Collaborate, Deliberate, Evaluate: How LLM Alignment Affects Coordinated Multi-Agent Outcomes

Extended Abstract

Abhijnan Nath
Colorado State University
Fort Collins, CO, USA
abhijnan.nath@colostate.edu

Carine Graff
Colorado State University
Fort Collins, CO, USA
carine.graff@colostate.edu

Nikhil Krishnaswamy
Colorado State University
Fort Collins, CO, USA
nkrishna@colostate.edu

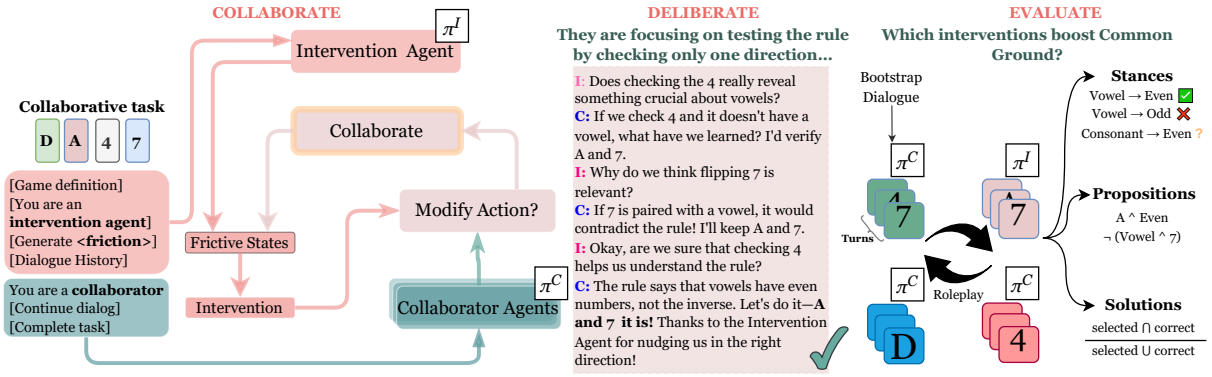

Figure 1: [L]: COLLABORATOR AGENTS complete tasks with an INTERVENTION AGENT in the loop. [C]: Sample collaborative roleplay from DeliData Wason Card task [8]. [R]: Common ground convergence and task outcomes with interventions provided by different agents.

ABSTRACT

Using the theoretical lens of the *modified-action MDP*, we show that common alignment techniques that are typically developed under single-user settings do not account for the dynamics of long-horizon multi-party interactions. We use a roleplay simulation methodology to quantify how AI partner interventions affect the trajectory of collaborative task dialogues. We show that interventions that are robust to action modification significantly outperform standard alignment in collaborative task support.

KEYWORDS

Multi-agent Coordination; Collaborative Problem Solving; Modified-Action MDP; Roleplay Simulation

ACM Reference Format:
Abhijnan Nath, Carine Graff, and Nikhil Krishnaswamy. 2026. Collaborate, Deliberate, Evaluate: How LLM Alignment Affects Coordinated Multi-Agent Outcomes: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/UQPO8536

This material is based in part upon work supported by Other Transaction award HR00112490377 from the U.S. Defense Advanced Research Projects Agency (DARPA) and by award DRL 2454151 from the U.S. National Science Foundation (NSF). Portions of this work were performed on the Colorado State University Data Science Research Institute high-performance computer *Riviera*. An extended version of this paper can be found at https://arxiv.org/abs/2509.05882 [14].

This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/UQPO8536

1 INTRODUCTION

Large Language Models (LLMs) are increasingly being integrated into "agentic" pipelines that help users achieve goals and solve problems. In doing so, they may interact and collaborate with *multiple* humans or other AIs. *Multi-party collaborations* challenge optimality assumptions; groups frequently succumb to *belief misalignment* and breakdown of *common ground* [1, 22]. We examine this problem with *intervention agents* designed to insert **friction** [7, 13, 16, 17] in collaborative problem solving, which plays a crucial role in successful human collaborations [5, 12, 20]. We perform a novel analysis of collaborative dynamics based on a modified-action MDP (MAMDP; [11]), and show that common LLM alignment methods lose optimality guarantees here. We use *roleplay simulation* to assess how different alignment techniques support both common ground construction and task solution correctness (Fig. 1). We experimentally show that interventions that are robust to action modification correlate with productive belief revision in multi-agent settings, benefit common ground convergence, and improve task outcomes.

