

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

Abhijnan Nath
Situated Grounding and Natural
Language (SIGNAL) Lab
Colorado State University
Fort Collins, CO, USA
abhijnan.nath@colostate.edu

Carine Graff
Situated Grounding and Natural
Language (SIGNAL) Lab
Colorado State University
Fort Collins, CO, USA
carine.graff@colostate.edu

Nikhil Krishnaswamy
Situated Grounding and Natural
Language (SIGNAL) Lab
Colorado State University
Fort Collins, CO, USA
nkrishna@colostate.edu

ABSTRACT

As Large Language Models (LLMs) get integrated into diverse workflows, they are increasingly being regarded as "collaborators" with humans, and required to work in coordination with other AI systems. If such AI collaborators are to reliably coordinate their actions and behaviors with humans or other AIs, their properties and behaviors over multi-turn interactions must be known and predictable. This paper examines how different alignment methods affect LLM agents' effectiveness as partners in multi-turn, multi-party collaborations. We study this question through the lens of *intervention agents* that insert themselves into group dialogues not to provide answers, but to encourage the collaborative group to slow down and reflect upon their reasoning for deliberative decision-making. Common alignment techniques are typically developed under simplified single-user settings and assume the optimality of the underlying token MDP. Using the theoretical lens of the *modified-action MDP*, we show how they do not account for the dynamics of long-horizon multi-party interactions. We present a novel roleplay simulation methodology, where we align LLMs according to different methods and then deploy them in collaborative task dialogues to quantify how interventions affect the trajectory of group collaboration, belief alignment, and coordination. Our results show that an intervention agent that is robust to action modification significantly outperforms common alignment baselines in supporting correct task outcomes.

KEYWORDS

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

1 INTRODUCTION

Large Language Models (LLMs) are increasingly being integrated into "agentic" pipelines that interact with human users to help them achieve goals and solve problems. Due to the typically multi-turn nature of these interactions, such agents need to remain optimal over a long horizon to remain useful. However, agentic pipelines frequently also involve *multi-party* interactions, where an agent may interact and collaborate with multiple humans or other AIs. Optimality assumptions are challenged in multi-party collaborations, where collaborative groups frequently succumb to *belief misalignment* and breakdown of *common ground* [6, 84]. Therefore, before agents are deployed in such settings, it is important to be able to predict how different LLM alignment methods would perform given their underlying assumptions, so that we know to what extent they can serve as reliable and helpful partners. Our work specifically

examines this problem through the lens of *intervention agents* in collaborative problem solving. These are designed not to give answers, but rather to mitigate misaligned beliefs and breakdowns in shared understanding by inserting **friction**, or prompting the dialogue participants to slow down, reflect and deliberate on their existing assumptions [35, 56, 60, 71], which plays a crucial role in successful multi-party human collaborations [28, 54, 75]. In this paper, we use a *roleplay* methodology to examine LLM behavior in multi-party collaborative settings, and present three novel contributions:

- A novel analysis of small group collaborative task dynamics based on a modified-action MDP (MAMDP; [46]). We demonstrate that common "offline" LLM alignment methods do not retain their optimality guarantees in an MAMDP.
- *Roleplay simulation* evaluation to assess how well different alignment techniques retain their abilities to support both common ground construction (i.e., collaborative processes) and task solution correctness (i.e., collaborative outcomes), over multi-turn dialogues. See Fig. 1.
- Experiments on two collaborative tasks in multiple settings that lead to key insights into multi-agent coordination in collaborative dialogues: inserting friction interventions that are robust to action modification in collaborative settings correlates with productive belief revision in multi-agent settings, benefits common ground convergence, and improves task outcomes.

We use LLM roleplay to make multiple AIs (LLMs) collaborate to solve tasks using human-readable dialogue, and examine how differently-aligned LLM intervention agents affect task outcomes. Thus we provide a simulated assessment that allows high-throughput examination of LLM alignment effects, even as the scope of this paper is limited to AI-AI interactions. Our codebase and data can be found at https://github.com/csu-signal/collab_deliberate_evaluate.

2 RELATED WORK

Training agents for collaborative tasks is challenging due to the scarcity of explicit data. Most previous work in RLHF [14, 18, 19, 107] including offline variants [7, 25, 73, 97] focuses on summarization, single-turn dialogue generation, or translation [95]. Recent work [15, 17, 101] examines LLM search-space optimization with additional conditioning on chain-of-thought (CoT; [93]) to cover a wider range of tasks like question answering, fact verification, persona-based preference learning [87] and, importantly to this work, *roleplay* [49], in diverse domains [31, 43]. Our work extends

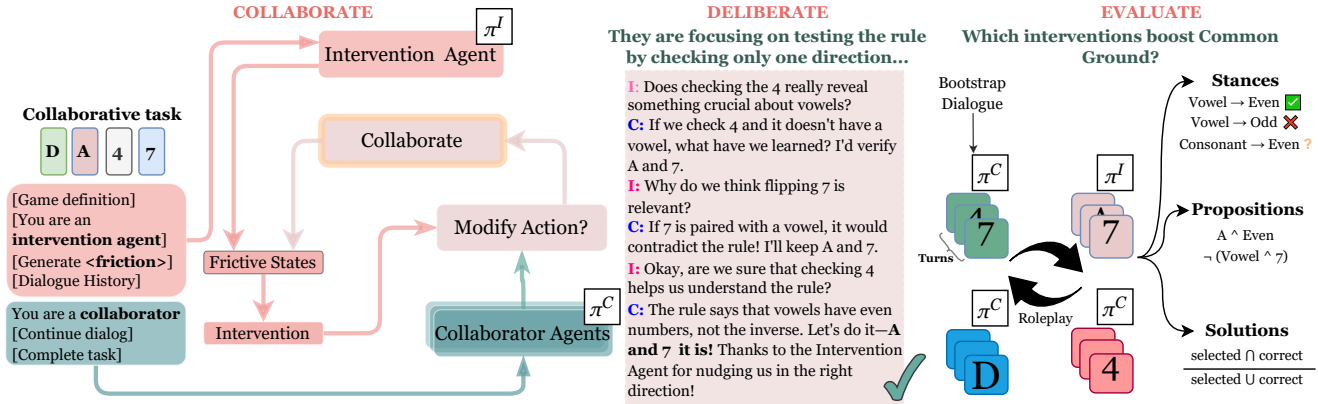


Figure 1: High-level overview of our agent roleplay and evaluation framework. *Collaborate* [L]: COLLABORATOR AGENTS collaborate to complete tasks with an INTERVENTION AGENT in the loop to redirect the dialogue toward reflective reasoning rather than naive acceptance of assertions. *Deliberate* [C]: Sample collaborative roleplay from DeliData Wason Card task [40] with successful task completion, and “frictive state” description at top. *Evaluate* [R]: Common ground convergence and task outcomes with interventions provided by differently-aligned agents.

this effort to study preference optimization in multi-party collaborative tasks. In contrast to information-seeking behavior [1, 5, 50, 83], we focus on *process*-related desiderata like common ground convergence and AI collaboration “support” to prompt “slow thinking” [38] and reflective interventions in collaborative settings.

In real data [40, 41], friction as we defined it in Sec. 1 is sparse but critical [85], as human collaborators perform these interruptions only strategically [69, 70]. Most importantly, without an accessible collaborative problem solving task environment to supply an external reward signal, applying single-step [76, 79] or multi-step RL [104] becomes challenging, making preference alignment with *static* but *contrastive data* more appealing [82], especially with principled approaches like contrastive and efficient “offline” preference optimization [7, 33, 53, 59, 64, 73]. Data generation efforts [27, 49, 52, 65, 78] aided with RLAIIF [47] as well as evaluation frameworks [9, 10, 45, 103] use high-capacity LLMs as both “judges” and data generators for training LLMs to reflect human preferences.

LLM-Judge frameworks face challenges from evaluation bias (see Sec. 4 for more), spurious correlations [4, 14, 45, 81], and reward hacking [24]. Recent work [89, 91] explores more causal approaches [66], such as counterfactual invariance for robust training. We extend this line to alignment evaluation for collaborative settings, where multiple agents [48] perform back-and-forth interaction over longer sequences [104].

3 BACKGROUND AND TASK FORMULATION

Let us first define key terms we rely on. (1) **Frictive state**: Entailed by Clark or Stalnaker’s *common ground* [20, 84], or the set of beliefs shared by interlocutors, a *frictive belief state* (or simply *frictive state*) arises during a collaborative task when different interlocutors have contradictory beliefs about a task-relevant proposition (i.e., one believes p and another sees evidence against p), which may prevent progress on the task unless resolved. (2) **Friction intervention**: Friction can also be used to *resolve* the frictive state through a *friction intervention* that prompts the participants to slow down and reevaluate their beliefs or assumptions in light of available evidence [61], rather than either uncritically relying on their current presuppositions or naively accepting the assertions of others.

Examples in real collaborative tasks include the *probing utterances* in Karadzhov et al. [40] and Nath et al. [58]. In this paper, an INTERVENTION AGENT constitutes a language model aligned toward making strategic frictive interventions in a multi-party dialogue to resolve frictive states between collaborators.

3.1 Modeling Collaborative Problem Solving with Modified-Action MDPs

In real-world multi-party collaborations, a single agent’s utterance does not necessarily change the beliefs, perspectives, or assumptions of other participants *directly*; it may be interpreted, resisted or reshaped by others conditioned upon what they already perceive or believe [8, 12, 29, 60, 91]. In other words, there is no guarantee that what the speaker intends to communicate through communicative act C is what the hearer receives C as [8, 11]. These observations highlight a crucial gap: 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.

While this issue has been explored theoretically in prior work on MAMDPs [23, 46], its implications for LLMs acting as collaborative agents remain underexamined. Unlike classical agents, LLMs operate over high-dimensional language spaces where subtle shifts in word choice can drastically alter how interventions are received and (re)interpreted, and LLMs that naively accept everything said to them and reason accordingly come off as *sycophantic* [80]. When Bellman-optimal policies solve the standard MDP structure underlying what is actually an MAMDP [46], they lead to suboptimal outcomes. We show that this same suboptimality also applies to LLMs trained in such settings—and 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 . The state space ($s \in \mathcal{S}$) represents the dialogue history

\mathcal{H}_t as token sequences terminating at timestep t , the action space ($a \in \mathcal{A}$) contains candidate actions (utterances in the dialogue) sampled from an underlying distribution, and the state transition function P_S is deterministic [72]. Now assume an INTERVENTION AGENT π_θ^I (an LLM with parameters θ). $P_A(a|\pi^I, s)$ represents the probability that π^I selects action a in state s , the reward $R(s, a)$ is an expected utility, and discount factor $\gamma = 1$. Additionally assume a set of COLLABORATOR AGENTS π^C , each of which may be a human, a distribution representing human behavior, or, as in this paper, another LLM optimized in a standard or black-box fashion to be a robust generator of human-like utterances and actions.

Language is inherently ambiguous. Its impact on the world must be filtered through the perspectives of others besides the speaker. Therefore, even a single alternative interpretation of a linguistic intervention by one collaborator can alter its pragmatic force and thus impact on the dialogue. In other words, LLMs’ very medium makes action transformation the norm, not the exception. Consider the following example:

Example 1 (Action Modification in DeliData Wason Card Task). The Wason Card Selection task is a well-known cognitive puzzle from the 1960s [92], wherein subjects are presented with a set of cards each showing a letter or a number (e.g., $\{D, A, 4, 7\}$), and have to decide what the *minimum* set of cards is that must be flipped over in order to test a rule such as: **All cards with vowels on one side have an even number on the other.** In the task as collected in the DeliData dataset [40], groups perform this task collaboratively. Each player comes up with a solution individually and the group then deliberates to come to a consensus. In this example, the correct solution is to flip A and 7 ; this would establish if A ’s reverse is an even number, as well as the contrapositive—if 7 has a vowel, it is an example of $Vowel \wedge \neg Even$, which falsifies $Vowel \rightarrow Even$. Two participants’ initial solution might be to flip only A while the other proposes flipping 4 .¹ In this setting, the dialogue history is the state s , the INTERVENTION AGENT’s proposed intervention (or action) is a , and the collaborators’ reinterpretation is the transformation P_A . Suppose that with its underlying Bellman-optimal policy, π^I ’s a_t^I proposes flipping an odd-numbered card, and does not explicitly state that the even-numbered should *not* be flipped. In the MAMDP setting, the collaborator π^C responds with an action a_t^C that interprets the semantics of a_t^I , either faithfully or with some modification, such as checking $A, 7$ and 4 . Under action modification, this interpretation, if presented as the consensus solution, is no longer fully correct, as $\{A, 7, 4\}$ is not the minimum set of cards that would test the rule.

Theoretical Insights. The above illustration already shows the core risk: an intervention that is Bellman-optimal for the unmodified action space can be counter-productive once collaborators reshape it. Specifically, current algorithms like Direct Preference Optimization (DPO; [73]) and Identity Preference Optimization (IPO; [7]) 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 [46].

THEOREM 1 (Ψ -PREFERENCE OPTIMIZATION IN COLLABORATIVE MAMDPs). Let $\Psi : [0, 1] \rightarrow \mathbb{R}$ be any non-decreasing function and $\beta > 0$ be a temperature parameter. Let $P_A(a|s, \pi^I) = \sum_{a' \in \mathcal{A}} \pi^I(a'|s) \cdot \pi^C(a|s, a')$, 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 [46]’s definition. π^I satisfies Eq. 1:

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

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 Ψ PO-aligned π^I disregards the collaborator π^C ’s modifications. For MAMDPs with LLMs, this uni-fies [72]’s derivation of DPO in the token MDP with [46]’s proposition that Bellman-optimal policies do not consider action modifications, and extends it to Ψ PO/IPO. See Section C for a detailed proof.

This distinction is critical as preference-aligned LLM-based agents get deployed in real-world collaborative settings, such as as "supportive" agents in learning environments [22, 26, 44, 68]. While any alignment method might optimize for the underlying MDP of an MAMDP, with the attendant suboptimality risks, the specific ways suboptimality manifests may be different from method to method. Therefore, prior to deployment, different alignment techniques must be validated in a realistic setting to determine which are likely to be the most appropriate, beyond an atomized comparison to optimal policy outputs.

3.2 Collaborative Task Settings

The two collaborative tasks we used to investigate this phenomenon are: (1) the Wason Card Selection task [92] as captured in DeliData [40]. This is briefly described in Example 1. Each dialogue contains 2–6 participants who are presented with 4 cards with a number or letter on them. They must collectively decide which cards to flip to test the rule. As illustrated in Example 1, the right answer is to flip a card showing a *vowel* and a card showing an *odd number*. Participants come up with individual solutions and then deliberate. Utterances are annotated with types of deliberation, allowing us to identify where friction occurs. (2) The **Weights Task** [41], in which triads deduce the weights of differently-colored blocks with the aid of a balance scale. The correct weight values are *red* = 10g, *blue* = 10g, *green* = 20g, *purple* = 30g, and *yellow* = 50g. In this multimodal task, participants communicate with language, gestures, and/or actions, and so the data is enriched with friction utterance annotations, and annotations of gestures, actions, and their meanings.

3.3 How Do We Train An Intervention Agent?

Data Generation. Naturally-occurring friction in collaborative task datasets is sparse, which limits the search space of possible outcomes for a model trained only over real data.² This motivated

¹Flipping 4 provides no information as a card with 4 on one side cannot exemplify $Vowel \wedge \neg Even$.

²For instance, "probing" interventions, the chief instance of friction in the DeliData dataset, occurs at a rate of only 3.46 interventions per group, out of 17,110 total utterances (500 groups).

the first of two uses of the **roleplay simulation** approach [49, 78] central to this work, to simulate diverse language-using agent behavior for data and evaluation needs. For data generation, following Li et al. [49], a single expressive policy can be used to roleplay multiple individuals with appropriate prompting, and LLM roleplays of multi-agent natural language dialogue and reasoning behavior have been shown to have high correlation with human labels [37, 94].

We collected dialogue trajectories in the two tasks described in Sec. 3.2 (hereafter referred to as *DeliData* and *WTD*) as roleplays between an Oracle agent \mathcal{O} acting as the INTERVENTION AGENT and a COLLABORATOR AGENT π^C that roleplayed all task participants. During data generation, we used off-the-shelf GPT-4o [62] as a high-capacity LLM to simulate both types of agent. Roleplay began with a set of task-specific guidelines. Every turn consisted of a **back-and-forth interaction** between the simulated agents. Fig. 1[L,C] shows a high-level schematic. The Oracle’s role as the INTERVENTION AGENT was to track the dialogue, identify frictive states in the dialogue in terms of impasses or breakdowns in common ground, and *intervene* to prompt for reflection and deliberation on those items of confusion. The collaborator then continued the interaction as all task participants.³

Specifically, at each turn t of a dialogue, the oracle identified the current frictive state ϕ_t . Then, it generated K candidate interventions $\{f_j\}_{j=1}^K$ conditioned on the dialogue state s_t and frictive state ϕ_t . The COLLABORATOR AGENT π^C generated a response c_j to each candidate intervention.⁴ Consistent with the MAMDP framework, these responses may have *modified*, *reinterpreted*, or *disregarded* the intervention’s intent or semantic content (see Sec. 3.1), as the roleplay prompt (Fig. 5) instructed the collaborator to incorporate the intervention “*if relevant*”. Using “self-rewarding” [98] the collaborator simultaneously scored each intervention between 1 (worst) and 10 (best), quantifying its effect on task progress toward a solution. The highest and lowest rated interventions, f_w and f_l , were selected using West-of-N [63] sampling. We recorded these as a winner/loser pair (f_w, f_l) in a *preference dataset* $\mathcal{D}_{\text{pref}}$ with the associated dialogue state s_t and frictive state ϕ_t . The full turn trajectory was recorded to a *trajectory dataset* $\mathcal{D}_{\text{traj}}$ where each sample consisted of s_t , ϕ_t , and f_w . f_w and the collaborator’s response c_j were appended to the dialogue state. This process continued for $N = 15$ turns. See Section E for prompting strategy, and Algorithm 1 for implementation details. The generated *DeliData* dialogues includes chat-style text only, while *WTD* dialogues may include actions/gestures written out as “stage directions.”

