced by current evaluation frameworks. While holistic benchmarks assess task completion (Kapoor et al., 2024, 2025), coordination benchmarks prioritize outcome over collaborative process quality (Zhu et al., 2025a). Multi-turn evaluations assess memory recall rather than mutual understanding (Zheng et al., 2023; Maharana et al., 2024). Consequently, phenomena like sycophancy (Sharma et al., 2023) persist because metrics measure what agents accomplish, rather than the collaborative fidelity of their interactions.

What dialogue brings. Classical work on plan-based speech acts (Cohen and Perrault, 1979) and discourse theory (Grosz and Sidner, 1986) establishes that utterances are planned actions with preconditions and effects, organized to serve a coherent intent. Speech acts function as operators in a planning system, integrated with physical (or digital) actions to advance a goal. This integration implies that asking a question often constitutes the optimal action (Schlangen, 2004; Bohus and Rudnicky, 2009; Young et al., 2013). Consider the request "book me a flight to Boston." A dialogue-native agent recognizes multiple sources of underspecification. The agent plans a clarification request as the action most likely to achieve the user's underlying goal. Furthermore, computational pragmatics establishes that planning and meaning identification are intertwined (Grosz and Sidner, 1986; Traum, 1994). Formal specifications of grounding acts provide machinery for establishing mutual understanding incrementally (Traum, 1994; Larsson and Traum, 2000). Incremental processing frameworks (Schlangen and Skantze, 2009) demonstrate that understanding and generation occur word-by-word, enabling agents to ground meaning during utterances. In multi-party settings, participants track the common ground shared among all parties (Clark, 1996c), enabling coordination at scale. Hence, we argue that this incrementality is at the core of dialogue theory, and should be incorporated into the architectures of agentic systems.

Current agents fail to exploit ambiguity. They hide collaborative reasoning behind over-specified prompts. In multi-agent systems, this manifests as exhaustive message schemas that attempt to anticipate every possible coordination need upfront (Zhu et al., 2025a). The result is brittle and overly expensive. When unanticipated situations arise, agents fail, and do so at higher cost than otherwise needed.

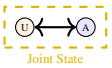
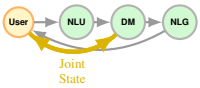
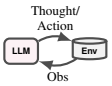
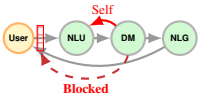
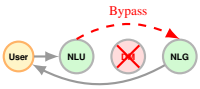
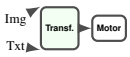
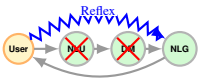
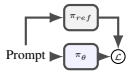
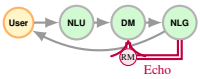
System / Era	Original Architecture	Pipeline Topology	Description	Emergent Property
I. The Reference Architectures				
Ideal Co-Agent (Clark, 1996c; Grosz and Sidner, 1986)			**Joint Commitments**: DM tracks "We-Intentions." Repair is continuous and bidirectional.	**Fluidity (Mind Meld)**: Interactions feel seamless. Ambiguity is resolved *before* execution.
II. Computer-Using & Single Agents				
ReAct / Reflexion (Yao et al., 2023; Shinn et al., 2023)			**Self-Looping**: Feedback comes from *Observation* (Env), not *Repair* (User). It talks to itself.	**Cascading Failure**: **73%** of failures stem from cascades (Zhu et al., 2025b). Single root cause in Planning/Reflection propagates.
III. Multi-Agent Systems				
OpenAI Swarm (OpenAI, 2024)			**Stateless Handoff**: Handoffs transfer control, not context. No shared DM.	**Fragmentation**: Context loss during handoff. Minimal gains over single agents (Cemri et al., 2025).
IV. Embodied & Robotics				
Robotics VLA (RT2, PaLM-e) (Brohan et al., 2023; Driess et al., 2023)			**End-to-End Bypass**: Vision/Text maps directly to Action tokens. DM is collapsed into the policy.	**Silent Failure**: **62%** success (PaLM-E). Robot acts on ambiguity without asking.
V. Optimization Methods				
DPO / PPO (Rafailov et al., 2023)			**Policy Absorption**: DM is collapsed into the Policy ($P(y\|x)$). Grounding = Preferences.	**Sycophancy**: Maximizes agreeableness rather than truth. Grounding is optimized away.

