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

Mert İnan, Malihe Alikhani, Anthony Sicilia
Northeastern University, Boston, MA, USA
West Virginia University, Morgantown, WV, USA
{inan.m, alikhani.m}@northeastern.edu, anthony.sicilia@mail.wvu.edu

Abstract

Large Language Model agents can seemingly plan and act, yet their language use is often treated primarily as an interface for instructing actions and reporting results. We argue that this framing is one important cause of recurrent coordination failures in human-facing and multi-agent settings, including ungrounded assumptions, silent goal misalignment, brittle protocol adherence, and failures to maintain or update shared dialogue state over time, a limitation previously linked to the absence of explicit common ground tracking in collaborative systems (Geib et al., 2022). Drawing from classical dialogue system research on joint action, common ground, grounding, repair, and incremental processing, we re-frame dialogue as part of the planning loop itself (rather than its output). We distill this re-framing into concrete implications for agentic architecture and evaluation, including explicit representations of shared commitments, clarification as a first class action available to the policy, and process metrics that approximate grounding behavior, repair, and commitment formation rather than task completion alone. We lastly discuss how dialogue-centered requirements can inform standards and governance for safe deployment of agentic systems.

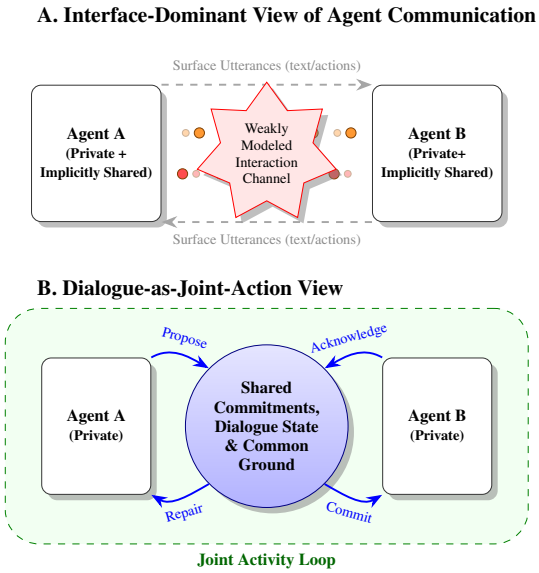


Figure 1: A schematic contrast between treating dialogue primarily as an interface for action execution versus as a mechanism for coordinating joint action. A) Agents use text-based generation for setting protocols and making reports, which can lead to coordination failures when grounding and shared state are weakly modeled. B) While agents built around the concept of dialogue as a joint action illustrate a perspective that suggests design directions for addressing common coordination failures of multi-agent systems. Some current systems may exhibit aspects of both views.

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 have presumed, incorrectly, that they are in a common time zone. This can be analyzed as a failure of *grounding*, the process of establishing mutual knowledge sufficient for the current purpose. The agents have executed API calls to schedule a meeting for "2 PM" without coordinating on the implicit time zone.

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. Even when faced with ambiguity or underspecification, these systems proceed as though interpretations are unproblematic, acting as if understanding and agreement have been reached, without considering evidence of whether this is the case. Prior work argues that planning and communication must be coupled in collaborative systems (Geib et al., 2022; Cohen, 2020), a perspective that informs and motivates our own suggestions.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 770–779

Addressing this fragility requires fundamentally rethinking how current agent architectures and evaluations 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*, many systems perform substantial latent reasoning before producing a surface utterance or action. Language functions as the output or summarization of an internal planning process. In contrast, we propose the *Joint Action View*, drawing on dialogue theory, computational linguistics, and adjacent fields.

Our view, based on Clark (1996b), frames conversation as the planning process itself for AI agents. In this setting, *Collaboration* constitutes the coordination of individual actions through public commitments or information state; *Repair* acts as the control mechanism by which agents correct misalignment; and, *Common Ground* is the evolving set of publicly grounded assumptions and commitments sufficient for the current purpose, which ultimately enables efficient future action.

Agency Is Social, Thus Needs Dialogue

Current agents treat dialogue as a planning output, in which, the system "reasons" internally,[1] then expresses conclusions through language. We argue for a complementary reversal of emphasis: 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. Thus, 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 such that it can reduce collaborative effort when repair is available—it is not solely an adversary to eliminate through upfront clarification. **Third**, memory and perception relevant to collaboration must be socially coordinated, requiring alignment with others' interpretations. These claims position dialogue as a space to clarify plans and future intentions—as opposed to a space for negotiating plans, reporting on them, or clarifying references—a view consistent with recent architectural work on human-AI teamwork (Geib et al., 2022). However, the treatment of dialogue as a space for plan formation is largely unimplemented in modern agents.