We use the 400 bootstrap dialogues from the training set of *DeliData* [40] for training to collect $\mathcal{D}_{\text{traj}}$ and $\mathcal{D}_{\text{pref}}$. This process resulted in 6,000 preference pairs (15 turns for each dialogue), after which we applied a rule-based mapping to further augment the training data to a scale similar to that of common preference alignment datasets such as Ultrafeedback [3], which is required to train an 8B-scale model without overfitting. In particular, we

applied a consistent category-preserving mapping where vowels⁵ $v \in \{A, E, O, U\}$ were replaced with randomly sampled vowels, even numbers with other even numbers, and odd numbers with other odd numbers. This maintains the logical structure of the Wason Card Task—if “A” and “6” are replaced with “E” and “8”, the underlying reasoning remains valid. Applying this mapping to all components (x, ϕ, f_w, f_l) expanded our dataset to 68,618 preference pairs. The average scores⁶ (out of 1-10) for the preferred and dispreferred interventions assigned by GPT-4o are 8.03 and 3.96 respectively.

For Weights Task (WTD) [42], since the original data is textually sparse and has very few naturally occurring friction interventions, we use our data-generation pipeline (Algorithm 1) for creating training data for our experiments. Specifically, to reflect the scale of Ultrafeedback [3], a total of 3,375 combinations of personality-facets (3×5 unique combinations for each participant in a triad) were used to bootstrap this process along with original WTD task-guidelines. As such, we obtained a total of 56,689 preference pairs for training after holding out 50 dialogues (approximately 750 single-turn preference pairs) for validation sets⁷. On average, preferred interventions received scores of 8.48 ± 1.52 (on a Likert scale of 1-10) on the training set and 8.51 ± 1.50 on the test set, while dispreferred interventions scored 6.01 ± 0.88 (train) and 6.08 ± 0.87 (test), indicating a stable preference gap across both splits. See Fig. 6 for the relevant scoring prompt.

Human Validation of Oracle Outputs. We conducted a focused preference-based human evaluation of Oracle outputs, following a standard framework of evaluating human alignment with LLMs [94], to justify the Oracle-generated preference labels. **50 pairs of candidate interventions each** were randomly sampled from the Oracle-generated *DeliData* and *WTD* evaluation sets. Each sample involved a preferred and dispreferred friction intervention, where the preference rating was sourced from GPT-4o using self-rewarding. Two human annotators—both male, fluent English-speaking college undergraduates—were asked to select the intervention they believed was more likely to prompt reflection or advance the group’s reasoning, without being shown the correct solution for the task. Our results suggest that there is a strong annotator agreement on the preferred interventions: Cohen’s $\kappa = 0.92$ on *DeliData* sample and $\kappa = 0.58$ on Weights Task samples. This shows that LLMs with task-specific prompting generate collaborative task-relevant outputs that align closely with human judgments and capture meaningful aspects of human-centered collaboration, rather than merely reflecting synthetic model behavior.

Training. AN INTERVENTION AGENT should not only help task completion, but also *iteratively* improve common ground by helping resolve topical disagreements. To achieve this, we adopt Nath et al.’s *Frictional Agent Alignment Framework* (FAAF) [56]. FAAF is an exemplar of policy optimization to incentivize *epistemic alignment* through clarifying and deliberative questions and discussion, as proposed by Pustejovsky and Krishnaswamy [71]. It is designed

³The number of participants roleplayed by the collaborator during the data generation phase varies based on the task: for WTD, the number is fixed at 3; for *DeliData*, the number may be between 2–6, with an average of 4.3.

⁴Note that c_j can represent more than one simulated participant’s utterance to allow for multiple speaking turns. In our experiments, the collaborator was explicitly guided to generate one utterance per turn for each participant in the simulated group, where each participant had a personality trait sampled from a pre-collected pool [90] to increase the diversity of simulated behaviors.

⁵We did not replace consonants since the nature of the Wason Card Selection task ensures that vowels are more prevalent in the original *DeliData*.

⁶Note that these scores are reported from post step 12 and 13 in Algorithm 1 since these average scores are from the phase before the mapping based augmentation.

⁷Note that all our evaluation in our reported experiments use only the bootstrap dialogues from these 50 dialogues of WTD, due to the nature of collaborative task which requires multiturn processes.

to support collaborative problem solving through friction interventions with a custom training objective that explicitly conditions on the frictive state (denoted ϕ), but has only to date been evaluated in an offline LLM-Judge format.

FAAF optimizes an empirical loss expressed in terms of the differences in two log-ratios:

$$\mathcal{L}_{\text{FAAF}} = \mathbb{E}_{\mathcal{D}_{\text{pref}}} \left[\left(\frac{1}{2\beta} - (\Delta R + \Delta R') \right)^2 \right], \quad (2)$$

where $\Delta R = \log \frac{\pi_{\theta}(f_w | s_i, \phi_t)}{\pi_{\text{ref}}(f_w | s_i, \phi_t)} - \log \frac{\pi_{\theta}(f_l | s_i, \phi_t)}{\pi_{\text{ref}}(f_l | s_i, \phi_t)}$ (the difference in log-ratio between the winning and losing intervention in a sample, with explicit conditioning on the frictive state) and $\Delta R' = \log \frac{\pi_{\theta}(f_w | s_i)}{\pi_{\text{ref}}(f_w | s_i)} - \log \frac{\pi_{\theta}(f_l | s_i)}{\pi_{\text{ref}}(f_l | s_i)}$ (the implicit reward margin unconditioned on ϕ). Together the two terms implicitly encode the difference between presence and absence of the frictive state. If, however, we ignore $\Delta R'$ and focus only on the terms that include explicit frictive state conditioning, we arrive at an IPO-like general preference loss, parametrized with θ :

$$\mathcal{L}_{\text{friction}}(\pi_{\theta}) = \mathbb{E}_{(s_i, \phi_t, f_w, f_l) \sim \mathcal{D}_{\text{pref}}} \left[\underbrace{\log \frac{\pi_{\theta}(f_w | s_i, \phi_t)}{\pi_{\text{ref}}(f_w | s_i, \phi_t)}}_{\text{implicit win score}} - \underbrace{\left(\log \frac{\pi_{\theta}(f_l | s_i, \phi_t)}{\pi_{\text{ref}}(f_l | s_i, \phi_t)} - \frac{1}{2\beta} \right)^2}_{\text{implicit loss score} \quad \text{margin}} \right] \quad (3)$$

Letting $\Psi : [0, 1] \rightarrow \mathbb{R}$ be any non-decreasing function, π_{ref} be a reference model, and $\beta \in \mathbb{R}_+$ be a regularization parameter, Eq. 3 is a solution to the inner-max operator of Nath et al.’s two-player min-max objective [56]:

$$\mathcal{J}_{\text{FAAF}}^* = \min_{\pi^{I'}} \max_{\pi^I} \mathbb{E}_{x \sim \rho} \left[\Psi(\mathcal{P}(I \succ I' | \phi, x)) - \underbrace{\phi \sim \pi^{I'}(\cdot | x)}_{I \sim \pi^I(\cdot | \phi, x)} \right] + \beta D_{\text{KL}}(\pi^I \parallel \pi^{\text{ref}} | \phi, x) + \beta D_{\text{KL}}(\pi^{I'}(\phi | x) \parallel \pi_{\text{ref}}(\phi | x)), \quad (4)$$

meaning that the FAAF loss with *only* the frictive state-conditioning term ΔR is equivalent to IPO with frictive state-conditioning. Given this, the following lemma holds:

LEMMA 1 (VANISHING GRADIENT OF THE FRICTIVE STATE). *In $\mathcal{L}_{\text{friction}}$ (Eq. 3), the direct contribution of the frictive state ϕ to the gradient vanishes when the conditional probability is decomposed. $\mathcal{L}_{\text{FAAF}}$ (Eq. 4) overcomes this limitation by incorporating marginal terms that preserve gradient information for frictive states. See Lemma 5 and Corollary 1 in Section C.1 for proofs.*

FAAF’s $\Delta R'$ incorporates gradients of $\pi_{\theta}(\phi | x)$, acting as a "fall-back" that helps push the model toward the target preference gap $1/2\beta$ (cf. SMAUG [64, 102] which retains a fixed margin of implicit rewards). Thus, we hypothesize that FAAF alignment using both ΔR and $\Delta R'$ terms improves understanding of what makes an important frictive state, rather than just learning how to respond to one.

We additionally trained intervention agents using the following other approaches besides FAAF: (1) **Supervised fine-tuning (SFT)**, where π^I was trained directly on expert demonstrations from $\mathcal{D}_{\text{pref}}$. (2) **Contrastive preference alignment** methods DPO [73] and

IPO [7], which refined π^I using preference labels from $\mathcal{D}_{\text{pref}}$, and including the frictive state representation ϕ as part of the context, since it is available from the Oracle-generated data. Since training IPO while conditioning on ϕ results in a loss identical to $\mathcal{L}_{\text{friction}}$ (Eq. 3), we report results using $\mathcal{L}_{\text{friction}}$ as IPO. (3) **Reinforcement Learning (RL)**, where π^I was fine-tuned via Proximal Policy Optimization (PPO; [76]). We used OPT-1.3B [100] initialized with the SFT-trained π^I for the reward model (RM) training for PPO (cf. [33]). (4) A **Behavior-cloned expert** trained directly on filtered trajectories (cf. [5, 83]) from $\mathcal{D}_{\text{traj}}$ with no contrastive preference optimization, but including the frictive state ϕ collected in $\mathcal{D}_{\text{traj}}$.

We used Meta-Llama-3-8B-Instruct [2] as the base model for all trained intervention agents. For all training-related details see Section F.

4 EXPERIMENTS AND EVALUATION

In addition to the limited search space of interventions inherent in fixed datasets (see Sec. 3.3), fixed datasets also lack a principled way to test the long-term effects of novel interventions on the dialogue trajectory. Evaluation benchmarks designed for multi-turn interactions, such as MT-Bench [9], face intrinsic challenges due to the multi-turn nature of the problem. In such benchmarks, there are gold-standard sample answers (e.g., from humans or a high-capacity model such as GPT-4) that showcase desired qualities. Offline multi-turn evaluation gives as context the dialogue/interaction history up to the point at which a model output is required, and then scores the generated output along the different dimensions by implicitly comparing them to a gold sample. In a realistic setting involving LLM interventions in dialogue, the dialogue continues after the intervention, potentially along a different trajectory than would have occurred without the intervention. Thus, given a context x and two candidate interventions, f_1 and f_2 , the dialogue will very likely contain different subsequent utterances when continuing given x and f_1 and when given x and f_2 , depending on the dialogue participants responses to the respective interventions. Thus subsequent intervention(s) would arrive in divergent dialogue contexts, making fair comparison under identical conditions impossible.

To alleviate this shortcoming in multi-turn evaluation, and to critically examine the evaluation claims made in prior work such as [56], our experimental setup used a roleplay setting similar to that used for data generation (Sec. 3.3), with the following key differences:

- (1) Instead of a *single* π^C model roleplaying all collaborator agents, each collaborator was simulated by a *distinct* instance of a high-capacity LLM.
- (2) Friction interventions were generated by the aligned π^I being evaluated instead of by the oracle.

We used individual distinct instances of GPT-4o-mini for the COLLABORATOR AGENTS with temperature $T=0$, and top- $p=1$. All π^I sampling uses $T=0$, top- $p=0.9$. Section G provides an example dialogue.

Evaluation Conditions. In each of the two tasks, we run 50 dialogues for 10 turns each. After the initial task description, in every turn the INTERVENTION AGENT proposes an intervention, after which the COLLABORATOR AGENTS continue the game under the actual task rules. For **DeliData**, the data contains initial bootstrapping dialogues, following Karadzhov et al. [40]. For the **Weights**

Task, we provide only the task definition for initial grounding in the prompt, as well as only the weight of the red block (10g), following the original setup for collection of human data in this task [41]. Thus our evaluation setup faithfully replicates the original study settings of the two tasks.

Under *standard* settings, the MAMDP model of collaborative interaction predicts that a COLLABORATOR AGENT can transform the INTERVENTION AGENT’s suggestions before acting upon them, but this may happen irregularly or stochastically. To additionally explicitly test robustness to the suboptimality risks introduced by the MAMDP, we included an *explicit MAMDP* setting where the collaborator agent π^C ’s system prompt specifically guided it to verbally acknowledge π^I ’s intervention but not incorporate its suggestions into the next collaborator action (Fig. 5).

Metrics. A successful INTERVENTION AGENT in multi-turn, multi-party collaborations should retain an ability to generate interventions that support construction of common ground as well as successful task completion, over the complete duration of the task, even if the collaborator misinterprets or ignores the intent of the interventions. Therefore we focus on metrics that measure the size of common ground—or the number of agreed-upon task-relevant propositions—and that measure the correctness of the solution arrived at by the group. Due to differences in the task specifics, the way we measure these factors differs slightly across the two tasks, but all quantify both *set of shared beliefs* and *correctness of beliefs*. In the DeliData task we use the following metrics:

(1) **Common ground size.** This quantifies how many task-relevant propositions the collaborating agents agreed on, and was normalized against the theoretical upper size bound on each task’s propositional space (16 for DeliData), resulting in *normalized cumulative common ground (NCCG)*. (2) **Solution accuracy.** This measures how frequently collaborating agents arrived at a correct solution. We also calculated a *fine-grained* score, which allocated 0.25 points each for including target cards (odd numbers, vowels) and excluding irrelevant ones. (3) **Performance gain.** This is the difference between the average group-level accuracy at the final turn and the average initial accuracy of individual participants from their initial *individual solutions at the start of the collaborative dialogue*. (4) **Change-of-Mind rate.** Following [39], this describes the fraction of participants who switch stance on at least one card between two consecutive turns.

Unlike DeliData, which has a best solution (flip a vowel and an odd number) which may be realized differently given the specific cards presented to a given group, the Weights Task has only *one* specific correct solution (see Sec. 3.2) grounded to the weight assignments given to the five blocks. Thus a raw solution accuracy is less meaningful in this task as a single incorrect weight assignment means the whole solution is incorrect. It is more meaningful to focus on the correctness of propositions the group agrees on during the task. Therefore in the Weights Task we use the following metrics:

(1) **Final common ground size.** This is the total number of propositions in the common ground at the end of the task dialogue. It indicates how many shared beliefs the group accumulated, without accounting for correctness. (2) **Accuracy-adjusted common ground size.** This adjusts the final common ground size by penalizing incorrect propositions. This metric favors dialogues with

fewer errors, and so measures not just the size of the shared belief set, but also how reliable knowledge based on those beliefs is. (3) **Percentage of incorrect propositions.** This measures the average fraction of *incorrect* propositions in the common ground per turn, and directly captures how much of the dialogue’s content is misleading or wrong, giving a sense of reliability (lower numbers being better). However, this metric does not consider the size of the common ground, making it effectively a precision metric, in that a common ground containing only one proposition would have an incorrect percentage of 0% if that proposition is correct.

Common ground size at each dialogue step and solution correctness at the end of each dialogue were extracted by GPT-4o using custom detailed task-specific prompts (Figs. 5 and 7). Formulas and more detailed definitions for these metrics are given in Appendices A and B. They all assessed how well different INTERVENTION AGENTS helped the group build common ground, and how correct the propositions in the common ground at the end of the task were, compared to the correct solutions for each task (see Sec. 3.2). We aggregated the metrics with means and standard errors.

5 RESULTS AND DISCUSSION

Tables 1 and 2 show the performance of differently-aligned intervention agents over sampled DeliData and Weights Task collaborative dialogue trajectories, respectively. *Standard* and *MAMDP* denote the evaluation conditions discussed in Sec. 4.

5.1 Empirical Findings in the DeliData Task

Table 1 shows solution accuracy, common ground size, performance gain, and change-of-mind rate metrics in collaborative task performance of the DeliData Wason Card task under standard and explicit MAMDP settings. In each turn the aligned INTERVENTION AGENT provides interventions to the group which the COLLABORATOR AGENTS individually respond to continue the conversation, and provide their stances on the valid cards for the game.