2 PROBLEM FORMULATION

We define **common ground**, **frictive state**, and **friction interventions** as in Nath et al. [13]. An INTERVENTION AGENT is an LLM aligned for friction interventions in a multi-party dialogue to strategically resolve frictive states between collaborators.

In real-world multi-party collaborations, a single agent's utterance may not directly change the beliefs, perspectives, or assumptions of other participants; it may be resisted or interpreted

DeliData

	Method	Coarse Acc.	Fine Acc.	NCCG	Perf. Gain	CoM Rate
Standard	SFT	0.355	0.806	0.204	0.244	0.260
	PPO	0.409	0.767	0.180	0.183	**0.322**
	BC	0.369	0.812	**0.210**	0.239	0.267
	DPO	0.418	0.831	0.209	0.243	0.264
	IPO	0.352	0.825	0.205	0.246	0.288
	FAAF	**0.485**	**0.851**	0.201	**0.260**	0.270
MAMDP	SFT	0.283	0.702	0.178	0.143	0.310
	PPO	0.382	0.763	0.181	0.191	0.304
	BC	0.474	0.809	0.179	0.236	0.302
	DPO	0.428	0.794	**0.201**	0.224	0.276
	IPO	0.391	0.774	0.192	0.197	0.272
	FAAF	**0.526**	**0.844**	0.196	**0.250**	**0.329**

WTD

	Method	Final CG	Adjusted CG	Incorrect %
Standard	SFT	4.267	3.571	12.407
	PPO	3.778	3.252	**6.966**
	BC	5.241	4.805	9.406
	DPO	**5.714**	**4.912**	16.649
	IPO	3.822	3.294	14.009
	FAAF	5.143	4.584	7.111
MAMDP	SFT	3.920	3.490	9.898
	PPO	5.160	4.504	13.361
	BC	4.167	3.837	6.490
	DPO	5.760	5.329	8.440
	IPO	4.160	3.635	**6.156**
	FAAF	**8.300**	**7.819**	7.837

Table 1: Performance comparison across intervention agents. Relevant metrics vary by dataset, including Coarse-grained Accuracy, Fine-grained Accuracy, Normalized Cumulative Common Ground, Performance Gain, Change-of-Mind Rate, Final CG size, accuracy-Adjusted CG size, and mean error rate (Incorrect %).

by others conditioned upon what they already perceive or believe [3, 4, 6, 16, 25]. Standard Bellman-optimal action policies assume a direct mapping from action to state change, which breaks down when the application of action to state change is mediated by other agents. To address this, we adopt the Modified-Action MDP (MAMDP) framework, which *explicitly* models how interventions are transformed before influencing the collaborative dialogue. Bellman-optimal policies that solve the standard MDP structure underlying what is actually an MAMDP [11] lead to suboptimal outcomes. We show that this same suboptimality also applies to LLMs trained in such settings. We then validate this insight empirically, highlighting the importance of accounting for action transformation when designing alignment objectives for LLM-based agents.

Formally, an MAMDP consists of a 6-tuple $\mathcal{M}_f = (\mathcal{S}, \mathcal{A}, P_S, P_A, R, \gamma)$, or equivalently, the 5-tuple of a standard MDP with additional parameter P_A. Now assume a INTERVENTION AGENT π^I_θ (an LLM with parameters θ). $P_A(a|\pi^I, s)$ represents the probability that π^I selects action a in state s. Additionally assume a set of COLLABORATOR AGENTS π^C, which in this paper is another LLM optimized to be a robust generator of human-like utterances and actions.

Current algorithms like DPO [19] and IPO [2] satisfy Bellman optimality conditions and have policy structures that retain the optimal policy formulation. We can show how they are suboptimal for collaborative settings because they disregard modifications made to the action space by π^C, and RL policies lose optimality guarantees when their actions are modified [11].