Many contemporary agents bypass collaborative planning during generation Rather than treating underspecification as an opportunity for coordination, agents often rely on implicit defaults, prompt conventions, or latent priors. They often revise goals weakly under uncertainty, especially without explicit grounding mechanisms. In multi-agent settings, they do not negotiate plans with collaborators, which can contribute to goal drift, where agents deviate from user intent as misinterpretations compound (Cemri et al., 2025). Cascading errors are widely observed, where early mistakes propagate to subsequent actions (Zhu et al., 2025b).

We view these failures as partly architectural and partly evaluative 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), many coordination benchmarks still emphasize outcomes more than observable grounding behavior (Zhu et al., 2025a). Multi-turn evaluations assess memory recall rather than mutual understanding (Zheng et al., 2023; Maharana et al., 2024). Even dialogue-specific comprehension benchmarks require reasoning beyond surface extraction (Sun et al., 2019), yet agent evaluations rarely measure whether understanding was actually mutual. This evaluative emphasis may help obscure phenomena like sycophancy (Sharma et al., 2023; Sicilia et al., 2025), deference, or premature agreement, 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.[2] Speech acts function as operators in a

[1] Many modern agents do not instantiate planning in the classical sense (Kambhampati et al., 2024). The argument here does not depend on whether internal reasoning conforms to such formulations, but adds a distinct shortfall.

[2] Grosz and Sidner (1986) treat a conversation as organized around goals and subgoals (intentional structure), arranged in a hierarchy (dominance and ordering relations), and interpreted relative to what is currently in focus (attentional state). While

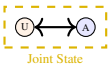
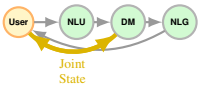
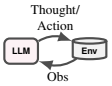
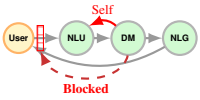
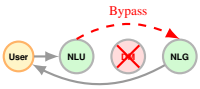
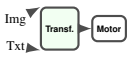
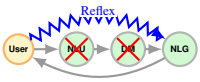
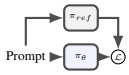
System / Era	Original Architecture	Pipeline Topology	Description	Emergent Property
I. The Reference Architectures				
Ideal Co-Agent (Clark, 1996c; Grosz and Sidner, 1986)			**Joint Commitments**: DM tracks "We-Intentions." Repair is continuous and bidirectional.	**Fluidity (Mind Meld)**: Interactions 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 `https://github.com/openai/swarm`			**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) (Zitkovich 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: A comparative analysis of some current conversational AI agents or related methods, and their mapping to the traditional dialogue system pipeline. The ideal architecture is shown at the top, providing comparison to other systems. Missing aspects lead to several emergent disadvantages (see Appendix A for more comparisons).

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). Computational models have demonstrated that cognitive representations of user intent can be integrated with reinforcement learning to plan targeted clarification actions (Khalid et al., 2020). Furthermore, works in discourse theory and plan-based dialogue models argue that planning and meaning interpretation are tightly coupled: understanding an utterance involves inferring the speaker's underlying goals, while producing an utterance involves choosing actions that advance one's goals (Grosz and Sidner, 1986; Traum, 1994). Formal specifications of grounding acts provide machinery for establishing mutual understanding incrementally (Larsson and Traum, 2000), along with incremental processing frameworks at a token-by-token level (Schlangen and Skantze, 2009). This enables grounding and repair during the utterance itself and motivates real-time integration. Deployed systems that include discourse coherence into modular task-oriented architectures have shown measurable improvements in user satisfaction (Atwell et al., 2024), suggesting that the dialogue-centered approach we advocate is practically viable. Meanwhile, in embodied conversation, speech and gesture must also be coordinated as joint actions, so planning must account for context across modalities (Cassell et al., 2000).

Current agent designs often suppress ambiguity management. Empirical evidence confirms this gap, where GPT-4 generates 60% fewer correct disambiguations compared to humans (Liu et al., 2023), and pragmatic reasoning remains systematically absent from grounded language systems (Fried et al., 2023). Even in controlled code generation settings, ambiguity in user goals degrades task success unless agents interactively resolve it (Inan et al., 2025b). In multi-agent systems, this weakness manifests through exhaustive message schemas that try to pre-encode coordination assumptions upfront (Zhu et al., 2025a). The result is brittle and overly expensive. When unanticipated situations arise, agents fail at a higher cost.