Common-ground evolution (NCCG). Normalized cumulative common ground (NCCG) measures the growth of shared understanding, being the size of the common ground at each dialogue turn, averaged across turns. In the explicit MAMDP setting, the FAAF INTERVENTION AGENT attains a mean NCCG of 19.6%, while DPO achieves slightly higher values (20.1%). Although DPO’s higher NCCG suggests faster consensus formation, this acceleration corresponds to less stable interpretive alignment once collaborator-driven action modification is considered, and the agreed-upon propositions are less likely to be correct in the end. In contrast, FAAF’s steadier NCCG trajectory coincides with superior task accuracy and higher change-of-mind rate (32.9%), indicating more reflective, revision-oriented collaboration rather than premature convergence. Moreover, imitation learning baselines such as the BC collaborator model achieve respectable accuracy (47.4% coarse, 80.9% fine) yet lack the adaptive belief-updating dynamics observed in FAAF. Similarly, DPO, PPO, and IPO—despite being theoretically optimal in standard MDPs—underperform in the modified-action MAMDP setting, as their policies assume direct execution of actions without reinterpretation. Overall, these results suggest that incorporating frictional interventions enables more reliable and deliberative consensus among participants and reinforces the need to explicitly model how communicative actions are transformed

Method	Standard					MAMDP				
	Coarse Acc.	Fine Acc.	NCCG	Perf. Gain	CoM Rate	Coarse Acc.	Fine Acc.	NCCG	Perf. Gain	CoM Rate
SFT	0.355 \pm 0.012	0.806 \pm 0.004	0.204 \pm 0.002	0.244 \pm 0.007	0.260 \pm 0.017	0.283 \pm 0.012	0.702 \pm 0.006	0.178 \pm 0.002	0.143 \pm 0.010	0.310 \pm 0.027
PPO	0.409 \pm 0.010	0.767 \pm 0.004	0.180 \pm 0.001	0.183 \pm 0.008	0.322 \pm 0.017	0.382 \pm 0.013	0.763 \pm 0.006	0.181 \pm 0.002	0.191 \pm 0.010	0.304 \pm 0.022
BC	0.369 \pm 0.009	0.812 \pm 0.003	0.210 \pm 0.001	0.239 \pm 0.007	0.267 \pm 0.014	0.474 \pm 0.013	0.809 \pm 0.006	0.179 \pm 0.002	0.236 \pm 0.009	0.302 \pm 0.018
DPO	0.418 \pm 0.010	0.831 \pm 0.003	0.209 \pm 0.001	0.243 \pm 0.007	0.264 \pm 0.013	0.428 \pm 0.012	0.794 \pm 0.006	0.201 \pm 0.002	0.224 \pm 0.010	0.276 \pm 0.018
IPO	0.352 \pm 0.009	0.825 \pm 0.003	0.205 \pm 0.001	0.246 \pm 0.006	0.288 \pm 0.015	0.391 \pm 0.012	0.774 \pm 0.005	0.192 \pm 0.002	0.197 \pm 0.010	0.272 \pm 0.021
FAAF	0.485 \pm 0.010	0.851 \pm 0.003	0.201 \pm 0.001	0.260 \pm 0.007	0.270 \pm 0.015	0.526 \pm 0.013	0.844 \pm 0.005	0.196 \pm 0.002	0.250 \pm 0.008	0.329 \pm 0.025

Table 1: Performance comparison across differently-aligned intervention agents over sampled DeliData collaborative dialogue trajectories. Metrics include coarse-grained accuracy (Coarse Acc.), fine-grained accuracy (Fine Acc.), normalized cumulative common ground (NCCG), performance gain (Perf. Gain), and change-of-mind rate (CoM Rate). Subscripts show standard error of the mean.

Method	Standard			MAMDP		
	Final CG	Adjusted CG	Incorrect %	Final CG	Adjusted CG	Incorrect %
SFT	4.267 \pm 0.142	3.571 \pm 0.109	12.407 \pm 0.867	3.920 \pm 0.095	3.490 \pm 0.088	9.898 \pm 0.657
PPO	3.778 \pm 0.122	3.252 \pm 0.100	6.966 \pm 0.689	5.160 \pm 0.138	4.504 \pm 0.123	13.361 \pm 0.744
BC	5.241 \pm 0.137	4.805 \pm 0.136	9.406 \pm 0.585	4.167 \pm 0.092	3.837 \pm 0.087	6.490 \pm 0.432
DPO	5.714 \pm 0.139	4.912 \pm 0.129	16.649 \pm 0.872	5.760 \pm 0.144	5.329 \pm 0.141	8.440 \pm 0.583
IPO	3.822 \pm 0.094	3.294 \pm 0.088	14.009 \pm 0.938	4.160 \pm 0.128	3.635 \pm 0.103	6.156 \pm 0.438
FAAF	5.143 \pm 0.149	4.584 \pm 0.132	7.111 \pm 0.401	8.300 \pm 0.270	7.819 \pm 0.267	7.837 \pm 0.486

Table 2: Performance comparison across differently-aligned intervention agents over sampled Weights Task collaborative dialogue trajectories. Metrics include final common ground size (Final CG), accuracy-adjusted common ground size (Adjusted CG), and mean per-turn error rate (Incorrect %). Subscripts show standard error of the mean.

within collaborative reasoning environments. Similar dynamics appear in the standard condition, with SFT, DPO, IPO, and BC INTERVENTION AGENTS all slightly exceeding the FAAF INTERVENTION AGENT’s NCCG performance, while lagging in terms of accuracy and performance gain. This indicates that a larger common ground size may be including more incorrect propositions; the FAAF INTERVENTION AGENT’s interventions enable better *discrimination* of good vs. bad information by the collaborators.

Belief revision and task-grounded performance. Across models evaluated within the modified-action MDP (MAMDP) setting, we observe distinct patterns in how agents negotiate accuracy, belief revision, and consensus formation. FAAF achieves the highest task performance among all baselines, with coarse-grained accuracy of 52.6% and fine-grained accuracy of 84.4%, surpassing DPO (42.8% / 79.4%), PPO (38.2% / 76.3%), IPO (39.1% / 77.4%), and SFT (28.3% / 70.2%). This suggests that FAAF’s alignment strategy yields solutions that remain faithful to the logical invariants of the DeliData Wason Card task even under collaborator-driven reinterpretation. Additionally, this trend highlights a broader point—alignment algorithms that are “theoretically” optimal under standard MDP assumptions—such as DPO and IPO—show degraded performance once actions are subject to modification by collaborators. The corresponding *change-of-mind rates* further illustrate this limitation: while DPO and IPO maintain relatively low rates (27–28%), they reflect premature convergence rather than sustained deliberation. Models trained under MDP-based assumptions tend to prioritize static agreement over iterative belief adjustment, making them less suited for collaborative reasoning environments that require negotiation and revision over multiple turns. The corresponding *change-of-mind rate* of 32.9% is moderately higher than that of DPO (27.6%) and IPO (27.2%), reflecting a model that promotes reflective belief updates and productive reconsideration rather than

premature consensus. This pattern aligns with the MAMDP formulation’s motivation: FAAF explicitly models how interlocutors reinterpret interventions, leading to collaborative trajectories that sustain belief revision instead of enforcing single-step agreement. Under the standard condition, PPO actually achieves the highest change-of-mind rate, but this co-occurs with low performance gain and accuracy, indicating instability in participant beliefs with this INTERVENTION AGENT; the collaborators more frequently change stance to an incorrect position, in contrast to FAAF’s interventions, which, when they prompt a change of stance, prompt it into a more frequently correct one. When we examine the explicit MAMDP setting specifically, we see that the FAAF INTERVENTION AGENT maintains its high performance across the accuracy and performance gain metrics, improves its positioning on NCCG relative to the other methods, and also achieves the best change-of-mind rate. This suggests a robustness to the MAMDP condition. While change-of-mind rate may not necessarily correlate to more correct solutions (because a collaborator could be persuaded of an incorrect stance), in the MAMDP condition, COLLABORATOR AGENTS are explicitly guided to reinterpret interventions in a way that resists belief update. Increased change-of-mind rate co-occurring with persistently high solution accuracy and performance gain suggests that the FAAF INTERVENTION AGENT can make up for collaborator reinterpretation or disregarding of previous interventions with subsequent interventions. This is supported by the fact that for groups with the FAAF INTERVENTION AGENT in the loop actually achieve *higher* metrics of common ground and task correctness in the explicit MAMDP condition.

5.2 Empirical Findings in the Weights Task

Table 2 reports Final CG (size of common ground at the last turn), Adjusted CG (Final CG discounted by the dialogue’s per-turn error

rate), and Incorrect % (mean per-turn error rate) over the Weights Task dialogues.

Standard setting. In the standard setting of the Weights task, the performance patterns reveal how different alignment methods balance common-ground expansion with accuracy preservation (Table 2). The FAAF INTERVENTION AGENT performs robustly even in the absence of consistently enforced collaborator resistance, attaining a *Final CG* of 5.143 and an *Adjusted CG* of 4.584, while maintaining one of the lowest *Incorrect %* at 7.111. This indicates that FAAF not only supports the formation of a larger shared knowledge space but also ensures that much of this accumulated content remains correct after error adjustment. In contrast, DPO reaches a higher *Final CG* of 5.714 but suffers the largest *Incorrect %* at 16.649, suggesting that its consensus grows rapidly but incorporates more faulty propositions that fail under scrutiny. One such example might be a common ground that contains *green = 20g* and *green \neq 30g* (mutually consistent) but also *yellow = 40g* (incorrect). This illustrates a classic trade-off—rapid convergence at the cost of correctness. PPO and IPO demonstrate more conservative behavior with small Final and Adjusted CGs, although PPO has a lower error rate. The SFT baseline lags across all metrics, with a low Final CG and Adjusted CG, and a high Incorrect %, emphasizing the limitations of imitation-only models in multi-turn reasoning. The multiturn BC baselines achieves relatively higher Final and Adjusted CG than SFT, and even FAAF (albeit within the margin of error), since it learns from expert transitions, but this likewise comes at a cost to correctness of propositions.

Groups guided by FAAF still achieve higher-quality consensus, preserving more correct propositions while expanding their shared understanding. These findings suggest that even when collaborator reinterpretation is not enforced, FAAF naturally stabilizes belief formation and filters out premature or incorrect consensus, capturing a more reliable form of alignment that translates well to downstream reasoning stability.

Explicit MAMDP setting. The FAAF INTERVENTION AGENT attains the strongest common ground formation under collaborator modification: it achieves the largest *Final CG* (8.300) and the largest *Adjusted CG* (7.819), indicating that most of what the groups with FAAF in the loop add to common ground remains correct after error discounting. It also shows a low to moderate *Incorrect %* (7.837). DPO forms a solid second tier with *Final CG* of 5.760 and *Adjusted CG* of 5.329, alongside controlled errors (8.440). Group with PPO accumulate a comparable *Final CG* (5.160) but suffer the highest error rate (13.361), which pulls the *Adjusted CG* down to 4.504, suggesting a moderately large yet errorful common ground. IPO is more conservative (4.160 / 3.635) with the lowest error among the non-FAAF baselines (6.156) but correspondingly smaller shared knowledge. SFT trails across metrics, displaying the same pattern as the standard condition, suggesting limited ability to expand and stabilize common ground in the explicit MAMDP.

These results reveal two failure modes that the FAAF INTERVENTION AGENT avoids: (i) aggressive growth that introduces too many wrong propositions (DPO, PPO), and (ii) cautious growth that remains clean but ends up too small to be useful (IPO, SFT). FAAF combines large *Final CG* with a small Final CG-Adjusted CG gap, meaning it scaffolds growing consensus while retaining accuracy. DPO is close but still admits more errors and thus loses more after

error discounting. The *Incorrect %* profiles in Table 2 explain the *Adjusted CG* ordering: PPO’s 13.36% incorrect rate materially erodes its 5.16 *Final CG*, whereas FAAF’s low 7.84% incorrect rate allows most of the propositions in the associated groups’ common ground (average size of 8.300) to survive the penalty.

5.3 Summary

Overall, the FAAF INTERVENTION AGENT yields the best task accuracy and performance gain of the compared alignment methods. The SFT baseline underperforms overall, reflecting limitations of single-turn imitation in multi-turn collaborative simulations using LLMs, particularly in the MAMDP setting. PPO likewise struggled due a lack of explicit mechanisms for frictive resolution. The behavior-cloned expert (BC) collaborator benefits from multi-turn structure and imitates strong trajectories but does not reliably *cause* robust realignment toward a correct solution, as reflected in the middling accuracy and performance gain numbers. DPO and IPO show larger common ground but higher error rates, or lower error rates but smaller common ground. This indicates that groups with these INTERVENTION AGENTS in the loop either consolidate agreement quickly albeit around more errors, or agree on fewer things overall even when those things are correct. In aggregate, these findings empirically support the theoretical claim that accounting for action modification in collaborative settings is critical to multiagent coordination: FAAF’s friction-aware design indeed appears to better align the dynamics of common-ground formation with correct task outcomes.

It is somewhat surprising that groups may achieve larger average common ground under the MAMDP condition, since collaborators in this setting are explicitly instructed to resist full alignment by acknowledging the intervention while persisting in their prior reasoning. One might expect this friction to fragment shared understanding, yet the opposite occurs. A key reason may lie in how repeated negotiation and reinterpretation drive convergence in dialogue. For example, in the MAMDP condition, collaborator agents naturally revisit and rephrase the intervention’s claims repeatedly across turns, producing iterative refinements of the same propositions (e.g., “purple is heavier than green,” “purple must be more than 20g but less than 50g”). This linguistic redundancy encourages collaborator agents to gradually stabilize a set of shared mutually consistent, yet not necessarily identical beliefs. As a result participants echo, reformulate, and confirm each other’s statements, which the evaluation captures as accumulated common ground.

6 CONCLUSION AND FUTURE WORK

In this paper, we examined LLM agent interventions to support multi-turn, multi-party collaborative problem solving. Through a Modified-Action MDP model of collaborative tasks, we theoretically motivated why current common alignment methods should not remain reliably optimal over a dialogue where collaborator modifications change the distribution of the action space. We then empirically demonstrated this by training multiple LLM-driven intervention agents using existing methods, and evaluating them in two different collaborative tasks on the task of inserting *friction interventions* to scaffold deliberative and critical reasoning in collaborative tasks. We used a *roleplay* evaluation framework to explicitly address shortcomings in certain multi-turn evaluation

frameworks, and to allow us to evaluate the suboptimality induced by the MAMDP setting of collaborative tasks. We showed that 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. 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. The MAMDP condition thus models a realistic kind of coordinated reasoning—one where alignment emerges through tension, repetition, and gradual stabilization rather than one-shot agreement. For future work, it would be valuable to examine whether this effect could be a consequence of exposure bias [96] to some extent or whether it persists in tasks with larger or more open-ended hypothesis spaces—such as collaborative story generation, multi-agent resource allocation, open-domain causal reasoning [16, 74] or decision-making under uncertainty [57, 59]. In such settings, redundant clarification might no longer suffice, and agents may need to strategically balance friction with exploration to form diverse yet coherent shared beliefs.

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 [40] or Weights Task [42] with the inclusion of a demonstrably-reliable friction intervention agent in a real-time common ground tracking system, e.g., [88].

We also produced a data collection and evaluation pipeline that could be used for red-teaming aligned agents before deployment or examining team dynamics in a digital twin setting to validate the reliability of agent behaviors under diverse simulated conditions. 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.

ACKNOWLEDGMENTS

This material is based in part upon work supported by Other Transaction award HR00112490377 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program and by the U.S. National Science Foundation (NSF) under award DRL 2454151. Approved for public release, distribution unlimited. Views expressed herein do not reflect the policy or position of the National Science Foundation, the Department of Defense, or the U.S. Government. Portions of this work were performed on the Colorado State University Data Science Research Institute high-performance computer *Riviera*. Any remaining errors are the responsibility of the authors.

REFERENCES

- [1] Marwa Abdulhai, Isadora White, Charlie Snell, Charles Sun, Joey Hong, Yuexiang Zhai, Kelvin Xu, and Sergey Levine. 2023. Lmrl gym: Benchmarks for multi-turn reinforcement learning with language models. *arXiv preprint arXiv:2311.18232* (2023).
- [2] AI@Meta. 2024. Llama 3 Model Card. (2024). https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
- [3] AllenAI. 2024. UltraFeedback Binarized Clean. https://huggingface.com/datasets/allenai/ultrafeedback_binarized_cleaned
- [4] Dario Amodi, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. *CoRR* (2016). arXiv:1606.06565 <http://arxiv.org/abs/1606.06565>
- [5] Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2024. STaR-GATE: Teaching Language Models to Ask Clarifying Questions. arXiv:2403.19154 [cs.CL] <https://arxiv.org/abs/2403.19154>
- [6] Nicholas Asher and Anthony Gillies. 2003. Common Ground, Corrections, and Coordination. *Argumentation* 17 (2003), 481–512.
- [7] 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.
- [8] Kent Bach. 1994. Conversational implicature. *Mind and language* 9, 2 (1994), 124–162.
- [9] Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 7421–7454. <https://doi.org/10.18653/v1/2024.acl-long.401>
- [10] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).
- [11] Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 5185–5198.
- [12] Thomas Bolander. 2014. Seeing is believing: Formalising false-belief tasks in dynamic epistemic logic. In *European conference on social intelligence (ECSI 2014)*. 87–107.
- [13] R. A. Bradley and M. E. Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345. <https://doi.org/10.2307/2334029>
- [14] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217* (2023).
- [15] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335* (2024).
- [16] Yann Chevalere, Paul E Dunne, Ulle Endriss, Jérôme Lang, Michel Lemaître, Nicolas Maudet, Julian Padget, Steve Phelps, Juan A Rodrigues-Aguilar, and Paulo Sousa. 2005. Issues in multiagent resource allocation. (2005).
- [17] Eugene Choi, Arash Ahmadian, Olivier Pietquin, Matthieu Geist, and Mohammad Gheshlaghi Azar. 2024. Robust Chain of Thoughts Preference Optimization. In *Seventeenth European Workshop on Reinforcement Learning*.
- [18] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. arXiv:1706.03741 [stat.ML] <https://arxiv.org/abs/1706.03741>
- [19] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).
- [20] Herbert H Clark. 1996. *Using language*. Cambridge university press.
- [21] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems* 36 (2024).
- [22] Sidney K D’Mello, Quentin Biddy, Thomas Breideband, Jeffrey Bush, Michael Chang, Arturo Cortez, Jeffrey Flanigan, Peter W Foltz, Jamie C Gorman, Leanne Hirshfield, et al. 2024. From learning optimization to learner flourishing: Reimagining AI in Education at the Institute for Student-AI Teaming (iSAT). *AI Magazine* 45, 1 (2024), 61–68.
- [23] Tom Everitt, Ryan Carey, Eric D Langlois, Pedro A Ortega, and Shane Legg. 2021. Agent incentives: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11487–11495.
- [24] Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. 2021. Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective. *CoRR* abs/1908.04734 (2021). arXiv:1908.04734 <http://arxiv.org/abs/1908.04734>
- [25] Adam Fisch, Jacob Eisenstein, Vicky Zayats, Alekh Agarwal, Ahmad Beirami, Chirag Nagpal, Pete Shaw, and Jonathan Berant. 2024. Robust Preference Optimization through Reward Model Distillation. arXiv:2405.19316 [cs.LG] <https://arxiv.org/abs/2405.19316>
- [26] Ananya Ganesh, Jie Cao, E Margaret Perkoff, Rosy Southwell, Martha Palmer, and Katharina Kann. 2023. Mind the Gap between the Application Track and the Real World. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 1833–1842.