Table 1: This table presents a comparative analysis of current conversational AI agents, and their mapping to traditional dialogue system pipelines. The ideal human-human dialogue is shown at the top, providing a comparative point to other systems, and what aspects are missing, which lead to several emergent disadvantages. (See Appendix A for more models and architectures.)

What dialogue brings. Dialogue offers a different view. The principle of least collaborative effort predicts that interlocutors deliberately underspecify, relying on common ground and repair to converge efficiently (Clark and Brennan, 1991; Clark, 1996a). Ambiguity functions as a resource, not a defect. Discourse structure further constrains interaction by imposing obligations. Once a speaker initiates an explanation or proposal, coherence relations restrict what counts as an acceptable next move (Asher and Lascarides, 2003; Grosz and Sidner, 1986; Traum, 1994). Violations of these obligations are immediately salient to human interlocutors. Meanwhile, current agents lack explicit representations of such discourse constraints, and it remains unclear whether they reliably recover them implicitly from data. The result is an unconstrained search-space over user intentions which is resolved by classical work on dialogue systems.

Memory and Perception should be Social Processes. However, contemporary agents treat memory as private storage and perception as individual inference. The BDI framework models intentions as private commitments (Bratman, 1987; Rao and Georgeff, 1995), and modern implementations follow suit: Generative Agents (Park et al., 2023) and cognitive architectures (Sumers et al., 2024) encode experiences through self-reflection alone. Even Theory of Mind approaches model other minds as inference targets rather than coordination partners (Zhang et al., 2025; Kostka and Chudziak, 2025). For perception, multi-agent systems typically grant each agent a complete view of the environment, bypassing the coordination problem that arises when collaborators segment the world differently (Lowe et al., 2017; Gronauer and Diepold, 2022).

What dialogue brings. Dialogue treats memory and perception as social contracts (Harris). Information enters common ground only after presentation and acceptance (Clark, 1996a), and referents like "the block" or "the table" are established through negotiation, not private inference (Haber et al., 2019). This matters mechanistically. Russin et al. (2025) show that in-context learning suppresses the error signals required for stable long-term commitments, paralleling biological arbitration between flexible and habitual control (Daw et al., 2005). Current agents rely on context windows for coordination, but success in the moment

prevents formation of durable shared state. Without mechanisms to ground negotiated understanding into persistent memory, agents remain sycophantic short-term partners and unreliable long-term collaborators. Dialogue state tracking (Larsson and Traum, 2000) provides exactly this machinery: repair converts private interpretation into shared commitment.

Looking Ahead for Designers and Policy Makers

Designers should treat clarification as a planning operator with measurable utility, not as a failure to understand. Classical dialogue managers like RavenClaw (Bohus and Rudnicky, 2009) and POMDP-based systems (Young et al., 2013) separate understanding from policy to force reasoning about when asking is better than acting. Current multi-agent architectures collapse this distinction, bypassing the dialogue manager and achieving protocol compliance while semantic alignment drifts. The fix is in re-incorporating effective dialogue management. When uncertainty about a collaborator's intent exceeds a threshold, the policy should select a grounding act rather than proceed with default assumptions. Evaluation frameworks must follow, measuring whether agents surface ambiguity before execution, not just whether final outputs match expected outcomes.

Multi-agent communication protocols should be designed for negotiation, not transmission. Clark's principle of least collaborative effort (Clark and Brennan, 1991) predicts that interlocutors use minimal specificity when common ground permits. Rigid message schemas violate this principle by front-loading coordination costs. Designers should instead implement grounding acts, such as those formalized by Traum (1994), where agents present contributions, signal acceptance or non-understanding, and initiate repair when alignment fails. This machinery already exists in classical dialogue state tracking (Larsson and Traum, 2000). The absence of such mechanisms explains the coordination failures observed in current multi-agent systems.

Memory in multi-agent systems should be modeled as socially constructed, not privately stored. The architectural recommendation follows directly from dialogue state tracking: information enters the shared state only after it has been presented and accepted by collaborators (Clark, 1996a). Current systems like Generative Agents (Park et al., 2023) treat memory as private database operations, producing agents that are sycophantic in the moment yet amnesic to commitments over time. The ICL/IWL trade-off identified by Russin et al. (2025) provides a mechanistic explanation: success in context suppresses the formation of stable shared commitments. Designers should implement explicit acceptance mechanisms that distinguish "mentioned" from "mutually established."