What dialogue brings. Dialogue offers a different view. The principle of least collaborative ef-

not formulated in planning terms, this structure is compatible with viewing discourse segments as goal-directed actions whose applicability depends on discourse context and whose effects update shared attentional and intentional state.

fort helps explain why interlocutors often leave information underspecified when repair is cheap and common ground is sufficient, relying on them to converge efficiently (Clark and Brennan 1991; 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). Violations of these obligations are immediately salient to human interlocutors. Discourse-theoretic approaches that organize dialogue structure through coherence relations have shown that surfacing enough information for errors and misunderstandings to become visible can improve task success without requiring explicit ambiguity tracking or formal grounding (Inan et al., 2025a; Atwell et al., 2024), pointing to how classical dialogue theory can productively inform LLM-based agent research. Meanwhile, current agents lack explicit representations of such discourse constraints, and it is unclear whether they reliably recover them implicitly from data. The result is an unconstrained search-space over user intentions—one which could be readily resolved by classical ideas from dialogue theory.

Contemporary agents treat memory as private storage and perception as individual inference. Indeed, early BDI formulations center private intentions (Bratman, 1987; Rao and Georgeff, 1995) and modern implementations of agents often follow suit.[3] This reflects a broader pattern whereby language understanding research has been constrained by treating language as separable from the physical and social contexts that give it meaning (Bisk et al., 2020). Generative Agents (Park et al., 2023) and cognitive architectures (Sumers et al., 2024) encode experiences through self-reflection alone. These architectures incorporate social signals only as inputs to individual memory, without representing memory as a shared, jointly updated state founded in explicit representations of common ground or commitments. Even Theory-of-Mind approaches primarily model other agents as inference targets, i.e., objects of belief and intention prediction (Zhang et al., 2025; Kostka and Chudziak, 2025), so that coordination is achieved indirectly through improved prediction. This fails to explicitly utilize shared or jointly constructed representation states, and therefore, leaves the agent prey to apparent uncertainties in the accuracy of belief inference (Sicilia et al., 2024; Sicilia and Alikhani, 2025). For perception, many multi-agent systems rely on centralized training signals with access to global state (Lowe et al., 2017; Gronauer and Diepold, 2022). However, because execution remains decentralized, agents must still act under partial and potentially misaligned observations, without explicit mechanisms for reconciling how collaborators segment the world. As a result, coordination is mediated through implicit policy alignment during training, which compounds uncertainty that could instead be resolved through explicit agreement at runtime.

What dialogue brings. On many grounding accounts of dialogue, information enters the working common ground through public presentation and sufficient evidence of uptake (Clark, 1996a), not through one agent's private inference alone. For example, in visually grounded dialogue, human interlocutors progressively compress descriptions of visual information across dialogue rounds, leveraging partner-specific expressions whose interpretation depends on shared interaction history (Haber et al., 2019). This points to a concrete design requirement for agents: context-sensitive interpretation is not enough unless the outcomes of clarification, correction, and acknowledgment are recorded in an explicit public state. Information state approaches (Larsson and Traum, 2000) can provide this missing split by distinguishing private beliefs and agendas from shared commitments, while grounding models specify how repair and acknowledgment update that shared record (Traum, 1994). By contrast, LLM attention mechanisms operating over context-windows are less explicit and perhaps too ephemeral to support collaborative commitments across interactions. This is evidenced by Imai et al. (2025), confirming that vision language models fail to exhibit grounding behaviors.

Looking Ahead for Designers and Policy Makers

Designers should treat clarification as a planning operator with measurable utility. Classical dialogue managers like RavenClaw (Bohus and Rudnicky, 2009) and POMDP-based systems (Young et al., 2013) separate understanding from action policies to force reasoning about when asking is better than (physically) acting. Current multi-agent

[3] It should be noted that these frameworks have also been extended to social commitments and joint intentions (Rao and Georgeff, 1992; Castelfranchi, 1995), but this view is not frequently taken in modern implementations of agents.

architectures collapse this distinction. The fix is in re-incorporating effective dialogue management. When uncertainty about a collaborator's intent exceeds a threshold, the agent's policy should select a grounding act rather than proceed with default assumptions. A key insight from traditional dialogue system architectures is that this may require explicit separation of "understanding" and "acting" components, or at least, treating understanding as a first-class action. Evaluation frameworks must follow, measuring whether agents surface ambiguity before execution, in addition to whether final outputs match expected outcomes. This principle connects to a broader body of work in which deliberate slowdowns and interruptions to automatic behavior have been shown to improve reflection, reduce errors, and support more mindful engagement across dialogue systems (Inan et al., 2025a), human-AI interaction (Chen and Schmidt, 2024), social media platforms (Ruiz et al., 2024), and technology-mediated nudging (Caraban et al., 2019). These findings suggest that the costs of clarification and repair, which current agent architectures treat as inefficiencies to minimize, can be productive.