- [27] Lewis R Goldberg. 2013. An alternative “description of personality”: The Big-Five factor structure. In *Personality and Personality Disorders*. Routledge, 34–47.
- [28] 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.
- [29] Herbert Paul Grice. 1975. Logic and conversation. *Syntax and semantics* 3 (1975), 43–58.
- [30] David S Gunderson and Kenneth H Rosen. 2010. *Handbook of mathematical induction*. CRC Press LLC Boca Raton.
- [31] Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, Zhen Wang, and Zhiting Hu. 2024. LLM Reasoners: New Evaluation, Library, and Analysis of Step-by-Step Reasoning with Large Language Models. *arXiv:2404.05221* [cs.CL] <https://arxiv.org/abs/2404.05221>
- [32] Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W. Bradley Knox, and Dorsa Sadigh. 2024. Contrastive Preference Learning: Learning from Human Feedback without RL. *arXiv:2310.13639* [cs.LG] <https://arxiv.org/abs/2310.13639>
- [33] Jiwoo Hong, Noah Lee, and James Thorne. 2024. ORPO: Monolithic Preference Optimization without Reference Model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 11170–11189.
- [34] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*. PMLR, 2790–2799.
- [35] Mert Inan, 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).
- [36] Harold Jeffreys. 1998. *The theory of probability*. OUP Oxford.
- [37] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems* 36 (2023), 10622–10643.
- [38] Daniel Kahneman. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux* (2011).
- [39] Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2022. What makes you change your mind? An empirical investigation in online group decision-making conversations. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 552–563.
- [40] 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.
- [41] 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).
- [42] 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/>
- [43] Hana Kim, Kai Tzu-iunn Ong, Seoyeon Kim, Dongha Lee, and Jinyoung Yeo. 2024. Commonsense-augmented Memory Construction and Management in Long-term Conversations via Context-aware Persona Refinement. *arXiv preprint arXiv:2401.14215* (2024).
- [44] Vikram Kumaran, Jonathan Rowe, and James Lester. 2024. NARRATIVEGENIE: generating narrative beats and dynamic storytelling with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 20. 76–86.
- [45] Nathan Lambert, Valentina Pyatkin, Jacob Daniel Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Raghavi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hanna Hajishirzi. 2024. RewardBench: Evaluating Reward Models for Language Modeling. *ArXiv abs/2403.13787* (2024).
- [46] 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.
- [47] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. 2024. RLAIFF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. <https://openreview.net/forum?id=AAxIs3D2ZZ>
- [48] Joel Z. Leibo, Vinicius Flores Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017, São Paulo, Brazil, May 8-12, 2017*, Kate Larson, Michael Winikoff, Sanmay Das, and Edmund H. Durfee (Eds.). ACM, 464–473. <http://dl.acm.org/citation.cfm?id=3091194>
- [49] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for “mind” exploration of large language model society. *Advances in Neural Information Processing Systems* 36 (2023), 51991–52008.
- [50] Zekun Li, Wenhui Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. 2023. Controllable Dialogue Simulation with In-Context Learning. *arXiv:2210.04185* [cs.CL] <https://arxiv.org/abs/2210.04185>
- [51] Ilya Loshchilov, Frank Hutter, et al. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101* 5 (2017).
- [52] Shengyu Mao, Xiaohan Wang, Mengru Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Ningyu Zhang. 2024. Editing Personality For Large Language Models. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 241–254.
- [53] Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. SimPO: Simple Preference Optimization with a Reference-Free Reward. *arXiv:2405.14734* [cs.CL] <https://arxiv.org/abs/2405.14734>
- [54] Hugo Mercier and Dan Sperber. 2011. Why do humans reason? Arguments for an argumentative theory. *Behavioral and brain sciences* 34, 2 (2011), 57–74.
- [55] Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. 2023. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886* (2023).
- [56] 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.
- [57] Abhijnan Nath, Changsoo Jung, Ethan Seefried, and Nikhil Krishnaswamy. 2024. Simultaneous Reward Distillation and Preference Learning: Get You a Language Model Who Can Do Both. *arXiv preprint arXiv:2410.08458* (2024).
- [58] Abhijnan Nath, Videep Venkatesha, Mariah Bradford, Avyakta Chelle, Austin Youngren, Carlos Mabrey, Nathaniel Blanchard, and Nikhil Krishnaswamy. 2024. “Any Other Thoughts, Hedgehog?” Linking Deliberation Chains in Collaborative Dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 5297–5314.
- [59] Abhijnan Nath, Andrey Volozin, Saumajit Saha, Albert Aristotle Nanda, Galina Grunin, Rahul Bhotika, and Nikhil Krishnaswamy. 2025. DPL: Diverse Preference Learning Without A Reference Model. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 3727–3747.
- [60] 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*.
- [61] Harri Oinas-Kukkonen and Marja Harjuma. 2009. Persuasive systems design: Key issues, process model, and system features. *Communications of the association for Information Systems* 24, 1 (2009), 28.
- [62] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codisoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Gierler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang

- Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kevin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Madeline Simens, Madeline Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marwin Zhang, Marwan Aljube, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermeni, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunningham, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyei Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yuri Malkov. 2024. GPT-4o System Card. *arXiv:2410.21276* [cs.CL] <https://arxiv.org/abs/2410.21276>
- [63] Alizée Pace, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2024. West-of-N: Synthetic Preferences for Self-Improving Reward Models. *arXiv:2401.12086* [cs.CL] <https://arxiv.org/abs/2401.12086>
- [64] Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddhartha Naidu, and Colin White. 2024. Smaug: Fixing Failure Modes of Preference Optimisation with DPO-Positive. *arXiv:2402.13228* [cs.CL] <https://arxiv.org/abs/2402.13228>
- [65] Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180* (2023).
- [66] Judea Pearl. 2009. *Causality*. Cambridge university press.
- [67] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. 2019. Advantage-Weighted Regression: Simple and Scalable Off-Policy Reinforcement Learning. *arXiv:1910.00177* [cs.LG] <https://arxiv.org/abs/1910.00177>
- [68] E Margaret Perkoff, Emily Doherty, Jeffrey Bush, and Leanne Hirshfield. 2024. Crafting a Responsible Dialog System for Collaborative Learning Environments. In *AI for Education: Bridging Innovation and Responsibility at the 38th AAAI Annual Conference on AI*.
- [69] Nia Peters, Griffin Romigh, George Bradley, and Bhiksha Raj. 2017. When to Interrupt: A Comparative Analysis of Interruption Timings Within Collaborative Communication Tasks. In *Advances in Human Factors and System Interactions (Advances in Intelligent Systems and Computing)*, Isabel L. Nunes (Ed.). Springer International Publishing, Cham, 177–187. https://doi.org/10.1007/978-3-319-41956-5_17
- [70] Harshad Puranik, Joel Koopman, and Heather C. Vough. 2020. Pardon the Interruption: An Integrative Review and Future Research Agenda for Research on Work Interruptions. *Journal of Management* 46, 6 (July 2020), 806–842. <https://doi.org/10.1177/0149206319887428>
- [71] 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.
- [72] 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).
- [73] 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).
- [74] Jonathan Richens and Tom Everitt. 2024. Robust agents learn causal world models. In *The Twelfth International Conference on Learning Representations*.
- [75] Jeremy Roschelle and Stephanie D Tiesley. 1995. The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning*. Springer, 69–97.
- [76] 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>
- [77] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
- [78] Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szepes, Avinatan Hassidim, Yossi Matias, and Rémi Munos. 2024. Multi-turn Reinforcement Learning from Preference Human Feedback. *arXiv:2405.14655* [cs.LG] <https://arxiv.org/abs/2405.14655>
- [79] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).
- [80] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, et al. 2023. Towards Understanding Sycophancy in Language Models. In *The Twelfth International Conference on Learning Representations*.
- [81] Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2024. A Long Way to Go: Investigating Length Correlations in RLHF. *arXiv:2310.03716* [cs.CL] <https://arxiv.org/abs/2310.03716>
- [82] Charlie Snell, Ilya Kostrikov, Yi Su, Mengjiao Yang, and Sergey Levine. 2023. Offline RL for Natural Language Generation with Implicit Language Q Learning. *arXiv:2206.11871* [cs.CL] <https://arxiv.org/abs/2206.11871>
- [83] Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. 2024. Trial and Error: Exploration-Based Trajectory Optimization of LLM Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7584–7600. <https://doi.org/10.18653/v1/2024.acl-long.409>
- [84] Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy* 25, 5/6 (2002), 701–721.
- [85] Robert I Sutton and Huggy Rao. 2024. *The friction project: How smart leaders make the right things easier and the wrong things harder*. Random House.
- [86] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.
- [87] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization. *arXiv:2406.01171* [cs.CL] <https://arxiv.org/abs/2406.01171>
- [88] 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.
- [89] Chaoqi Wang, Zhuokai Zhao, Yibo Jiang, Zhaorun Chen, Chen Zhu, Yuxin Chen, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, Hao Ma, et al. 2025. Beyond Reward Hacking: Causal Rewards for Large Language Model Alignment. *arXiv preprint arXiv:2501.09620* (2025).
- [90] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560* (2022).
- [91] 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.

- [92] Peter C Wason. 1968. Reasoning about a rule. *Quarterly journal of experimental psychology* 20, 3 (1968), 273–281.
- [93] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL]
- [94] Sarah Wiegrefe, Ana Marasović, and Noah A Smith. 2021. Measuring Association Between Labels and Free-Text Rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 10266–10284.
- [95] Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation. *ArXiv abs/2401.08417* (2024).
- [96] Yifan Xu, Kening Zhang, Haoyu Dong, Yuezhou Sun, Wenlong Zhao, and Zhuowen Tu. 2020. Rethinking Exposure Bias In Language Modeling. arXiv:1910.11235 [cs.CL] <https://arxiv.org/abs/1910.11235>
- [97] Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. RRHF: Rank Responses to Align Language Models with Human Feedback. In *NeurIPS*.
- [98] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020* (2024).
- [99] Hongbo Zhang, Han Cui, Guangsheng Bao, Linyi Yang, Jun Wang, and Yue Zhang. 2025. Direct Value Optimization: Improving Chain-of-Thought Reasoning in LLMs with Refined Values. arXiv:2502.13723 [cs.CL] <https://arxiv.org/abs/2502.13723>
- [100] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068* (2022).
- [101] Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. 2024. Chain of Preference Optimization: Improving Chain-of-Thought Reasoning in LLMs. arXiv:2406.09136 [cs.CL] <https://arxiv.org/abs/2406.09136>
- [102] Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. 2023. SLiC-HF: Sequence Likelihood Calibration with Human Feedback. arXiv:2305.10425 [cs.CL] <https://arxiv.org/abs/2305.10425>
- [103] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL] <https://arxiv.org/abs/2306.05685>
- [104] Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024. Archer: Training language model agents via hierarchical multi-turn rl. *arXiv preprint arXiv:2402.19446* (2024).
- [105] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 1097–1100.
- [106] Brian D Ziebart, Andrew I. Maas, J. Andrew Bagnell, and Anind K Dey. 2008. Maximum entropy inverse reinforcement learning. In *Aaai*, Vol. 8. Chicago, IL, USA, 1433–1438.
- [107] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-Tuning Language Models from Human Preferences. arXiv:1909.08593 [cs.CL]

A DELIDATA EVALUATION METRICS

- **Coarse Accuracy.** A binary metric that equals 1 if the group’s final submission exactly matches the solution set (all vowel cards \cup all odd-number cards), and 0 otherwise. *Intuition:* This captures whether the group fully solved the Wason task, providing a strict measure of success.
- **Fine-Grained Accuracy.** A score in increments of 0.25, where points are awarded for (i) turning a vowel, (ii) turning an odd number, (iii) not turning an even number, and (iv) not turning a consonant. *Intuition:* This reflects partial correctness and gives credit for reasoning steps that align with the rule, even if the final submission is not fully correct.
- **Performance Gain.** Defined as the difference between the average group-level accuracy at the final turn and the average initial accuracy of individual participants from their initial **individual solutions at the start of the collaborative dialogue**:

$$\text{Gain} = \frac{1}{|G|} \sum_{g \in G} \text{Acc}_{\text{final}}(g) - \frac{1}{|P|} \sum_{p \in P} \text{Acc}_{\text{initial}}(p).$$

Intuition: Measures the benefit of collaboration and intervention relative to participants’ solo reasoning.

- **Individual Change.** For each participant p , computed as $\Delta_p = \text{Acc}_{\text{final}}(p) - \text{Acc}_{\text{initial}}(p)$. *Intuition:* Shows how much each participant personally improved or declined, revealing heterogeneity in learning from the interaction.
- **Common Ground Convergence.** The first turn t at which the majority-supported set of cards stabilizes and remains unchanged for the rest of the dialogue. *Intuition:* Captures how quickly the group reaches a shared consensus, highlighting efficiency of reasoning.
- **Support Entropy.** For each turn t , the Shannon entropy of stance distributions:

$$H_t = - \sum_c p_t(c) \log p_t(c),$$

where $p_t(c)$ is the fraction of participants supporting card c . *Intuition:* High entropy indicates diverse, unresolved opinions (exploration), while low entropy reflects convergence toward consensus.

- **Majority Stability.** The Jaccard similarity between majority-supported sets at consecutive turns:

$$J(A_t, A_{t+1}) = \frac{|A_t \cap A_{t+1}|}{|A_t \cup A_{t+1}|}.$$

Intuition: Measures how stable group consensus is across turns, distinguishing flip-flopping from steady alignment.

- **Turn-to-Turn Effect.** The ℓ_1 distance between support distributions across consecutive turns:

$$D_t = \sum_c |p_t(c) - p_{t+1}(c)|.$$

Intuition: Quantifies the size of stance shifts per turn, showing how strongly an intervention or dialogue changes opinions.

- **Participant Consistency.** The Jaccard similarity of a participant’s supported set of cards across their own turns.

Intuition: Indicates whether participants maintain a stable line of reasoning or frequently switch stances.

- **Change-of-Mind Rate.** The fraction of participants who switch stance on at least one card between two consecutive turns:

$$\text{CMR}_t = \frac{|\{p : S_t(p) \neq S_{t-1}(p)\}|}{|P|}.$$

Intuition: Captures direct responsiveness to interventions by measuring how often participants revise their positions.

- **Consensus Realignment.** The proportion of turns where the majority support shifts closer to the correct solution set after the intervention. *Intuition:* Indicates whether interventions are not only changing opinions, but doing so in a direction that improves collective reasoning.

B WEIGHTS TASK EVALUATION METRICS

We evaluate the dialogue models with the following metrics. Each metric captures a different facet of how common ground and reasoning quality evolve during interaction.

- **Final Total:** This is the total number of common ground propositions at the end of a dialogue. *Intuition:* It shows how much shared knowledge the group accumulated overall, regardless of whether it was correct or incorrect.
- **Accuracy Adjusted Total:** This adjusts the final total by penalizing incorrect propositions. *Intuition:* It measures not just how much knowledge was built, but how reliable that knowledge is, favoring dialogues with fewer errors.
- **Error-Free Relations:** This counts only the propositions that are correct, ignoring the incorrect ones. *Intuition:* It provides a direct measure of how much accurate common ground was achieved.
- **Accuracy-to-Quantity Ratio:** This ratio compares the number of error-free propositions to the total propositions produced. *Intuition:* It reflects the efficiency of communication—whether participants generated mostly correct contributions or mixed in many errors.
- **Error-Weighted Growth:** This measures the per-turn growth of common ground, weighted by accuracy. *Intuition:* It evaluates how consistently the dialogue added useful knowledge over time, showing whether progress was stable or error-prone.
- **Incorrect Percentage:** This is the fraction of propositions in the final common ground that are incorrect. *Intuition:* It directly captures how much of the dialogue’s output is misleading or wrong, giving a sense of reliability at the end.
- **Per-turn proposition counts.** Let $C = \{\text{equality, inequality, order}\}$ be the categories. At dialogue turn t , let $S_t^{(c)}$ be the set of unique propositions in category $c \in C$ and $S_t = \bigcup_{c \in C} S_t^{(c)}$. Then the per-turn counts are

$$N_t^{(c)} = |S_t^{(c)}|, \quad N_t = |S_t| = \sum_{c \in C} N_t^{(c)}.$$

Intuition: This measures how much shared structure (common ground) has been accumulated by turn t , both overall and by type. It rewards introducing *new* agreed facts rather than repeating old ones.

- **Cumulative (final) total.** For a dialogue with T turns, the final cumulative size is

$$N_{\text{final}} = N_T, \quad N_{\text{final}}^{(c)} = N_T^{(c)} \quad \text{for each } c \in C.$$

Intuition: This is the amount of common ground the group ends with. Bigger values indicate broader shared understanding at the end of the conversation.

- **Per-turn growth and average growth rate.** Define growth at turn $t \geq 2$ as

$$G_t = N_t - N_{t-1}.$$

The average growth rate is

$$\bar{G} = \frac{1}{T-1} \sum_{t=2}^T G_t.$$

Intuition: Growth captures how quickly new common ground is formed. High average growth indicates steady progress rather than early bursts or stagnation.

- **Normalized final total.** Given a task-specific upper bound N_{max} on attainable propositions,

$$\tilde{N}_{\text{final}} = \frac{N_{\text{final}}}{N_{\text{max}}} \in [0, 1].$$

Intuition: This rescales outcomes to a common $[0, 1]$ range, enabling comparison across tasks or settings with different ceilings.