THEOREM 1 (Ψ-PREFERENCE OPTIMIZATION IN COLLABORATIVE MAMDPS). *Let $\Psi : [0, 1] \to \mathbf{R}$ be any non-decreasing function and $\beta > 0$ be a temperature parameter. Let $P_A(a|s, \pi^I) = \sum_{a' \in A} \pi^I(a'|s) \cdot \pi^C(a|s, a')$, and represent modifications to the probability distribution over the action space by a collaborator policy π^C, and let π^I be an* INTERVENTION AGENT *policy trained via Ψ-preference optimization in a collaborative MAMDP $\mathcal{M}_f = (\mathcal{M}, P_A)$ with MDP $\mathcal{M}$ and P_A following [11]'s definition. π^I satisfies Eq. 1:*

$$\pi^I(a|s) = \frac{\exp(Q^I(s, a)/\beta)}{\sum_{a'} \exp(Q^I(s, a')/\beta)} \tag{1}$$

where Q^I satisfies the Bellman optimality equation for the underlying MDP $\mathcal{M}$. Thus π^I is optimal only when actions are sampled without modification. The Bellman-optimality of ΨPO-aligned π^I disregards the collaborator π^C's modifications. For MAMDPs with LLMs, this unifies [18]'s derivation of DPO in the token MDP with [11]'s proposition that Bellman-optimal policies do not consider action modifications, and extends it to ΨPO/IPO.

3 EXPERIMENTS AND RESULTS

We largely follow the roleplay methodology given in Nath and Krishnaswamy [15] for data generation and evaluation, except that instead of a *single* π^C model roleplaying all collaborator agents, each collaborator was simulated by a *distinct* instance of GPT-4o. We evaluate in two collaborative tasks: the Wason Card Selection task [26] as captured in **DeliData** [8], and the **Weights Task** [9]. We trained intervention agents π^I using **Supervised fine-tuning (SFT)**, **DPO** [19], **IPO** [2], **PPO**; [21], **behavior cloning** [23], and **FAAF** [13], an alignment method specifically designed for friction interventions in collaborative tasks. We used `Meta-Llama-3-8B-Instruct` as the base model for all trained intervention agents. We evaluated under *standard* MAMDP settings, where action modification by the collaborators *may* happen stochastically, and under *explicit* MAMDP settings, where action modification *does* happen deterministically due to the roleplay prompt.

Results are given in Table 1. The FAAF alignment method, specifically designed for friction interventions, indeed outperforms other methods on facilitating a balance of group common ground convergence and correct task solutions. Additionally, it demonstrates robustness to collaborator action modification or resistance to belief update. The MAMDP condition thus models a realistic kind of coordinated reasoning—one where alignment emerges through negotiation, debate, and gradual stabilization.

4 CONCLUSION

Our study emphasizes that in multiagent collaboration, as in human-human collaboration, the collaborative *process* is as important as the *outcome*. Our findings suggest that friction, rather than obstructing alignment, can paradoxically deepen it by promoting iterative clarification. To perform a controlled, high-throughput evaluation, we used an LLM roleplay methodology. The next logical step is studying agent interventions with real human subjects, e.g., by reproducing the studies of the Wason task [8] or Weights Task [10] with the inclusion of a demonstrably-reliable friction intervention agent in a real-time common ground tracking system, e.g., [24]. We also hope this study raises awareness of the utility of "friction" to prompt deliberation and accountable decision making in multiagent and human-AI systems, and shows that slower AI interactions can also be positive ones.

REFERENCES

[1] Nicholas Asher and Anthony Gillies. 2003. Common Ground, Corrections, and Coordination. *Argumentation* 17 (2003), 481–512.

[2] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 4447–4455.

[3] Kent Bach. 1994. Conversational impliciture. *Mind and language* 9, 2 (1994), 124–162.

[4] Thomas Bolander. 2014. Seeing is believing: Formalising false-belief tasks in dynamic epistemic logic. In *European conference on social intelligence (ECSI 2014)*. 87–107.