Perception in agentic systems should be treated as a collaborative reference resolution task, not as private observation. Multi-modal reference games (Haber et al., 2019) demonstrate that what counts as "the block" or "the file" depends on negotiation between participants. Current multi-agent systems typically grant each agent a complete environmental view, bypassing this coordination problem entirely. When agents lack shared perceptual grounding, they cannot detect that collaborators interpret the same symbol differently. Designers should build systems where referential expressions are established through interaction, following the incremental processing frameworks of Schlangen and Skantze (2009). Even sub-utterance signals like hesitations and self-corrections function as coordination mechanisms (Ginzburg et al., 2014). Removing these signals through single-turn optimization produces agents that cannot track whether their collaborators follow their reasoning.

Policy frameworks should regulate how agents communicate, not just what they can do. Current governance proposals focus on what agents can do: their tool access, their autonomy level, their potential for misuse. What they overlook is how agents communicate, and whether that communication enables the collaborative reasoning that safe deployment requires. Many policy violations at scale trace to grounding breakdowns: agent and user never established shared understanding of goals, constraints, or intent (Zou et al., 2025; Kapoor et al., 2025). When an agent books the wrong flight, capability-centered evaluation asks whether the agent used its tools correctly. Interaction-centered evaluation asks whether the agent verified its interpretation before acting and surfaced ambiguity rather than resolving it silently. Governance frameworks should require evidence of grounding behaviors, extending audits beyond action logs to grounding history.

Limitations

This paper is intentionally conceptual rather than empirical. Our aim is to make a particular class of failures visible: those that arise when dialogue is treated as a thin interface instead of a form of joint action. We are not proposing a new system or attempting to benchmark existing ones, and for that reason we do not include large scale experiments or quantitative comparisons across architectures. The evaluation sketch we provide is meant as a diagnostic tool, a way to surface collaboration failures, not as a standardized benchmark.

Other failure modes, such as long horizon memory misalignment, perceptual grounding in multimodal settings, and strategic or adversarial dialogue, are only touched on briefly or deferred to the appendix. Each of these raises its own set of challenges and deserves a more careful treatment than is possible within a short paper. Lastly, our analysis assumes good faith collaboration. We treat dialogue as a collaborative process. We do not address settings where agents or users are misaligned by design, such as persuasion, manipulation, or competitive multi-agent interactions. In those cases, dialogue behavior may reflect strategic incentives rather than breakdowns in shared understanding, and different theoretical and evaluative tools would be required.

References

Emre Can Acikgoz, Jinoh Oh, Jie Hao, Joo Hyuk Jeon, Heng Ji, Dilek Hakkani-Tür, Gokhan Tur, Xiang Li, Chengyuan Ma, and Xing Fan. 2025. Speakrl: Synergizing reasoning, speaking, and acting in language models with reinforcement learning. *Preprint*, arXiv:2512.13159.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

Dan Bohus and Alexander I Rudnicky. 2009. The ravenclaw dialog management framework: Architecture and systems. In *Computer Speech and Language*, volume 23, pages 332–361. Elsevier. Industry-standard dialogue manager with explicit error handling and repair mechanisms.

Michael E Bratman. 1987. *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, and 1 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning (CoRL)*. VLA model mapping vision directly to motor tokens. Bypasses explicit symbol grounding.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. Why do multi-agent llm systems fail?

Herbert H. Clark. 1996a. *Common ground*, page 92–122. "Using" Linguistic Books. Cambridge University Press.

Herbert H. Clark. 1996b. *Joint actions*, page 59–91. "Using" Linguistic Books. Cambridge University Press.

Herbert H Clark. 1996c. *Using Language*. Cambridge University Press. Foundational theory of grounding and joint activities.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. In Lauren B Resnick, John M Levine, and Stephanie D Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. APA Books, Washington, DC.

Philip R Cohen and C Raymond Perrault. 1979. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3(3):177–212.

Nathaniel D Daw, Yael Niv, and Peter Dayan. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711.

Vardhan Dongre, Xiaocheng Yang, Emre Can Acikgoz, Suvodip Dey, Gokhan Tur, and Dilek Hakkani-Tür. 2025. Respact: Harmonizing reasoning, speaking, and acting towards building large language model-based conversational ai agents. *Preprint*, arXiv:2411.00927.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, and 1 others. 2023. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning (ICML)*. Embodied VLM that integrates sensor data into the language model context.