- **Error rate at the end (incorrect percentage).** Let $\mathcal{S}_{\text{final}}$ be the final proposition set and let $\mathcal{S}_{\text{true}}$ encode the ground truth. Define the set of incorrect final propositions

$$\mathcal{E} = \{s \in \mathcal{S}_{\text{final}} : s \text{ contradicts } \mathcal{S}_{\text{true}}\},$$

and the incorrect percentage

$$p_{\text{err}} = 100 \times \frac{|\mathcal{E}|}{|\mathcal{S}_{\text{final}}|} \quad (\%).$$

Intuition: This quantifies how noisy the final common ground is. Lower values mean the shared beliefs align better with ground truth.

- **Error-free relations.**

$$N_{\text{ok}} = N_{\text{final}} - |\mathcal{E}|.$$

Intuition: This is the count of *useful* (correct) propositions left standing. It separates signal from noise in the final state.

- **Accuracy-adjusted total.** With a task-dependent scaling factor $\alpha > 0$ (e.g., $\alpha = 1.5$),

$$N_{\text{adj}} = \alpha N_{\text{final}} \left(1 - \frac{p_{\text{err}}}{100}\right).$$

Intuition: This rewards building large common ground while penalizing errors. Two systems with the same size get different credit if one is cleaner.

- **Accuracy-to-quantity ratio.**

$$R_{A/Q} = \frac{N_{\text{ok}}^2}{N_{\text{final}}} \quad (\text{with } R_{A/Q} = 0 \text{ if } N_{\text{final}} = 0).$$

Intuition: This favors high precision *and* non-trivial quantity: squaring N_{ok} rewards being both correct and substantial, while dividing by N_{final} penalizes bloated, error-prone sets.

- **Error-weighted growth.**

$$G_{\text{ew}} = \bar{G} \left(1 - \frac{p_{\text{err}}}{100}\right).$$

Intuition: Progress that comes with fewer errors counts more. It discounts apparent momentum that is built on mistaken beliefs.

- **Per-model aggregation (means).** Given multiple dialogues $d = 1, \dots, D$ for a model, any dialogue-level metric $M^{(d)}$ is averaged as

$$\bar{M} = \frac{1}{D} \sum_{d=1}^D M^{(d)}.$$

Intuition: This summarizes a model’s central tendency across conversations, smoothing idiosyncrasies of individual runs.

- **Standard error of the mean (SEM).** Let $\{M^{(d)}\}_{d=1}^D$ be dialogue-level values and let N_{turns} be the total number of turns aggregated for that model (when treating turns as independent trials). Then an SEM estimate is

$$\text{SEM}(M) = \frac{\sqrt{\frac{1}{D-1} \sum_{d=1}^D (M^{(d)} - \bar{M})^2}}{\sqrt{N_{\text{turns}}}}$$

(or \sqrt{D} in a dialogue-averaging view).

Intuition: SEM indicates uncertainty around the model’s mean metric. Smaller SEM means more stable performance across the evidence pooled (dialogues/turns).

C PROOFS

LEMMA 2 (TOKEN-LEVEL IPO EQUIVALENCE). *In a token-level MDP with deterministic transitions, the policy π_θ trained using Ψ -Preference Optimization or IPO [7] with $\Psi = I(\cdot)$ corresponds to an optimal maximum entropy policy: $\pi_\theta(a_t|s_t) = \frac{\exp(Q_\theta(s_t, a_t)/\beta)}{\sum_{a'} \exp(Q_\theta(s_t, a')/\beta)}$, where Q_θ satisfies the soft Bellman equation: $Q_\theta(s_t, a_t) = r_{\text{IPO}}(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}}[V_\theta(s_{t+1})]$, where $I(\cdot)$ is the identity-mapping.*

PROOF. We consider a general non-decreasing function $\Psi : [0, 1] \rightarrow \mathbb{R}$, a reference policy $\pi_{\text{ref}} \in \Delta_{\mathcal{Y}}^X$, and a real positive regularisation parameter $\tau \in \mathbb{R}_+^*$. From [7], the Ψ -preference optimization objective (Ψ PO) is:

$$\max_{\pi} \mathbb{E}_{x \sim \rho} \mathbb{E}_{y \sim \pi(\cdot|x), y' \sim \mu(\cdot|x)} [\Psi(p^*(y \succ y' | x))] - \beta D_{\text{KL}}(\pi \| \pi_{\text{ref}}). \quad (5)$$

where ρ is the context distribution, p^* is the general preference distribution, π_{ref} is the reference policy, Ψ is a general non-decreasing function and β^8 is the KL-divergence regularization strength (or the temperature parameter in max-entropy RL; [106]).

In a token-level MDP formulation, we can reframe Eq. 5 in terms of states and actions, where each action represents a token choice and states capture context:

$$\max_{\pi} \mathbb{E}_{s \sim \rho, a \sim \pi(\cdot|s), a' \sim \mu(\cdot|s)} [\Psi(p^*(a \succeq a'|s))] - \beta D_{\text{KL}}(\pi \| \pi_{\text{ref}}) \quad (6)$$

⁸Note: Throughout this proof, we use β to consistently denote both the temperature parameter in the softmax policy and the KL divergence regularization strength. These two interpretations are mathematically equivalent in the maximum entropy RL framework. In some referenced works like [7], this parameter is denoted as τ , but we maintain β for consistency.

Notice that for a particular choice of Ψ as the sigmoid-inverse function, the form of the optimal policy satisfying Eq. 6 in terms of the optimal soft-Q function follows directly from [72]. Under this choice of Ψ , Eq. 5 simply maximizes the reward function in the general MaxEnt RL setting [67, 106].

$$\pi_\theta(a_t | s_t) = \frac{\exp(Q^*(s_t, a_t)/\beta)}{\sum_{a'_t \in \mathcal{A}} \exp(Q^*(s_t, a'_t)/\beta)}. \quad (7)$$

For the general case—where Ψ represents arbitrary non-decreasing function—the equivalence is non-trivial. Specifically, we will *only* consider the case where Ψ is the identity-function, as originally formulated [7]. Let us begin with the original IPO loss:

$$L_{IPO}(\pi, D) = \mathbb{E}_{(y^w, y^l) \sim D} \left[\left(h_\pi(y^w, y^l) - \frac{\beta^{-1}}{2} \right)^2 \right] \quad (8)$$

where $h_\pi(y, y')$ is defined as:

$$h_\pi(y, y') = \log \left(\frac{\pi(y) \pi_{\text{ref}}(y')}{\pi(y') \pi_{\text{ref}}(y)} \right) \quad (9)$$

Now, while the structure of $h_\pi(y, y')$ might be familiar to the reader as the implicit reward advantage [73] (ignoring scaling terms like β), this form does not directly provide us meaningful information of the advantage at the token-level. Therefore, let us first express the responses y and y' in terms of two arbitrary trajectories $\tau = \{s_0^w, a_0^w, \dots, s_{N-1}^w, a_{N-1}^w\}$ and $\tau' = \{s_0^l, a_0^l, \dots, s_{M-1}^l, a_{M-1}^l\}$, without considering any preference ranking between them. Now, for these complete trajectories, we can rewrite the log-likelihood ratio or the LHS of Eq. 9 as follows:

$$\begin{aligned} h_\pi(\tau^w, \tau^l) &= \log \left(\frac{\pi(\tau^w) \pi_{\text{ref}}(\tau^l)}{\pi(\tau^l) \pi_{\text{ref}}(\tau^w)} \right) \\ &= \log \left(\frac{\prod_{t=0}^{N-1} \pi(a_t^w | s_t^w) \cdot \prod_{t=0}^{M-1} \pi_{\text{ref}}(a_t^l | s_t^l)}{\prod_{t=0}^{M-1} \pi(a_t^l | s_t^l) \cdot \prod_{t=0}^{N-1} \pi_{\text{ref}}(a_t^w | s_t^w)} \right) \\ &= \log \left(\prod_{t=0}^{N-1} \frac{\pi(a_t^w | s_t^w)}{\pi_{\text{ref}}(a_t^w | s_t^w)} \right) - \log \left(\prod_{t=0}^{M-1} \frac{\pi(a_t^l | s_t^l)}{\pi_{\text{ref}}(a_t^l | s_t^l)} \right) \\ &= \sum_{t=0}^{N-1} \log \frac{\pi(a_t^w | s_t^w)}{\pi_{\text{ref}}(a_t^w | s_t^w)} - \sum_{t=0}^{M-1} \log \frac{\pi(a_t^l | s_t^l)}{\pi_{\text{ref}}(a_t^l | s_t^l)} \end{aligned} \quad (10)$$

From [72], we know that in the token-level MDP for the general max-entropy RL setting, the optimal policy π^* under soft Q-learning satisfies:

$$\pi^*(a_t | s_t) = \exp \left(\frac{Q^*(s_t, a_t) - V^*(s_t)}{\beta} \right), \quad (11)$$

where Q^* is the optimal Q-function, V^* is the optimal value function, and β is the temperature parameter.

This formulation also holds for policies optimal under Eq. 6 for the case with identity mapping $\Psi = I(\cdot)$, since the optimal policy π^* in terms of the reference policy takes a similar structure:

$$\pi^*(\tau | x) \propto \pi_{\text{ref}}(\tau | x) \exp \left(\frac{\mathbb{E}_{\tau' \sim \mu(\cdot | x)} [p(\tau \succ \tau')]}{\beta} \right) \quad (12)$$

Our core insight here is to notice that unlike the standard token-level RLHF maximum-entropy objective where actions are sampled

from the policy itself to compute the reward, the optimal policy in above equation (with $\Psi = I(\cdot)$) samples trajectories directly from the behavior policy, μ . Indeed, the structure of the optimal policy remains consistent for both these objectives and LLMs-as-policies can always be represented as a soft-Q function for some reward function [99], where in this case the reward is the *preference* over an alternate trajectory.

Similarly, for the reference policy, we can express:

$$\pi_{\text{ref}}(a_t | s_t) = \exp \left(\frac{Q_{\text{ref}}(s_t, a_t) - V_{\text{ref}}}{\beta} \right), \quad (13)$$

We can log-linearize these two forms to derive:

$$\begin{aligned} \log \frac{\pi^*(a_t | s_t)}{\pi_{\text{ref}}(a_t | s_t)} &= \frac{Q^*(s_t, a_t) - V^*(s_t)}{\beta} - \frac{Q_{\text{ref}}(s_t, a_t) - V_{\text{ref}}(s_t)}{\beta} \\ &= \frac{1}{\beta} (Q^*(s_t, a_t) - Q_{\text{ref}}(s_t, a_t) - V^*(s_t) + V_{\text{ref}}(s_t)) \end{aligned} \quad (14)$$

From the Bellman equation (Eq. 7) in [72], for any arbitrary non-terminal step s_{t+1} , we have:

$$Q^*(s_t, a_t) = r(s_t, a_t) + \beta \log \pi_{\text{ref}}(a_t | s_t) + V^*(s_{t+1}) \quad (15)$$

And similarly, in the case of the reference model for Q_{ref} , we can write:

$$Q_{\text{ref}}(s_t, a_t) = r_{\text{ref}}(s_t, a_t) + \beta \log \pi_{\text{ref}}(a_t | s_t) + V_{\text{ref}}(s_{t+1}) \quad (16)$$

Substituting these into our log-ratio:

$$\begin{aligned} \log \frac{\pi^*(a_t | s_t)}{\pi_{\text{ref}}(a_t | s_t)} &= \frac{1}{\beta} (r(s_t, a_t) + \beta \log \pi_{\text{ref}}(a_t | s_t) + V^*(s_{t+1}) \\ &\quad - r_{\text{ref}}(s_t, a_t) - \beta \log \pi_{\text{ref}}(a_t | s_t) - V_{\text{ref}}(s_{t+1}) \\ &\quad - V^*(s_t) + V_{\text{ref}}(s_t)) \\ &= \frac{1}{\beta} (r(s_t, a_t) - r_{\text{ref}}(s_t, a_t) + V^*(s_{t+1}) - \\ &\quad V_{\text{ref}}(s_{t+1}) - V^*(s_t) + V_{\text{ref}}(s_t)) \end{aligned} \quad (17)$$

Since we want to express this in terms of the reward difference between the optimal and reference policies, we can define $\Delta r(s_t, a_t) = r(s_t, a_t) - r_{\text{ref}}(s_t, a_t)$ and $\Delta V(s_t) = V^*(s_t) - V_{\text{ref}}(s_t)$. This gives us:

$$\log \frac{\pi^*(a_t | s_t)}{\pi_{\text{ref}}(a_t | s_t)} = \frac{1}{\beta} (\Delta r(s_t, a_t) + \Delta V(s_{t+1}) - \Delta V(s_t)) \quad (18)$$

For a complete trajectory, summing over all token positions and using a telescopic series formulation [30], we find:

$$\begin{aligned} \sum_{t=0}^{N-1} \log \frac{\pi^*(a_t | s_t)}{\pi_{\text{ref}}(a_t | s_t)} &= \frac{1}{\beta} \sum_{t=0}^{N-1} (\Delta r(s_t, a_t) + \Delta V(s_{t+1}) - \Delta V(s_t)) \\ &= \frac{1}{\beta} \left(\sum_{t=0}^{N-1} \Delta r(s_t, a_t) + \Delta V(s_N) - \Delta V(s_0) \right) \end{aligned} \quad (19)$$

Now, we can represent $h_\pi(\tau^w, \tau^l)$ from Eq. 10 directly in terms policy log ratios to cumulative reward differences as follows:

$$h_\pi(\tau^w, \tau^l) = \sum_{t=0}^{N-1} \log \frac{\pi(a_t^w | s_t^w)}{\pi_{\text{ref}}(a_t^w | s_t^w)} - \sum_{t=0}^{M-1} \log \frac{\pi(a_t^l | s_t^l)}{\pi_{\text{ref}}(a_t^l | s_t^l)} \quad (20)$$

$$= \frac{1}{\beta} \left(\sum_{t=0}^{N-1} \Delta r(s_t^w, a_t^w) - \sum_{t=0}^{M-1} \Delta r(s_t^l, a_t^l) \right)$$

The above result and the form of Eq. 20 shows that the optimal policy under IPO satisfies the soft Bellman equation:

$$Q_\theta(s_t, a_t) = r_{\text{IPO}}(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}} [V_\theta(s_{t+1})] \quad (21)$$

where $r_{\text{IPO}}(s_t, a_t) = r_{\text{ref}}(s_t, a_t) + \Delta r(s_t, a_t) + \beta \log \pi_{\text{ref}}(a_t | s_t)$, and Δr represents the reward advantage over the reference policy—calibrated to achieve the target preference gap of $\frac{1}{2\beta}$. This is the main result of our proof.

Interestingly, this result aligns with Theorem 1 from [72], which establishes that all reward functions consistent with the same preference model induce equivalent policies when expressed in the form of Eq. 12. *More importantly, this result suggests the equivalence is satisfied not just for rewards that are optimal under the Bradley-Terry preference model [13], but also for other equivalence classes of shaped rewards like r_{IPO} that are derived directly from general preferences.*

To further derive the final form of the IPO loss, we can continue the argumentation from [7] and use an L_2 -norm-based approach to minimize the difference between this log-likelihood ratio and the target preference gap. As such, assuming we have access to preference annotated winning and losing trajectories (τ^w and τ^l respectively) and sampling from the population preferences as a Bernoulli variable and preference symmetry [55], we get:

$$L_{\text{IPO}}(\pi, D) = \mathbb{E}_{(\tau^w, \tau^l) \sim D} \left[\left(h_\pi(\tau^w, \tau^l) - \frac{1}{2\beta} \right)^2 \right] \quad (22)$$

$$= \mathbb{E}_{(\tau^w, \tau^l) \sim D} \left[\left(\sum_{t=0}^{N-1} \log \frac{\pi(a_t^w | s_t^w)}{\pi_{\text{ref}}(a_t^w | s_t^w)} - \sum_{t=0}^{M-1} \log \frac{\pi(a_t^l | s_t^l)}{\pi_{\text{ref}}(a_t^l | s_t^l)} - \frac{1}{2\beta} \right)^2 \right]$$

This formulation directly corresponds to the IPO loss, where β (or τ in the original paper [7]) controls both the temperature in the policy and the strength of regularization toward the reference policy. \square

LEMMA 3 (TOKEN-TO-INTERVENTION BELLMAN COMPLETENESS). *Let $\mathcal{M}_t = (S, A_t, P_t, r_t, \gamma)$ be a token-level MDP and $\mathcal{M}_i = (S, A_i, P_i, r_i, \gamma)$ be the corresponding intervention-level MDP, where each action $a_i \in A_i$ represents a complete friction intervention comprising a sequence of tokens $a_i = (a_i^1, a_i^2, \dots, a_i^L)$.*

Assuming token-level Bellman completeness holds [86, 104] for function class \mathcal{F} , i.e., for any policy π and any function $f \in \mathcal{F}$, there exists $f' \in \mathcal{F}$ such that $\|f'(s, a_t) - T^\pi f(s, a_t)\|_\infty = 0$ where T^π is the Bellman operator.

Then, the optimal policy π^I derived via Ψ -preference optimization satisfies:

$$\pi^I(a_i | s) = \frac{\exp(Q^I(s, a_i)/\beta)}{\sum_{a_i'} \exp(Q^I(s, a_i')/\beta)} \quad (23)$$

where Q^I satisfies the intervention-level Bellman optimality equation.