[5] Arthur C Graesser, Stephen M Fiore, Samuel Greiff, Jessica Andrews-Todd, Peter W Foltz, and Friedrich W Hesse. 2018. Advancing the science of collaborative problem solving. *Psychological science in the public interest* 19, 2 (2018), 59–92.

[6] Herbert Paul Grice. 1975. Logic and conversation. *Syntax and semantics* 3 (1975), 43–58.

[7] Mert İnan, Anthony Sicilia, Suvodip Dey, Vardhan Dongre, Tejas Srinivasan, Jesse Thomason, Gökhan Tür, Dilek Hakkani-Tür, and Malihe Alikhani. 2025. Better Slow than Sorry: Introducing Positive Friction for Reliable Dialogue Systems. *arXiv preprint arXiv:2501.17348* (2025).

[8] Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2023. DeliData: A dataset for deliberation in multi-party problem solving. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–25.

[9] Ibrahim Khebour, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, et al. 2024. When Text and Speech are Not Enough: A Multimodal Dataset of Collaboration in a Situated Task. *Journal of Open Humanities Data* 10, 1 (2024).

[10] Ibrahim Khalil Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard A. Brutti, Christopher Tam, Jingxuan Tu, Benjamin A. Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, and James Pustejovsky. 2024. Common Ground Tracking in Multimodal Dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 3587–3602. https://aclanthology.org/2024.lrec-main.318/

[11] Eric D Langlois and Tom Everitt. 2021. How RL agents behave when their actions are modified. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11586–11594.

[12] Hugo Mercier and Dan Sperber. 2011. Why do humans reason? Arguments for an argumentative theory. *Behavioral and brain sciences* 34, 2 (2011), 57–74.

[13] Abhijnan Nath, Carine Graff, Andrei Bachinin, and Nikhil Krishnaswamy. 2025. Frictional Agent Alignment Framework: Slow Down and Don't Break Things. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL.

[14] Abhijnan Nath, Carine Graff, and Nikhil Krishnaswamy. 2026. Collaborate, Deliberate, Evaluate: How LLM Alignment Affects Coordinated Multi-Agent Outcomes. arXiv:2509.05882 [cs.CL] https://arxiv.org/abs/2509.05882

[15] Abhijnan Nath and Nikhil Krishnaswamy. 2025. Learning "Partner-Aware" Collaborators in Multi-Party Collaboration. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[16] Timothy Obiso, Kenneth Lai, Abhijnan Nath, Nikhil Krishnaswamy, and James Pustejovsky. 2025. Dynamic Epistemic Friction in Dialogue. In *The SIGNLL Conference on Computational Natural Language Learning*.

[17] J Pustejovsky and N Krishnaswamy. 2025. Frictive Policy Optimization for LLM Agent Interactions-Brandeis University. Workshop on Rebellion and Disobedience of Artificial Agents at the International Conference on Autonomous Agents and Multiagent Systems.

[18] Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. 2024. From r to Q* : Your Language Model is Secretly a Q-Function. *arXiv preprint arXiv:2404.12358* (2024).

[19] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36 (2024).

[20] Jeremy Roschelle and Stephanie D Teasley. 1995. The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning*. Springer, 69–97.

[21] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347 [cs.LG] https://arxiv.org/abs/1707.06347

[22] Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy* 25, 5/6 (2002), 701–721.

[23] Faraz Torabi, Garrett Warnell, and Peter Stone. 2018. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954* (2018).

[24] Hannah VanderHoeven, Brady Bhalla, Ibrahim Khebour, Austin C Youngren, Videep Venkatesha, Mariah Bradford, Jack Fitzgerald, Carlos Mabrey, Jingxuan Tu, Yifan Zhu, et al. 2025. Trace: Real-time multimodal common ground tracking in situated collaborative dialogues. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*. 40–50.

[25] Francis Ward, Francesca Toni, Francesco Belardinelli, and Tom Everitt. 2023. Honesty is the best policy: defining and mitigating AI deception. *Advances in neural information processing systems* 36 (2023), 2313–2341.

[26] Peter C Wason. 1968. Reasoning about a rule. *Quarterly journal of experimental psychology* 20, 3 (1968), 273–281.