Multi-agent communication protocols should be designed for negotiation. Clark and Brennan's principle of least collaborative effort predicts that interlocutors use minimal specificity when common ground permits. Rigid message schemas and exhaustive prompts violate this principle by front-loading coordination costs. Designers should instead implement grounding acts, where agents present contributions, signal acceptance or non-understanding, and initiate repair when alignment fails. The machinery which already exists in classical dialogue state tracking (Larsson and Traum, 2000), can be an example model.

Memory in multi-agent systems should be modeled as socially constructed. Shared memory in multi agent systems should distinguish private memory from publicly-grounded commitments. The architectural recommendation follows directly from classical dialogue state tracking (e.g., information states) that 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 and amnesic to commitments over time. Designers should implement explicit acceptance mechanisms that distinguish what is *mentioned* or *inferred* from what is *mutually established*.

Perception in agentic systems should be treated as a collaborative reference resolution task. Multimodal reference games (Haber et al., 2019) demonstrate that accurate co-reference to mutually perceived objects depends on negotiation between participants. Current multi-agent systems often learn under a complete environmental view, and infer global states at runtime, bypassing this coordination problem entirely. However, without this coordination, it is difficult to reliably detect when collaborators may 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), and nonverbal cues such as facial displays of uncertainty have been shown to be interpretable by human interlocutors in both human and agent interactions (Oh and Stone, 2007). Removing these signals through single-turn optimization produces agents that cannot track whether their collaborators follow their reasoning.

Policy frameworks should evaluate communicative behavior alongside capabilities. 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: the 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, while in fact, the error originated upstream. Interaction-centered evaluation instead asks whether the agent verified its interpretation before acting and surfaced ambiguity rather than resolving it silently. The demonstrated capacity of models to produce fluent yet misleading content (Huang et al., 2022) and the systemic nature of model deployment (Bommasani et al., 2022) reinforces the need for governance frameworks that evaluate communicative grounding processes, extending audits beyond action logs to grounding history.

Acknowledgement

We thank Matthew Stone for his guidance and mentorship throughout this work. His expertise in dialogue theory and computational models of communication significantly shaped our thinking and improved this paper.

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.

Katherine Atwell, Mert Inan, Anthony B. Sicilia, and Malihe Alikhani. 2024. Combining discourse coherence with large language models for more inclusive, equitable, and robust task-oriented dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3538–3552, Torino, Italia. ELRA and ICCL.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735. Association for Computational Linguistics.

Dan Bohus and Alexander I. Rudnicky. 2009. The RavenClaw dialog management framework: Architecture and systems. *Computer Speech Language*, 23(3):332–361.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, and 95 others. 2022. On the opportunities and risks of foundation models. *Preprint*, arXiv:2108.07258.

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

Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15. ACM.

Justine Cassell, Matthew Stone, and Hao Yan. 2000. Coordination and context-dependence in the generation of embodied conversation. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 171–178, Mitzpe Ramon, Israel. Association for Computational Linguistics.

Cristiano Castelfranchi. 1995. Commitments: From individual intentions to groups and organizations. In *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)*, pages 41–48, San Francisco, CA. AAAI Press.

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?

Zeya Chen and Ruth Schmidt. 2024. Exploring a behavioral model of "positive friction" in human-AI interaction. In *HCII 2024*, volume 14713 of *LNCS*, pages 3–22. Springer.

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.

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. 2020. Back to the future for dialogue research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13514–13519.

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

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)*.

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/Innovative Applications of Artificial Intelligence*, AAAI '98/IAAI '98, page 567–572, USA. American Association for Artificial Intelligence.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-referential self-improvement via prompt evolution. *Preprint*, arXiv:2309.16797.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12619–12640, Singapore. Association for Computational Linguistics.

Christopher Geib, Denson George, Baber Khalid, Richard Magnotti, and Matthew Stone. 2022. An integrated architecture for common ground in collaboration. In *Proceedings of the Tenth Annual Conference on Advances in Cognitive Systems*, Arlington, VA.

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.

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.

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*.

Kung-Hsiang Huang, Fatemehsadat Mireshghallah, Swabha Mitra, and Kathleen McKeown. 2022. Faking fake news for real fake news detection: Propaganda-loaded training data generation. *arXiv preprint arXiv:2203.05386*.