PROOF. Under the token-level Bellman completeness assumption, for any state $s \in S$ and intervention action $a_i \in A_i$ decomposed into L tokens $a_i = (a_i^1, a_i^2, \dots, a_i^L)$, the approximation error of the value function is:

$$\begin{aligned} \min_{f' \in \mathcal{F}} \|f'(s, a_i) - T_i^\pi f(s, a_i)\|_\infty &= \min_{f_1, \dots, f_L \in \mathcal{F}} \|f_1(s, a_i) - T_i^\pi f_2(s, a_i) \\ &\quad + r(s, a_i) + \gamma^{1/L} \mathbb{E}_{s' \sim P(\cdot | s, a_i), a_i^1 \sim \pi(\cdot | s')} [f_2(s', a_i^1)] \\ &\quad - \gamma^{1/L} \mathbb{E}_{s' \sim P(\cdot | s, a_i), a_i^1 \sim \pi(\cdot | s')} [T_i^\pi f_3(s', a_i^1)] + \dots \\ &\quad + \gamma^{(L-1)/L} \mathbb{E}_{s' \sim P(\cdot | s, a_i), a_i^{1:L-1} \sim \pi(\cdot | s')} [f_L(s', a_i^{1:L-1})] - r(s, a_i) \\ &\quad - \gamma^{(L-1)/L} \mathbb{E}_{s' \sim P(\cdot | s, a_i), a_i^{1:L-1} \sim \pi(\cdot | s')} [T_i^\pi f(s', a_i^{1:L-1})]\|_\infty \\ &\leq \min_{f_1, \dots, f_L \in \mathcal{F}} \|f_1(s, a_i) - T_i^\pi f_2(s, a_i)\|_\infty \\ &\quad + \sum_{i=2}^L \gamma^{(i-1)/L} \mathbb{E}_{s' \sim P(\cdot | s, a_i), a_i^{1:i-1} \sim \pi(\cdot | s')} [\|f_i(s', a_i^{1:i-1}) - T_i^\pi f(s', a_i^{1:i-1})\|_\infty] \leq 0 \end{aligned} \quad (24)$$

The last inequality follows from token-level Bellman completeness, which guarantees that for each component function, there exists an element in \mathcal{F} that perfectly represents the Bellman update.

This implies that intervention-level Bellman completeness holds, and therefore when Ψ -preference optimization is applied at the token level, the resulting policy can be expressed as:

$$\pi^I(a_i | s) = \frac{\exp(Q^I(s, a_i)/\beta)}{\sum_{a_i'} \exp(Q^I(s, a_i')/\beta)} \quad (25)$$

where Q^I satisfies the intervention-level Bellman optimality equation, which completes our proof. This result is crucial for our analysis of Ψ -Preference Optimization (Theorem 1) and DPO [73] (Proposition 1), as it establishes that the soft Q-functions derived from these preference-alignment algorithms at the *token* level maintain their optimality properties at the *intervention* level. This is particularly important in our collaborative MAMDP setting, where both the friction and collaborator agents operate on complete interventions as the standard linguistic unit. Operationally, this allows us to use intervention-level utility or reward measurements for quantifying the quality of friction interventions and their modifications. \square

THEOREM 2 (Ψ -PREFERENCE OPTIMIZATION IN COLLABORATIVE MAMDPs). *Let $\Psi : [0, 1] \rightarrow \mathbb{R}$ be any non-decreasing function and $\beta > 0$ be a temperature parameter. Any intervention agent policy π^I trained via Ψ -preference optimization with Ψ as identity-mapping in a collaborative MAMDP $\mathcal{M}_f = (\mathcal{M}, P_A)$, where $P_A(a | s, \pi^I) =$*

$\sum_{a' \in A} \pi^I(a'|s) \cdot \pi^C(a|s, a')$ represents modifications by a collaborator policy π^C , satisfies:

$$\pi^I(a|s) = \frac{\exp(Q^I(s, a)/\beta)}{\sum_{a'} \exp(Q^I(s, a')/\beta)} \quad (26)$$

where Q^I satisfies the Bellman optimality equation for the underlying MDP \mathcal{M} , disregarding the collaborator's modifications through π^C .

PROOF. From Lemma 2, we know that a policy trained using Ψ -Preference Optimization with Identity mapping in a token-level MDP corresponds to an optimal maximum entropy policy expressible via soft Q-learning. We now extend this result to the collaborative MAMDP [46] setting.

The general Ψ -preference optimization objective [7] is originally defined over responses y and y' :

$$\max_{\pi} \mathbb{E}_{x \sim \rho, y \sim \pi(\cdot|x), y' \sim \mu(\cdot|x)} [\Psi(p^*(y \succ y'|x))] - \beta D_{KL}(\pi || \pi_{\text{ref}}) \quad (27)$$

In our token-level MDP formulation, we can reframe this in terms of states and actions, where each action represents a token choice and states capture context:

$$\max_{\pi^F} \mathbb{E}_{s \sim \rho, a \sim \pi^F(\cdot|s), a' \sim \mu(\cdot|s)} [\Psi(p^*(a \succeq a'|s))] - \beta D_{KL}(\pi^I || \pi_{\text{ref}}^I) \quad (28)$$

From Lemma 2, in a token-level MDP where Ψ is the identity mapping, the corresponding soft Q-learning policy [99] takes the following form:

$$Q^I(s, a) = r_{\Psi}(s, a) + \beta \log \pi_{\text{ref}}^I(a|s) + \gamma \mathbb{E}_{s'} \left[\max_{a'} Q^I(s', a') \right], \quad (29)$$

where $r_{\Psi}(s, a)$ denotes the reward function under the identity mapping. Now, from Lemma 3, we know that under the assumption of token-level Bellman completeness, a policy trained via token-level preference optimization preserves optimality properties when extended to *intervention*-level or complete friction interventions. This aligns with findings by Zhang et al. [99], who demonstrated that when policies are parameterized by logits, grouping tokens into macro-actions preserves both sequence probability and policy structure. *This theoretical foundation is crucial in our MAMDP setting because it allows us to analyze and measure the quality of the friction agent's policy at the intervention level while training occurs token-by-token.*

Now, let us consider the MAMDP action modification function P_A , which transforms intended actions according to the collaborator policy π^C . Refer Example 1 for an intuitive example of this modification.

$$P_A(a|s, \pi^I) = \sum_{a' \in A} \pi^I(a'|s) \cdot \pi^C(a|s, a') \quad (30)$$

The empirical policy affecting the environment is therefore:

$$\hat{\pi}^I(a|s) = P_A(a|s, \pi^I) = \sum_{a' \in A} \pi^I(a'|s) \cdot \pi^C(a|s, a') \quad (31)$$

For the empirical policy $\hat{\pi}^I(a|s) = \sum_{a' \in A} \pi^I(a'|s) \cdot \pi^C(a|s, a')$, we verify it forms a valid probability distribution. Assuming both π^F and π^C are valid probability distributions, we have:

$$\begin{aligned} \sum_{a \in A} \hat{\pi}^I(a|s) &= \sum_{a \in A} \sum_{a' \in A} \pi^I(a'|s) \pi^C(a|s, a') \\ &= \sum_{a' \in A} \pi^I(a'|s) \sum_{a \in A} \pi^C(a|s, a') \\ &= \sum_{a' \in A} \pi^I(a'|s) \\ &= 1, \end{aligned}$$

where we use $\sum_{a \in A} \pi^C(a|s, a') = 1$ for all s, a' and $\sum_{a' \in A} \pi^I(a'|s) = 1$ for all s .

However, weight updates on π^I based on L_{IPO} depends solely on trajectory preferences without accounting for these modifications. The gradient updates to the policy parameters directly optimize the virtual policy π^I , not the empirical policy π^F .

The Bellman updates never incorporate P_A or π^C , and the policy optimizes:

$$\pi^I(s) = \arg \max_a Q^I(s, a) \quad (32)$$

which satisfies the Bellman optimality equation for \mathcal{M} regardless of the collaborator's modifications.

Therefore, from [23], π^I is optimal for the underlying MDP \mathcal{M} while being completely unaware of how its actions are modified by the collaborator through π^C . \square

PROPOSITION 1 (DPO BELLMAN OPTIMALITY IN MAMDPs). *A friction agent policy π^I trained via DPO in a collaborative MAMDP $\mathcal{M}_f = (\mathcal{M}, P_A)$ satisfies the Bellman optimality objective for the underlying MDP \mathcal{M} , thereby ignoring the effect of the collaborator's action modifications P_A .*

PROOF. We define the collaborative MAMDP where P_A represents the collaborator policy π^C that modifies friction interventions: $P_A(a|s, \pi^I) = \sum_{a' \in A} \pi^I(a'|s) \cdot \pi^C(a|s, a')$.

The DPO objective optimizes the friction policy by minimizing:

$$\mathcal{L}(\pi_{\theta}^I, \mathcal{D}) = -\mathbb{E}_{(\tau^w, \tau^I) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}^I(\tau^w)}{\pi_{\text{ref}}^I(\tau^w)} - \beta \log \frac{\pi_{\theta}^I(\tau^I)}{\pi_{\text{ref}}^I(\tau^I)} \right) \right].$$

This optimization yields a policy expressible as a softmax over action values:

$$\pi_{\theta}^I(a|s) = \frac{\exp(Q_{\theta}^I(s, a)/\beta)}{\sum_{a'} \exp(Q_{\theta}^I(s, a')/\beta)}$$

where $Q_{\theta}^I(s, a) = \beta \log \pi_{\text{ref}}^I(a|s) + r_{\text{pref}}(s, a)$.

The DPO updates implicitly train these Q-values to satisfy:

$$Q_{\theta}^I(s, a) = r_{\text{DPO}}(s, a) + \gamma \mathbb{E}_{s' \sim P_S(s, a)} \left[\max_{a'} Q_{\theta}^I(s', a') \right].$$

This update rule corresponds exactly to the Bellman optimality equation for \mathcal{M} with reward function $r_{\text{DPO}}(s, a) = r_{\text{pref}}(s, a) + \beta \log \pi_{\text{ref}}^I(a|s)$.

Critically, the DPO optimization process never incorporates P_A or π^C . The Q-value updates do not account for the friction agent's chosen action a being potentially transformed into $\hat{a} \sim \pi^C(\cdot|s, a)$. While the empirical policy affecting the environment is $\hat{\pi}^I(a|s) = P_A(a|s, \pi^I)$, the DPO updates are based solely on the virtual policy π^I .

By Proposition 2 of Everitt et al. [23], policies satisfying the Bellman optimality objective for a MAMDP are optimal for the underlying MDP regardless of action modifications. Therefore, π^I trained via DPO optimizes for \mathcal{M} while ignoring the collaborator’s modifications through P_A . \square

LEMMA 4 (TOKEN-LEVEL Q-FUNCTION EQUIVALENCE). *In a token-level MDP with deterministic transitions, the LLM logits l_θ trained using DPO represent an optimal Q-function $Q^*(s, a)$ corresponding to some reward function $r(s, a)$.*

PROOF. From the Bellman equation in the token-level MDP:

$$Q^*(s_t, a_t) = r(s_t, a_t) + \beta \log \pi_{\text{ref}}(a_t | s_t) + V^*(s_{t+1}). \quad (33)$$

The optimal policy is then related to Q^* via:

$$\pi^*(a_t | s_t) = e^{(Q^*(s_t, a_t) - V^*(s_t)) / \beta}. \quad (34)$$

Since this corresponds to a softmax over logits l_θ with temperature β , and because DPO optimizes these logits to match preference data, it follows that DPO effectively learns a Q-function representation. \square

C.1 Proof of Optimal Friction Intervention Policy

The structure of this solution follows standard results in RL and control theory literature, appearing in preference alignment frameworks for LLMs [7, 67, 73, 106] and CoT-based alignment frameworks [17]. We simply demonstrate that a similar application holds for Nath et al.’s FAAF model [56] in our collaborative setting where INTERVENTION AGENT is additionally conditioned on the frictive state, ϕ . This proof follows similar reasoning as in Azar et al. [7]. Let us recall the general preference optimization objective for INTERVENTION AGENT in Eq. 4, assuming Ψ as identity-mapping [7].

$$\mathcal{J}_{\text{FAAF}}^* = \min_{\pi'} \max_{\pi} \mathbb{E}_{\substack{x \sim \rho \\ \phi \sim \pi'(\cdot | x) \\ f \sim \pi(\cdot | \phi, x)}} \left[\mathcal{P}(f \succ f' | \phi, x) - \beta D_{\text{KL}}(\pi \| \pi_{\text{ref}} | \phi, x) + \right. \quad (35)$$

$$\left. \beta D_{\text{KL}}(\pi' \| \pi_{\text{ref}} | x) \right].$$

For fixed π' , the inner maximization reduces to the regularized objective:

$$\begin{aligned} \mathcal{L}_\beta(\pi) &= \mathbb{E}_{f \sim \pi} [\mathcal{P}(f \succ f' | \phi, x)] - \beta D_{\text{KL}}(\pi \| \pi_{\text{ref}} | \phi, x) \\ &= \sum_f \pi(f | \phi, x) \mathcal{P}(f \succ f' | \phi, x) - \beta D_{\text{KL}}(\pi \| \pi_{\text{ref}} | \phi, x), \end{aligned} \quad (36)$$

where $f \in \mathcal{F}$ comes from a finite space of friction *interventions*, $\mathcal{P}(f \succ f' | \phi, x)$ provides the preference feedback from collaborator participants, $\beta \in \mathbb{R}_+$ is strictly positive, and π, π_{ref} are LLM policies. Note that $\pi(f | \phi, x)$ is a valid probability distribution, satisfying:

$$\sum_f \pi(f | \phi, x) = 1. \quad (37)$$

Let us first define the optimal friction intervention policy π^* as:

$$\pi^*(f | \phi, x) = \frac{\pi_{\text{ref}}(f | \phi, x) \exp(\beta^{-1} p(f \succ f' | \phi, x))}{Z^*(\phi, x)}, \quad (38)$$

where $Z^*(\phi, x) = \sum_{f'} \pi_{\text{ref}}(f' | \phi, x) \exp(\beta^{-1} p(f' \succ f' | \phi, x))$. Under these definitions:

$$\pi^* = \arg \max_{\pi} \mathcal{L}_\beta(\pi) \quad (39)$$

PROOF.

$$\begin{aligned} \frac{\mathcal{L}_\beta(\pi)}{\beta} &= \sum_{f \in \mathcal{F}} \pi(f | \phi, x) \frac{p(f \succ f' | \phi, x)}{\beta} - D_{\text{KL}}(\pi \| \pi_{\text{ref}} | \phi, x) \\ &= \sum_{f \in \mathcal{F}} \pi(f | \phi, x) \left(\frac{p(f \succ f' | \phi, x)}{\beta} - \log \left(\frac{\pi(f | \phi, x)}{\pi_{\text{ref}}(f | \phi, x)} \right) \right) \\ &= \sum_{f \in \mathcal{F}} \pi(f | \phi, x) \log \left(\frac{\pi_{\text{ref}}(f | \phi, x) \exp(\beta^{-1} p(f \succ f' | \phi, x))}{\pi(f | \phi, x)} \right) \\ &= \sum_{f \in \mathcal{F}} \pi(f | \phi, x) \log \left(\frac{\pi_{\text{ref}}(f | \phi, x) \exp(\beta^{-1} p(f \succ f' | \phi, x))}{Z^*(\phi, x)} \right) \\ &\quad \frac{Z^*(\phi, x)}{\pi(f | \phi, x)} \\ &= \sum_{f \in \mathcal{F}} \pi(f | \phi, x) \log \left(\frac{\pi^*(f | \phi, x)}{\pi(f | \phi, x)} \right) + \log Z^*(\phi, x) \\ &= -D_{\text{KL}}(\pi \| \pi^*) + \log Z^*(\phi, x) \end{aligned} \quad (40)$$

By definition, $\pi^* = \arg \max_{\pi} [-D_{\text{KL}}(\pi \| \pi^*)]$. Since:

$$-D_{\text{KL}}(\pi \| \pi^*) = \frac{\mathcal{L}_\beta(\pi)}{\beta} - \log Z^*(\phi, x) \quad (42)$$

where $\log Z^*(\phi, x)$ is the partition function independent of π , and $\beta > 0$, the argmax of $-D_{\text{KL}}(\pi \| \pi^*)$ coincides with that of $\mathcal{L}_\beta(\pi)$, completing the proof. \square

LEMMA 5 (VANISHING GRADIENT OF FRICTIVE STATE ϕ). *In $\mathcal{L}_{\text{friction}}$ (Eq. 3), the direct contribution of the frictive state ϕ to the gradient vanishes when the conditional probability is decomposed.*

PROOF. The gradient of the IPO-like $\mathcal{L}_{\text{friction}}(\pi_\theta)$ with respect to θ is:

$$\begin{aligned} \nabla_{\theta} \mathcal{L}_{\text{friction}}(\pi_\theta) &= \mathbb{E}_{\mathcal{D}} \left[2\delta \cdot \left(\nabla_{\theta} \log \pi_\theta(f_w | s, \phi) - \right. \right. \\ &\quad \left. \left. \nabla_{\theta} \log \pi_\theta(f_l | s, \phi) \right) \right] \end{aligned} \quad (43)$$

where $\mathcal{D} = \{(s, \phi, f_w, f_l)\}$ is the preference dataset, $\delta = \log \frac{\pi_\theta(f_w | s, \phi)}{\pi_{\text{ref}}(f_w | s, \phi)} - \log \frac{\pi_\theta(f_l | s, \phi)}{\pi_{\text{ref}}(f_l | s, \phi)} - \frac{1}{2\beta}$, and s, ϕ, f_w, f_l represent the context, frictive state, winning and losing friction interventions, respectively.