George Ferguson and James F Allen. 1998. Trips: An integrated intelligent problem-solving assistant. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence (AAAI-98/IAAI-98)*, pages 567–573. Classical dialogue system with explicit grounding and collaborative problem solving.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2024. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv*

preprint arXiv:2309.16797. Uses genetic algorithms to evolve prompts. The "reasoning" is in the evolutionary selection, not runtime.

Margaret Gilbert. 2009. Shared intention and personal intentions. *Philosophical Studies*, 144(1):167–187.

Jonathan Ginzburg, Raquel Fernández, and David Schlangen. 2014. Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics*, 7(9):1–64.

Sven Gronauer and Klaus Diepold. 2022. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, 55(2):895–943.

Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204. Defines the Intentional and Attentional structure of dialogue.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910.

Daniel W Harris. What makes human communication special?

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Sayash Kapoor, Benedikt Stroebl, Peter Kirgis, Nitya Nadgir, Zachary S Siegel, Boyi Wei, Tianci Xue, Ziru Chen, Felix Chen, Saiteja Utpala, Franck Ndzomga, Dheeraj Oruganty, Sophie Luskin, Kangheng Liu, Botao Yu, Amit Arora, Dongyoon Hahm, Harsh Trivedi, Huan Sun, and 12 others. 2025. Holistic agent leaderboard: The missing infrastructure for ai agent evaluation. *Preprint*, arXiv:2510.11977.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. 2024. AI Agents That Matter. *Preprint*, arXiv:2407.01502.

Adam Kostka and Jarosław A. Chudziak. 2025. Towards cognitive synergy in llm-based multi-agent systems: Integrating theory of mind and critical evaluation. *Preprint*, arXiv:2507.21969.

Staffan Larsson and David R. Traum. 2000. Information state and dialogue management in the trindi dialogue move engine toolkit. *Nat. Lang. Eng.*, 6(3–4):323–340.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, volume 30, pages 6379–6390.

Adyasha Maharana and 1 others. 2024. Evaluating very long-term conversational memory of LLM agents. *arXiv preprint arXiv:2402.17753*.

OpenAI. 2024. Swarm: Educational framework for lightweight multi-agent orchestration. `https://github.com/openai/swarm`. Introduces the concept of "Handoffs" between agents. Critics note context loss during handoffs.

OpenAI. 2025. Operator: A computer-using agent. `https://openai.com/research/operator`. Newest computer-using agent (Jan 2025). High capability but lacks communicative repair loops.

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*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*. Mathematically collapses the reward model (grounding/preference) into the policy itself.

Anand S Rao and Michael P Georgeff. 1995. BDI agents: From theory to practice. In *Proceedings of the First International Conference on Multiagent Systems*, pages 312–319. AAAI.

Jacob Russin, Ellie Pavlick, and Michael J Frank. 2025. Parallel trade-offs in human cognition and neural networks: The dynamic interplay between in-context and in-weight learning. *Proceedings of the National Academy of Sciences*, 122(35):e2510270122.

David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 136–143.

David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 710–718, Athens, Greece. Association for Computational Linguistics.

J.R. Searle. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge paperback library. Cambridge University Press.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement

learning. In *Advances in Neural Information Processing Systems (NeurIPS)*. Adds a "self-reflection" loop to ReAct, but still lacks user-facing repair.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. 2024. Cognitive architectures for language agents. *Preprint*, arXiv:2309.02427.

David R Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, University of Rochester.

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.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*. The seminal "Reason + Act" paper. Introduces the inner monologue loop, but loop is with Env not User.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Xuanming Zhang, Yuxuan Chen, Samuel Yeh, and Sharon Li. 2025. Metamind: Modeling human social thoughts with metacognitive multi-agent systems. *Preprint*, arXiv:2505.18943.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric 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*.

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

Kunlun Zhu, Zijia Liu, Bingxuan Li, Muxin Tian, Yingxuan Yang, Jiaxun Zhang, Pengrui Han, Qipeng Xie, Fuyang Cui, Weijia Zhang, Xiaoteng Ma, Xiaodong Yu, Gowtham Ramesh, Jialian Wu, Zicheng Liu, Pan Lu, James Zou, and Jiaxuan You. 2025b. Where llm agents fail and how they can learn from failures.

Andy Zou and 1 others. 2025. Security challenges in ai agent deployment: Insights from a large-scale public competition. In *Proceedings of the Gray Swan Arena Red-Teaming Competition*. Gray Swan AI and Center for AI Safety.