Saki Imai, Mert Inan, Anthony B. Sicilia, and Malihe Alikhani. 2025. Measuring how (not just whether) VLMs build common ground. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 441–451, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Mert Inan, Anthony Sicilia, Suvodip Dey, Vardhan Dongre, Tejas Srinivasan, Jesse Thomason, Gokhan Tur, Dilek Hakkani-Tür, and Malihe Alikhani. 2025a. Better slow than sorry: Introducing positive friction for reliable dialogue systems.

Mert Inan, Anthony Sicilia, Alex Xie, Saujas Vaduguru, Daniel Fried, and Malihe Alikhani. 2025b. Identifying and interactively refining ambiguous user goals for data visualization code generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Paul Saldyt, and Anil B Murthy. 2024.

Position: Llms can't plan, but can help planning in llm-modulo frameworks. In *Forty-first International Conference on Machine Learning*.

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.

Baber Khalid, Malihe Alikhani, and Matthew Stone. 2020. Combining cognitive modeling and reinforcement learning for clarification in dialogue. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4417–4428, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore. Association for Computational Linguistics.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6382–6393, Red Hook, NY, USA. Curran Associates Inc.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.

Insuk Oh and Matthew Stone. 2007. Understanding RUTH: Creating believable behaviors for a virtual human under uncertainty. In *Digital Human Modeling, ICDHM 2007*, volume 4561 of *Lecture Notes in Computer Science*, pages 443–452, Berlin, Heidelberg. Springer.

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, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Anand S. Rao and Michael P. Georgeff. 1992. Social plans: A preliminary report. In *Decentralized AI 3 – Proceedings of the Third European Workshop on Modelling Autonomous Agents and Multi-Agent Worlds (MAAMAW-91)*, pages 57–76, Amsterdam. Elsevier/North-Holland.

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.

Nicolas Ruiz, Gabriela Molina León, and Hendrik Heuer. 2024. Design frictions on social media: Balancing reduced mindless scrolling and user satisfaction. In *Proceedings of Mensch und Computer 2024*, pages 442–447. ACM.

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)*.

Anthony Sicilia and Malihe Alikhani. 2025. Evaluating theory of (an uncertain) mind: Predicting the uncertain beliefs of others from conversational cues. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8007–8021, Vienna, Austria. Association for Computational Linguistics.

Anthony Sicilia, Mert Inan, and Malihe Alikhani. 2025. Accounting for sycophancy in language model uncertainty estimation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7866–7881, Albuquerque, New Mexico. Association for Computational Linguistics.

Anthony Sicilia, Hyunwoo Kim, Khyathi Chandu, Malihe Alikhani, and Jack Hessel. 2024. Deal, or no deal (or who knows)? forecasting uncertainty in conversations using large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11700–11726, Bangkok, Thailand. Association for Computational Linguistics.

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

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

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)*.

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.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, and 35 others. 2023. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR.

Andy Zou, Maxwell Lin, Eliot Krzysztof Jones, Micha V. Nowak, Mateusz Dziemian, Nick Winter, Valent Nathanael, Ayla Croft, Xander Davies, Jai Patel, Robert Kirk, Yarin Gal, Dan Hendrycks, J Zico Kolter, and Matt Fredrikson. 2025. Security challenges in AI agent deployment: Insights from a large scale public competition. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Detailed Comparison Table

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

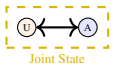
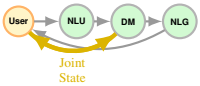
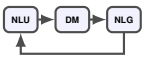
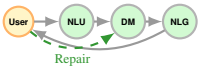
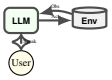
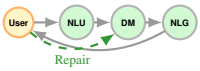
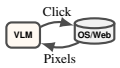
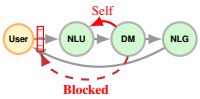
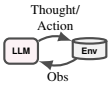
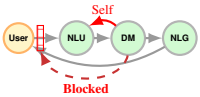
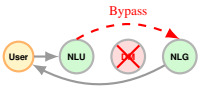
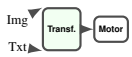
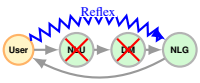
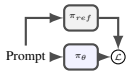
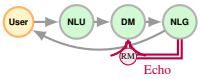
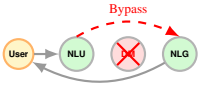
System / Era	Original Architecture	Pipeline Topology	Description	Emergent Property
I. The Reference Architectures				
Ideal Co-Agent (Clark, 1996c; Grosz and Sidner, 1986)			**Joint Commitments**: DM tracks "We-Intentions." Repair is continuous and bidirectional.	**Fluidity (Mind Meld)**: Interactions 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 https://openai.com/research/operator			**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 https://github.com/openai/swarm			**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) (Zitkovich 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., 2023)			**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.