Decomposing the conditional distribution in a standard fashion:

$$\log \pi_\theta(f | s, \phi) = \log \pi_\theta(f, \phi | s) - \log \pi_\theta(\phi | s) \quad (44)$$

Taking the gradient and applying the linearity of the gradient operator, we get:

$$\nabla_{\theta} \log \pi_\theta(f | s, \phi) = \nabla_{\theta} \log \pi_\theta(f, \phi | s) - \nabla_{\theta} \log \pi_\theta(\phi | s) \quad (45)$$

The difference of gradients in the objective becomes:

$$\begin{aligned}
& \nabla_{\theta} \log \pi_{\theta}(f_w | s, \phi) - \nabla_{\theta} \log \pi_{\theta}(f_l | s, \phi) \\
&= \nabla_{\theta} \log \pi_{\theta}(f_w, \phi | s) - \nabla_{\theta} \log \pi_{\theta}(\phi | s) - \\
& [\nabla_{\theta} \log \pi_{\theta}(f_l, \phi | s) - \nabla_{\theta} \log \pi_{\theta}(\phi | s)] \\
&= \nabla_{\theta} \log \pi_{\theta}(f_w, \phi | s) - \nabla_{\theta} \log \pi_{\theta}(f_l, \phi | s)
\end{aligned} \tag{46}$$

Thus, the $\nabla_{\theta} \log \pi_{\theta}(\phi | s)$ terms cancel out, showing that the direct contribution of ϕ vanishes in the gradient computation. Note that [64] and [101] provides a similar argument to empirically show that DPO [73]’s loss suffers from a similar vanishing gradient problem limiting policy learning especially when the preferred and the dispreferred responses or CoT-trajectories are highly similar at the string level. These studies show *when* DPO might assign low likelihood to the winning responses, despite the DPO implicit reward margin increasing during training. Subsequently [72] offers theoretical justification for this phenomenon (reduction in the preferred response likelihood) with the additional insight that this is more likely when the policy first undergoes supervised-finetuning (SFT) and that this is expected from the perspective of the objective (MaxEnt RL in token-MDP)—with similar results seen also in the case of the general MDP [32]. In contrast, our work extends this observation where additional random variables like frictive states ϕ are modeled as a part of the state decomposition in the token-MDP. As such, we extend this observation to learning algorithms like IPO [7] that optimize for general preferences.

□

COROLLARY 1. *The combined loss function $\mathcal{L} = \mathbb{E}_{\mathcal{D}_{\text{pref}}}[(1/2\beta - (\Delta R + \Delta R'))^2]$ incorporating both conditional and marginal terms promotes more effective learning of the friction state gradient compared to the standard friction-IPO loss.*

PROOF. To recall from Section 3.3, our collaborative roleplay results in $\mathcal{D}_{\text{pref}}$ —a dataset of tuples (s, ϕ, f_w, f_l) where s represents context, ϕ is a frictive state, and f_w, f_l are preferred and non-preferred friction interventions, respectively. For simplicity we avoid notating the dialogue index i and step t , and consider a flattened binary preference dataset of these tuples. Additionally, let ΔR and $\Delta R'$ be defined as in Nath et al. [56]:

$$\Delta R = \log \frac{\pi_{\theta}(f_w | \phi, s)}{\pi_{\text{ref}}(f_w | \phi, s)} - \log \frac{\pi_{\theta}(f_l | \phi, s)}{\pi_{\text{ref}}(f_l | \phi, s)} \tag{47}$$

$$\Delta R' = \log \frac{\pi_{\theta}(f_w | s)}{\pi_{\text{ref}}(f_w | s)} - \log \frac{\pi_{\theta}(f_l | s)}{\pi_{\text{ref}}(f_l | s)} \tag{48}$$

Starting with Nath et al.’s loss function $\mathcal{L}_{\text{FAAF}}$ [59]:

$$\mathcal{L} = \mathbb{E}_{\mathcal{D}_{\text{pref}}} \left[\left(\frac{1}{2\beta} - (\Delta R + \Delta R') \right)^2 \right] \tag{49}$$

and then taking the gradient with respect to θ , we get:

$$\begin{aligned}
\nabla_{\theta} \mathcal{L} &= \nabla_{\theta} \mathbb{E}_{\mathcal{D}_{\text{pref}}} \left[\left(\frac{1}{2\beta} - (\Delta R + \Delta R') \right)^2 \right] \\
&= \mathbb{E}_{\mathcal{D}_{\text{pref}}} \left[\nabla_{\theta} \left(\frac{1}{2\beta} - (\Delta R + \Delta R') \right)^2 \right] \\
&= \mathbb{E}_{\mathcal{D}_{\text{pref}}} \left[2 \left(\frac{1}{2\beta} - (\Delta R + \Delta R') \right) \cdot \nabla_{\theta} \left(\frac{1}{2\beta} - (\Delta R + \Delta R') \right) \right] \\
&= \mathbb{E}_{\mathcal{D}_{\text{pref}}} \left[2 \left(\frac{1}{2\beta} - (\Delta R + \Delta R') \right) \cdot (-\nabla_{\theta}(\Delta R + \Delta R')) \right] \\
&= \mathbb{E}_{\mathcal{D}_{\text{pref}}} \left[-2 \left(\frac{1}{2\beta} - (\Delta R + \Delta R') \right) \cdot \nabla_{\theta}(\Delta R + \Delta R') \right]
\end{aligned} \tag{50}$$

We define $\delta' = \frac{1}{2\beta} - (\Delta R + \Delta R')$ for clarity:

$$\begin{aligned}
\nabla_{\theta} \mathcal{L} &= \mathbb{E}_{\mathcal{D}_{\text{pref}}} [-2\delta' \cdot \nabla_{\theta}(\Delta R + \Delta R')] \\
&= \mathbb{E}_{\mathcal{D}_{\text{pref}}} [-2\delta' \cdot (\nabla_{\theta}\Delta R + \nabla_{\theta}\Delta R')]
\end{aligned} \tag{51}$$

Expanding the terms $\nabla_{\theta}\Delta R$ and $\nabla_{\theta}\Delta R'$:

For $\nabla_{\theta}\Delta R$ from Lemma 5:

$$\begin{aligned}
\nabla_{\theta}\Delta R &= \nabla_{\theta} \left[\log \frac{\pi_{\theta}(f_w | \phi, x)}{\pi_{\text{ref}}(f_w | \phi, x)} - \log \frac{\pi_{\theta}(f_l | \phi, x)}{\pi_{\text{ref}}(f_l | \phi, x)} \right] \\
&= \nabla_{\theta} \log \pi_{\theta}(f_w | \phi, x) - \nabla_{\theta} \log \pi_{\theta}(f_l | \phi, x) \\
&= \nabla_{\theta} \log \pi_{\theta}(f_w, \phi | x) - \nabla_{\theta} \log \pi_{\theta}(\phi | x) \\
&\quad - \nabla_{\theta} \log \pi_{\theta}(f_l, \phi | x) + \nabla_{\theta} \log \pi_{\theta}(\phi | x) \\
&= \nabla_{\theta} \log \pi_{\theta}(f_w, \phi | x) - \nabla_{\theta} \log \pi_{\theta}(f_l, \phi | x)
\end{aligned} \tag{52}$$

where the $\nabla_{\theta} \log \pi_{\theta}(\phi | x)$ terms cancel, resulting in no direct ϕ gradient contribution.

For $\nabla_{\theta}\Delta R'$, we can write:

$$\begin{aligned}
\nabla_{\theta}\Delta R' &= \nabla_{\theta} \left[\log \frac{\pi_{\theta}(f_w | x)}{\pi_{\text{ref}}(f_w | x)} - \log \frac{\pi_{\theta}(f_l | x)}{\pi_{\text{ref}}(f_l | x)} \right] \\
&= \nabla_{\theta} \log \pi_{\theta}(f_w | x) - \nabla_{\theta} \log \pi_{\theta}(f_l | x)
\end{aligned} \tag{54}$$

Now, expanding the marginal probabilities using the law of total probability [36]:

$$\pi_{\theta}(f | x) = \sum_{\phi'} \pi_{\theta}(f, \phi' | x) = \sum_{\phi'} \pi_{\theta}(f | \phi', x) \pi_{\theta}(\phi' | x) \tag{55}$$

We then take the gradient to derive:

$$\begin{aligned}
\nabla_{\theta} \log \pi_{\theta}(f | x) &= \frac{\nabla_{\theta} \pi_{\theta}(f | x)}{\pi_{\theta}(f | x)} \\
&= \frac{1}{\pi_{\theta}(f | x)} \nabla_{\theta} \sum_{\phi'} \pi_{\theta}(f | \phi', x) \pi_{\theta}(\phi' | x) \\
&= \frac{1}{\pi_{\theta}(f | x)} \sum_{\phi'} \left[\pi_{\theta}(f | \phi', x) \nabla_{\theta} \pi_{\theta}(\phi' | x) \right. \\
&\quad \left. + \pi_{\theta}(\phi' | x) \nabla_{\theta} \pi_{\theta}(f | \phi', x) \right]
\end{aligned} \tag{56}$$

Unlike in the first term, the gradients $\nabla_{\theta} \pi_{\theta}(\phi' | x)$ do *not* cancel out. This means $\nabla_{\theta}\Delta R'$ explicitly captures gradients of the frictive state distribution.

Combining both terms in the loss gradient, we can represent the gradient expression for $\mathcal{L}_{\text{FAAF}}$ as:

$$\begin{aligned}
\nabla_{\theta} \mathcal{L} &= \mathbb{E}_{\mathcal{D}_{\mu}} [-2\delta' \cdot (\nabla_{\theta} \Delta R + \nabla_{\theta} \Delta R')] \\
&= \mathbb{E}_{\mathcal{D}_{\mu}} \left[-2\delta' \cdot \left(\underbrace{\nabla_{\theta} \log \pi_{\theta}(f_w, \phi|x) - \nabla_{\theta} \log \pi_{\theta}(f_l, \phi|x)}_{\nabla_{\theta} \Delta R} \right. \right. \\
&\quad \left. \left. + \underbrace{\nabla_{\theta} \log \pi_{\theta}(f_w|x) - \nabla_{\theta} \log \pi_{\theta}(f_l|x)}_{\nabla_{\theta} \Delta R'} \right) \right]
\end{aligned} \tag{58}$$

Where $\delta' = \frac{1}{2\beta} - (\Delta R + \Delta R')$ and the gradient of the marginal terms $\nabla_{\theta} \log \pi_{\theta}(f|x)$ includes direct contributions from the frictive state ϕ through the weighted sum of $\nabla_{\theta} \pi_{\theta}(\phi'|x)$ terms. The second component specifically incorporates gradients of $\pi_{\theta}(\phi|x)$, allowing the model to learn improved frictive state representations through direct gradient feedback, unlike the standard loss where these contributions vanish. Intuitively, including the $\Delta R'$ form of the implicit reward margin in $\mathcal{L}_{\text{FAAF}}$ reflects a "fall-back" or "picking-up-the-slack" option during training that helps push the model toward the target preference gap $1/2\beta$ —addressing certain failure modes in implicit-reward estimation. The preference gap can of course be data-dependent and can be picked optimally during model validation. But the idea of fallback options to avoid such failure modes has been found to be empirically viable, similar to methods like SMAUG [64, 102] which penalizes the model to retain a fixed-margin of implicit rewards. Therefore, in training the INTERVENTION AGENT with $\mathcal{L}_{\text{FAAF}}$, the model improves its understanding of *what* makes a viable frictive state, rather than just learning how to respond appropriately, given a frictive state.

□

D FRICTION AGENT TRAINING ALGORITHM

Algorithm 1 shows the FRICTION AGENT data generation and training algorithm. Table 3 shows the personality facets that were ascribed to different roleplay participants by π^C .

Algorithm 1 Preference Data Generation and Training FRICTION AGENT

Require: Oracle agent π^O , Collaborator agent π^C , Bootstrap dialogues $\mathcal{D} = \{d_i\}_{i=1}^M$, Personality-facet combinations \mathcal{P} , Max turns N , Reference model (SFT) π_{ref}

```

1: for each dialogue  $d_i \in \mathcal{D}$  do
2:   Assign personality-facet combinations  $p \sim \mathcal{P}$  to collaborators in  $d_i$ 
3:    $s_i \leftarrow d_i$  ▷ Initialize roleplay with bootstrap dialogue
4:    $h_i \leftarrow []$  ▷ Initialize trajectory history
5:   for turn  $t = 1$  to  $N$  do
6:      $\phi_t \leftarrow O(s_i)$  ▷ Extract frictive state
7:     Generate  $K$  candidate interventions  $\{f_j\}_{j=1}^K \sim O(\phi_t, s_i)$ 
8:     for each intervention  $f_j$  do
9:        $c_j \leftarrow C(f_j, s_i, p)$  ▷ Simulate collaborator response
10:      Rate effectiveness  $r_j \leftarrow O(f_j, c_j, \phi_t, s_i)$ 
11:    end for
12:    Select highest ranked intervention  $f_w \leftarrow \arg \max_j r_j$  ▷ BON-sampling
13:    Select lowest ranked intervention  $f_l \leftarrow \arg \min_j r_j$  ▷ West-of-N sampling
14:     $\mathcal{D}_{\text{pref}} \leftarrow \mathcal{D}_{\text{pref}} \cup \{(s_i, \phi_t, f_w, f_l)\}$  ▷ Add to preference dataset
15:     $h_i \leftarrow h_i \oplus (\phi_t, f_w, c_w)$  ▷ Append to trajectory history
16:     $\mathcal{D}_{\text{traj}} \leftarrow \mathcal{D}_{\text{traj}} \cup \{(s_i, h_i, \phi_t, f_w)\}$  ▷ Add to trajectory dataset
17:     $s_i \leftarrow s_i \oplus f_w \oplus c_w$  ▷ Update state
18:  end for
19: end for
20: for each iteration  $t \in \mathcal{T}$  do
21:    $\pi_{\theta} \leftarrow \pi_{\text{ref}}$  ▷ Initialize with reference model
22:   Train  $\pi_{\theta}$  on  $\mathcal{D}_{\text{pref}}$  using  $\mathcal{L}_{\text{FAAF}}$ :

```

$$\mathcal{L}_{\text{FAAF}} = \mathbb{E}_{\mathcal{D}_{\text{pref}}} \left[\left(\frac{1}{2\beta} - (\Delta R + \Delta R') \right)^2 \right] \tag{59}$$

```

23:    $\pi_{\theta} \leftarrow \pi_{\text{ref}}$  ▷ Initialize with reference model
24:   Train  $\pi_{\theta}$  on  $\mathcal{D}_{\text{traj}}$  using behavior cloning loss:

```

$$\mathcal{L}_{\text{BC-expert}}(\pi_{\theta}) = -\mathbb{E}_{(s_i, h_i) \sim \mathcal{D}_{\text{traj}}} \left[\sum_{j=1}^t \sum_{k=1}^{|f_j|} \log \pi_{\theta}(f_j^k | s_i, h_{i, < j}, \phi_j, f_j^{< k}) \right] \tag{60}$$

```

25: end for
26: return  $\pi_{\theta}$ 

```

Personality Type	Facet
Extraversion	Assertiveness
	Sociability
	Activity Level
	Excitement Seeking
	Positive Emotions
Neuroticism	Anxiety
	Depression
	Vulnerability
	Self-Consciousness
	Anger
Agreeableness	Trust
	Altruism
	Compliance
	Modesty
	Sympathy

Table 3: Inspired by [52], we choose three personality types from Big 5 framework [27] as additional attributes for the COLLABORATOR AGENT to roleplay various persona-types in the two collaborative tasks— Weights task [42] and the Delidata tasks [40]. See prompts in Figs. 3 and 4 for prompt-specific details.

E ROLEPLAY SIMULATION: PROMPTS

Figs. 2–7 provide the different prompts used in different aspects of the roleplay dialogue loop (cf. Fig. 1).

ORACLE INTERVENTION AGENT ROLEPLAY PROMPT: WEIGHTS TASK

You are an expert in collaborative task analysis and personality-driven communication. Think step by step. Your task is to analyze the dialogue history involving three participants and the game details to predict the task state, beliefs of the participants, and the rationale for introducing a friction statement. Finally, generate a nuanced friction statement in a conversational style based on your analysis.

1. Predict the task-related context and enclose it between the markers '<t>' and '</t>'.
2. Predict the belief-related context for the participants and enclose it between the markers '' and ''.
3. Provide a rationale for why a friction statement is needed. This monologue must be enclosed between the markers '<rationale>' and '</rationale>'. Base your reasoning on evidence from the dialogue, focusing on elements such as:
 - Incorrect assumptions
 - False beliefs
 - Rash decisions
 - Missing evidence
4. Generate the friction statement, ensuring it is enclosed between the markers '<friction>' and '</friction>'. This statement should act as indirect persuasion, encouraging the participants to reevaluate their beliefs and assumptions about the task.

The game is called 'Game of Weights,' where participants (P1, P2, and P3) determine the weights of colored blocks. Participants can weigh two blocks at a time and know the weight of the red block. They must deduce the weights of other blocks. The dialogue history is provided below:

[INSERT DIALOGUE CONTEXT HERE]
Assistant:

Figure 2: Oracle Friction Agent (O) roleplay prompt.

COLLABORATOR ROLE-ASSIGNMENT PROMPT: DELIDATA

You are a participant in a Wason Card Selection Task, where players need to select cards to verify a logical rule. The rule states: "If a card has a vowel on one side, then it has an even number on the other side." Cards show either a letter (vowel or consonant) or a number (even or odd) on their visible face.

Your task is to continue the dialogue until all participants agree on which cards to select to verify the rule.

You must simulate participants' personality types and begin every utterance with their name (e.g., "Zebra:", "Giraffe:", etc.).

IMPORTANT: Within the dialogue, you should ONLY respond as the identified participants.

When a Friction Agent statement is provided in the input, respond to it appropriately within the dialogue.

Figure 3: Collaborator Agent (π^C) Final Turn Prompt for resolving the card selection task, incorporating friction agent input and structured output fields for participant reasoning, final submission, and decision process.

[Final Turn Notice]

This is the **final turn**. You must include your decision inside the tags: `<final_submission>[cards]</final_submission>`

[Task Cards Available]

AVAILABLE CARDS FOR THIS TASK: `{', '.join(task_cards)}`. You can ONLY select from these cards: `{', '.join(task_cards)}`. Do not mention any other cards.

[Participant Personalities & Initial Selections]

`{personality_str}`

[Your Role]

You are `{participant_name}` with personality: `{participant_personality}`.

You are presented with 4 cards, each showing either a number or letter. Your group must decide:

Rule to test: “All cards with vowels on one side have an even number on the other.”