A Detailed Comparison Table

We present the full table with additional example architectures in Table 2.

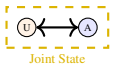
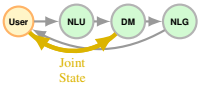
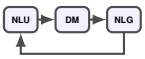
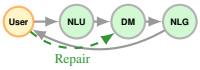
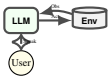
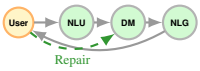
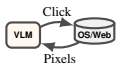
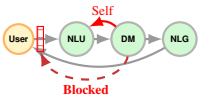
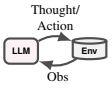
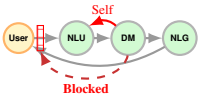
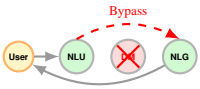
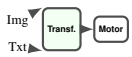
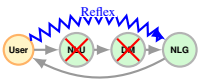
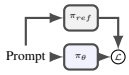
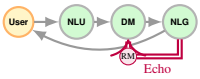
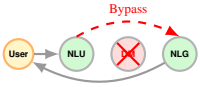
System / Era	Original Architecture	Pipeline Topology	Description	Emergent Property
I. The Reference Architectures				
Ideal Co-Agent (Clark, 1996c; Grosz and Sidner, 1986)			**Joint Commitments**: DM tracks "We-Intentions." Repair is continuous and bidirectional.	**Fluidity (Mind Meld)**: Interactions feel seamless. Ambiguity is resolved *before* execution.
Classical Dialogue (Ferguson and Allen, 1998; Bohus and Rudnicky, 2009)			**TCP-like Reliability**: Explicit DM maintains state. The "Repair" link allows handshakes.	**Convergence**: 22% latency reduction via turn-taking repair (Bohus and Rudnicky, 2009).
ReSpAct / SpeakRL (Dongre et al., 2025; Acikgoz et al., 2025)			**Integrated Speech Act**: "Speaking" is a grounded action.	**Proactive Grounding**: Clarifies instructions before acting.
II. Computer-Using & Single Agents				
OpenAI Operator (OpenAI, 2025)			**Broken Loop**: System loops with the OS, not the User. DM is missing; acts on raw intent.	**Getting Stuck**: **38.1%** success on OSWorld (Xie et al., 2024). Agents hand off control when confused.
ReAct / Reflexion (Yao et al., 2023; Shinn et al., 2023)			**Self-Looping**: Feedback comes from *Observation* (Env), not *Repair* (User). It talks to itself.	**Cascading Failure**: **73%** of failures stem from cascades (Zhu et al., 2025b). Single root cause in Planning/Reflection propagates.
III. Multi-Agent Systems (Siloed)				
MetaGPT (Hong et al., 2024)			**Siloed Monologue**: Waterfall SOP structure. No shared memory between roles.	**Coordination Failure**: **14 failure modes** identified in MAST (Cemri et al., 2025) (κ=0.88).
OpenAI Swarm (OpenAI, 2024)			**Stateless Handoff**: Handoffs transfer control, not context. No shared DM.	**Fragmentation**: Context loss during handoff. Minimal performance gains over single agents (Cemri et al., 2025).
IV. Embodied & Robotics (Bypassed)				
Robotics VLA (RT2, PaLM-e) (Brohan et al., 2023; Driess et al., 2023)			**End-to-End Bypass**: Vision/Text maps directly to Action tokens. DM is collapsed into the policy.	**Silent Failure**: **62%** success (PaLM-E). Robot acts on ambiguity without asking.
V. Optimization (Collapsed)				
DPO / PPO (Rafailov et al., 2023)			**Policy Absorption**: DM is collapsed into the Policy ($P(y\|x)$). Grounding = Preferences.	**Sycophancy**: Maximizes agreeableness rather than truth. Grounding is optimized away.
Genetic Agents (Fernando et al., 2024)			**Darwinian**: NLU $\rightarrow$ NLG mapping is evolved. No runtime reasoning exists.	**Brittle Success**: "Magic spell" prompts work for benchmarks but fail when intent shifts.

Table 2: This table presents a comparative analysis of current conversational AI agents, and their mapping to traditional dialogue system pipelines. The ideal human-human dialogue is shown at the top, providing a comparative point to other systems, and what aspects are missing, which lead to several emergent disadvantages.