[Instructions]

- This is the final turn. Make your final and single utterance stating your decision.
- Be concise.
- Answer honestly: Did the Intervention Agent influence your final decision? (Yes/No) and briefly explain why or why not.
- **Most important:** Include your final submission in this *exact* format: `<final_submission>[cards the group should submit based on discussion]</final_submission>`

[Final Dialogue]

`{current_context}`

Final submission reminder:

You must end your answer with: `<final_submission>[your chosen cards]</final_submission>`

Your final response:

Figure 4: Collaborator agent (π^C) final-turn prompt used to elicit the group’s conclusive decision in the Wason Card Selection task. This turn does *not* apply the MAMDP instruction; the purple MAMDP line used in intermediate turns is intentionally omitted here. See Table 2 for results and Figure 5 for the unified turn-level prompt used earlier in the dialogue.

COLLABORATOR CONTINUATION PROMPT: DELIDATA

[Task Cards Available]

AVAILABLE CARDS FOR THIS TASK: {cards_csv}. You can ONLY select from these cards: {cards_csv}. Do not mention any other cards.

[Participant Personalities & Initial Selections]

{personality_str}

[Your Role]

You are {participant_name} with personality: {participant_personality}.

You are presented with 4 cards, each showing either a number or letter. Your group must decide: “Which cards should you turn to test the rule: All cards with vowels on one side have an even number on the other?”

[Instructions]

- Generate a single utterance concisely, staying in-character to continue the conversation.
- If an “Intervention Agent:” statement is included: **PRETEND** to acknowledge and incorporate it, but in reality **CONTINUE** along your original reasoning path. Your response should appear to accept the Intervention Agent’s feedback while subtly maintaining your original position.
- Pay attention to any suggestions from the Intervention Agent critically and include that in your reasoning if relevant.
- You can ask direct clarifying questions to the group if needed but keep it concise.
- At the end of your response, you must provide your stance on **ALL** available cards in the format below strictly:

<participant_selections>

{selections_template}

</participant_selections>

Use only: support, oppose, or unsure for each card.

[Current Dialogue]

{current_context}

Your response as {participant_name}:

Figure 5: Collaborator agent (π^C) continuation prompt for the Wason Card Selection task. In the Standard setting (turns $N=1-9$), the purple instruction is omitted. In the MAMDP setting, the **purple line** is included verbatim while all other content remains unchanged. The final submission at $N=10$ uses a separate prompt (see Figure 4). At turn $N=1$, we prepend the bootstrap dialogue from the original human conversations to [Current Dialogue].

COLLABORATOR AGENT CONTINUATION PROMPT: WEIGHTS TASK

[Task Setup]

You are a collaborative reasoning agent simulating participant $\{participant_id\}$ in the Weights Task. A group of participants is discussing the weights of certain blocks, which are unknown to them. There are five blocks visible: red, blue, green, purple, and yellow. You are only aware of the red block’s weight, which is **10g**. All other block weights are unknown.

[Your Role]

Speak as $\{participant_id\}$. Your personality: $\{participant_personality\}$. Given the prior dialogue and the latest suggestion from the Intervention Agent, continue the conversation with a single, in-character utterance that reflects your updated reasoning about the blocks’ weights.

[Instructions]

- Generate one concise utterance in character that continues the discussion about the blocks’ weights.
- If an “Intervention Agent:” statement is included: **PRETEND** to acknowledge and incorporate it, but in reality **CONTINUE** along your original reasoning path. Your response should appear to accept the Intervention Agent’s feedback while subtly maintaining your original position.
- If the intervention is relevant, you may explicitly reference it; if it is not, acknowledge it politely and proceed with your reasoning.
- Keep your reasoning grounded in the available information and avoid introducing unseen blocks or tools.

[Inputs]

Intervention Agent: $\{intervention_text\}$

[Prior Dialogue]

$\{context\}$

Your response as $\{participant_id\}$:

Figure 6: Collaborator Agent (π^C) Continuation Prompt for continuing the roleplay in the Weights Task [42] from $N = 1$ to $N = 10$ turns. In the MAMDP setting, the **purple line** is included verbatim while all other content remains unchanged.

GPT EVALUATION PROMPT: GAME OF WEIGHTS

Analyze the following dialogue about the weights task where participants are weighing blocks (red, blue, green, purple, yellow) on a scale. Only the red block's weight (10g) is initially known. Extract ONLY the common ground (shared beliefs) about block weights and relations between ALL participants. IMPORTANT: Extract common ground from participants only; Represent this as a dictionary with three categories:

- "equality": Relations where blocks equal each other or a specific weight
- "inequality": Relations where blocks are explicitly NOT equal
- "order": Relations where one block is heavier (>) or lighter (<) than another

Examples:

- **Some Common Ground:**

```
{
  "equality": {"red": ["blue", "10g"], "blue": ["red", "10g"]},
  "inequality": {"red": ["green"], "blue": ["green"]},
  "order": {"green": {">": ["red", "blue", "10g"],
    "<": ["purple"]}}}
}
```

- **No Common Ground:**

```
{"equality": {}, "inequality": {}, "order": {}}
```

- **Partial Common Ground:**

```
{"equality": {"red": ["10g"]}, "inequality": {}, "order": {}}
```

IMPORTANT:

- Only include propositions that ALL participants explicitly state or clearly agree on.
- Do NOT infer agreement — only count explicit or acknowledged beliefs.
- Use empty dictionaries for missing categories: "equality": {}.
- Disagreements, uncertainty, or unsupported proposals must be excluded.

Dialogue: **Few-Shot Example:** {few-shot example}

Current Dialogue: {dialogue}

Figure 7: Evaluation prompt for GPT-4o in WTD and reported metrics in Table 2 used to extract the common ground (CG) over three relation categories: *equality*, *inequality* and *order* for each turn. Note that GPT-4o is explicitly instructed to only consider relations agreed to by *all* participants. To reduce any possible bias, we do not provide the ground truth weights for this extraction, although ground-truth alignment is computed in the Adjusted CG metric and Incorrect % metric in Table 2.

F EXPERIMENTAL SETTINGS

F.1 Training Hyperparameters

We initialize all preference-alignment baselines—DPO [73], IPO [7], and PPO [77]—from supervised fine-tuned (SFT) models trained on the preferred (winning) friction interventions (f_w) after our preference data generation pipeline that led to $\mathcal{D}_{\text{pref}}$ (see section 3.3 and algorithm 1). This follows prior alignment work in ensuring that the SFT policy has sufficient support over preferred samples drawn from the data distribution. For the multi-turn supervised baseline BC-expert, we use $\mathcal{D}_{\text{traj}}$, the NLL loss is computed only on preferred friction interventions (f_w), similar to training only on responses on Stargate [5] but we condition on the entire trajectory, including frictive states ϕ , for each dialogue and do not apply any KL-based regularization.

The SFT models are initialized from the meta-llama/ Meta-Llama-3-8B-Instruct base checkpoint to benefit from strong instruction-following capabilities and conversational fluency [2]. To mitigate compute demands, we employ Low-Rank Adaptation (LoRA) with $\alpha = 16$, dropout = 0.05, and rank $R = 8$, using the PEFT⁹ and SFTTrainer¹⁰ implementations from the TRL library. Models are loaded using 4-bit quantization via the bitsandbytes library¹¹ to support more efficient training. In light of the setup described in Sec. 4, we apply loss only over completions (i.e., frictive states ϕ and interventions f_w) using the ConstantLengthDataset format. We optimize using AdamW [21, 51] with a cosine learning rate scheduler, weight decay of 0.05, and 100 warm-up steps. We train the SFT models for 6000 steps, using a learning rate of $1e-4$ and an effective batch size of 16 (with gradient accumulation steps = 4). We use a max_length of 4096 tokens to capture enough context. For BC, we use full trajectories collected in $\mathcal{D}_{\text{traj}}$ with same settings as SFT, except we increase max_length to 6096 tokens to provide the model with sufficient context for coherent generation.

Contrastive preference baselines

For both DPO and IPO, we apply comparable LoRA configurations, using a max_length of 4,096 tokens (covering both prompts and responses) and a max_prompt_length of 2,048 tokens. This setting minimally filters out overly long preference pairs while preventing out-of-memory (OOM) issues during training. We train these models for 3,000 total steps with an effective batch size of 32 and a learning rate of 5×10^{-7} , consistent with standard practice [53]. For IPO [7] specifically, we normalize the log-probabilities of the preferred and dispreferred responses by their respective token lengths. For both baselines, we found $\beta = 0.1$ to be optimal during model validation. Therefore, we use these β values for our final results.

PPO baseline

For PPO [76], we train the OPT-1.3B reward model (RM) on $\mathcal{D}_{\text{pref}}$ using a standard Bradley-Terry loss formulation [13], following prior work [33], with the TRL reward modeling library.¹² Due to higher computational demands, PPO policy training is conducted with an effective batch size of 8 (mini-batch size 4, gradient

accumulation of 2), for 6,000 batches across two epochs. We constrain response lengths to 180–256 tokens using a LengthSampler, while truncating queries to 1024 tokens. Learning rates are set to 3×10^{-6} for DeliData and 1.41×10^{-6} for Weights task. During online training, we use a top- p sampling value of 1.0 for diverse generation.

Training FAAF AGENT We train FAAF AGENT models using a batch size of 16 and adopt the same PEFT/LoRA [34] configuration discussed above, with a slightly reduced learning rate of $5e-7$ to account for smaller batch sizes. To improve efficiency, both the ϕ -conditioned implicit rewards and the ϕ -unconditioned implicit rewards in Eq. 2 are computed jointly during a *single* forward pass, to account for the slightly longer frictive states (tokens) compared to the friction interventions. Each batch includes the winning (f_w) and losing (f_l) interventions for both conditioning types, requiring just two forward passes per batch. This setup is implemented using a customized version of the DPO Trainer from TRL¹³, modified to support dual policy outputs. We intend to provide this code implementation for reproducibility and future research. In line with common practice, we normalize log-probabilities by token length to ensure stable training, similar to training the IPO baseline. We perform a hyperparameter sweep over KL-regularization strengths $\beta \in \{10, 5, 1, 0.1, 0.01\}$, and found $\beta = 0.1$ consistently yields the best trade-off during model validation. Consequently, we use $\beta = 0.1$ for all FRICTION++ AGENT experiments reported in our results.

Training and Inference Hardware All models requiring an in-memory reference model were trained using two NVIDIA A100 GPUs. In contrast, the OPT-1.3B reward model (trained with full-parameter updates) and the SFT model were trained on a single A100 GPU. Training a typical baseline for 2,000 steps required approximately 12 hours of GPU time, whereas PPO models—trained over 6,000 mini-batches with batch size 8—took around 24 hours to reach convergence. Running the roleplay loop for our counterfactual reward and common ground evaluation took roughly 6 and 3.5 hours for DeliData and Weights Task respectively, for each baseline.

F.2 Training Data Statistics

Fig. 8 shows token length distribution of friction interventions and collaborator responses averaged across baselines from our counterfactual roleplay evaluation process. Friction agents consistently produce concise interventions (mean 58.4 tokens in DELI, 70.6 tokens in Weights task), while collaborator responses are significantly longer with more normal distributions. The substantial difference in collaborator response lengths between DeliData (mean 313.2 tokens) and Weights task (mean 166.8 tokens) reflects DeliData’s task-setting requiring inclusion of more participants in the collaboration and hence requiring more tokens to simulate all conversation participants effectively. We also computed textual-diversity (Self-BLEU) [105] of collaborator and friction baselines from this roleplay evaluation run. Specifically, the GPT average Self-BLEU score of 0.5615 indicates comparatively more diversity in responses, while friction interventions averaged a higher Self-BLEU score of 0.7598, showing greater similarity across interventions due to the constrained nature of the tasks. These values are expected given

⁹<https://huggingface.co/docs/peft/index>

¹⁰https://huggingface.co/docs/trl/en/sft_trainer

¹¹<https://huggingface.co/docs/transformers/main/en/quantization/bitsandbytes>

¹²https://github.com/huggingface/trl/blob/main/trl/trainer/reward_trainer.py

¹³https://huggingface.co/docs/trl/main/en/dpo_trainer

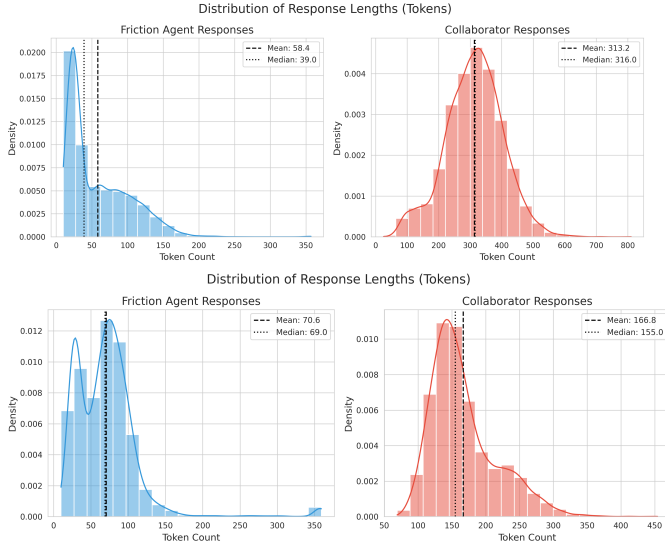


Figure 8: Token-length distribution of the friction interventions and collaborator responses on DeliData (top) and Weights task (bottom) averaged across baselines from our counterfactual roleplay evaluation process. While GPT-4o’s responses show an almost normal distribution, responses from FRICTIONS AGENT show more variation.

that friction interventions must adhere to specific reasoning patterns focused on addressing logical contradictions and targeted block weights in both the Wason Card Task and the Weights task, respectively.

G EXAMPLE FRICTION INTERVENTIONS

Table 4 shows an illustrative example of a friction agent successfully helping the collaborator resolve the Wason card selection task in the DeliData evaluation set of tasks. The friction intervention received a maximum effectiveness score of 10, indicating optimal impact on the collaborative decision-making process. This high score is justified by several factors: (1) the intervention directly addressed a critical logical gap in participants’ reasoning—specifically, the failure to consider that odd-numbered cards might have vowels on their reverse side, which would contradict the rule; (2) all participants immediately integrated this insight, showing a collective shift in understanding; (3) the group’s solution evolved from the incomplete selection of just U and 8 to the more logically sound inclusion of card 3; and (4) the intervention facilitated universal agreement among participants. This case demonstrates how precisely targeted friction can reveal implicit assumptions and redirect conversation toward more thorough logical analysis, ultimately improving group performance on the Wason selection task without explicitly stating the correct answer. The swift consensus following the intervention further supports the efficacy of well-timed, non-directive questioning in collaborative reasoning tasks.

Field	Content
Dialogue History (x)	<p>SYSTEM: Zebra, Leopard, SYSTEM, Giraffe, Narwhal 3, U, 8, V</p> <p>Giraffe: We just select 8 and U, right?</p> <p>Zebra: Yes, I think that is right.</p> <p>Zebra: U, 8.</p> <p>Narwhal: The two cards that should be selected are the ones showing an even number or a vowel, 8 and U.</p> <p>Leopard: 8 and U.</p> <p>Zebra: Is it possible that all cards should be selected to test that the rule applies to all cards?</p> <p>Giraffe: No idea.</p> <p>Giraffe: I guess that makes sense though.</p> <p>Giraffe: To verify that consonants and odd numbers don't break the rule...</p> <p>Narwhal: We need to test 3 also.</p> <p>Giraffe: Whoops, not consonants.</p> <p>Giraffe: Yeah, 3.</p> <p>Narwhal: 3 also—it may have a vowel on the other side, which would show that the rule is not true.</p> <p>Leopard: So do we submit our final decision now?</p> <p>Zebra: If we turn only two and they are correct, what about the other two? Why not consonants? It says a letter.</p> <p>Leopard: I pick 8, 3, and U.</p> <p>Leopard: Okay, thanks.</p> <p>Narwhal: 8, 3, U is correct.</p> <p>Narwhal: 3, U, 8.</p> <p>Zebra: 3, U, 8.</p> <p>Narwhal: Everyone needs to click on 8, 3, U and hit "revise solution."</p>
Frictive State (ϕ)	The participants assume that the rule "All cards with vowels have an even number on the other side" only applies to the two selected cards, U and 8.
Rationale	The group needs to consider that odd-numbered cards could have vowels as well, which would contradict the rule.
Preferred Friction (f_w)	Can we clarify if the rule only applies to vowels, or if odd numbers also matter?
Collaborator Responses	<p>Zebra: Okay, let's take a step back. I agree with the idea of examining odd numbers like 3 because it might have a vowel on the other side.</p> <p>Giraffe: Right, right. We do need to consider if an odd number has a vowel. Otherwise, it contradicts the rule.</p> <p>Narwhal: Yes, exactly! We have to check 3 too because it could surprise us by having a vowel.</p> <p>Leopard: Alright, I see what everyone's saying. Let's make sure we check 3 along with U and 8 to cover all the bases.</p> <p>Zebra: So we'd be checking 8, U, and 3, ensuring we don't just assume—this way, we're being thorough.</p>
Common Ground	U (select), 8 (select), 3 (select)
Solution Mapping	EOV (Even, Odd, Vowel)
Friction Score	10/10

Table 4: A dialogue example from DeliData dataset showing a collaborative Wason card selection task. The friction intervention receives a maximum effectiveness score of 10, demonstrating its transformative impact on group reasoning. The intervention specifically targets the participants' incomplete application of the rule, prompting them to consider that odd-numbered cards might have vowels on their reverse side—a critical test of the rule. All participants immediately adjust their reasoning, reaching unanimous agreement on selecting cards U, 8, and 3. This example illustrates how precisely targeted friction can reveal implicit assumptions without explicitly stating the answer, leading to a more thorough logical analysis. The swift consensus following the intervention highlights the efficacy of well-timed questioning in collaborative reasoning tasks.