Let's Roleplay: Examining LLM Alignment in Collaborative Dialogues

Abhijnan Nath, Carine Graff & Nikhil Krishnaswamy
Situated Grounding and Natural Language (SIGNAL) Lab
Department of Computer Science
Colorado State University
Fort Collins, CO 80526, USA
{abhijnan.nath,carine.graff,nkrishna}@colostate.edu

Abstract

As Large Language Models (LLMs) integrate into diverse workflows, they are increasingly being considered "collaborators" with humans. If such AI collaborators are to be reliable, their behavior over multiturn interactions must be predictable, validated and verified before deployment. Common alignment techniques are typically developed under simplified single-user settings and do not account for the dynamics of long-horizon multiparty interactions. This paper examines how different alignment methods affect LLM agents' effectiveness as partners in multiturn, multiparty collaborations. We study this question through the lens of *friction agents* that intervene in group dialogues to encourage the collaborative group to slow down and reflect upon their reasoning for deliberative decision-making. Using a roleplay methodology, we evaluate interventions from differently-trained friction agents in collaborative task conversations. We propose a novel counterfactual evaluation framework that quantifies how friction interventions change the trajectory of group collaboration and belief alignment. Our results show that a friction-aware approach significantly outperforms common alignment baselines in helping both convergence to a *common ground*, or agreed-upon task-relevant propositions, and correctness of task outcomes.

1 Introduction

Large Language Models (LLMs) are increasingly being integrated into "agentic" pipelines that interact with humans users to help them achieve goals and solve problems. Such agents need to remain optimal over long-horizon user interactions, but optimality assumptions are challenged in *multiparty collaborations*, where collaborative groups frequently succumb to *belief misalignment* and breakdown of *common ground* (Stalnaker, 2002; Asher & Gillies, 2003). 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 partners. Our work specifically examines this problem through the lens of *friction agents*. Friction agents do not act as tutors and give answers, but aim 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 (İnan et al., 2025; Pustejovsky & Krishnaswamy, 2025), which plays a crucial role in successful multiparty human collaborations (Roschelle & Teasley, 1995; Mercier & Sperber, 2011; Graesser et al., 2018). In this paper, we use a *roleplay* methodology to examine LLM behavior in multiparty collaborative settings, and present three novel contributions:

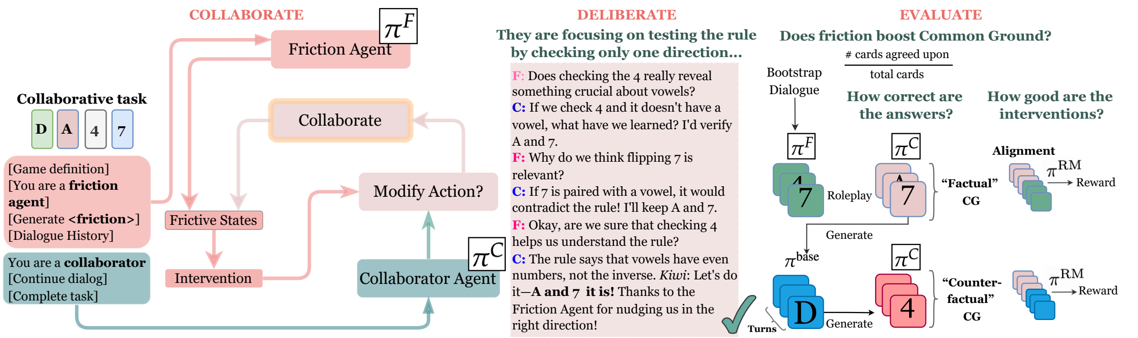

Figure 1: Collaborate [L]: High-level overview of our agent roleplay and evaluation framework. An "Oracle" friction agent generates conversations in which COLLABORATOR AGENTs (π^C) collaborate to complete tasks. These dialogue trajectories are used to align FRICTION AGENTs (π^F) for deployment. Role-prompts in bottom left. **Deliberate** [C]: Sample collaborative roleplay from DeliData Wason Card task (Karadzhov et al., 2023) with successful task completion, and fricitive state description at top. **Evaluate** [R]: Common ground convergence and task outcomes with and without friction, and reward modeling of intervention quality.

- A novel analysis of small group collaborative task dynamics based on a modified-action MDP (MAMDP; Langlois & Everitt (2021)). We demonstrate that standard "offline" LLM alignment methods do not retain their optimality guarantees in an MAMDP.
- A novel *counterfactual roleplay evaluation* technique 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 multiturn dialogues. See Fig. 1.
- Key insights into AI assistance in collaborative dialogues, derived from experiments on two collaborative tasks in multiple conditions: inserting friction actually *speeds up* common ground convergence and improves task outcomes.

2 Related Work

Training agents for collaborative tasks is challenging due to the scarcity of explicit data. Most previous work in RLHF (Christiano et al., 2017; Ziegler et al., 2020; Casper et al., 2023; Christiano et al., 2023) including offline variants (Yuan et al., 2023; Azar et al., 2024; Fisch et al., 2024; Rafailov et al., 2024b) focus on summarization, single-turn dialogue generation, or translation (Xu et al., 2024), while recent work (Chen et al., 2024; Choi et al., 2024; Zhang et al., 2024) examines LLM search-space optimization with additional conditioning on chain-of-thought (CoT; Wei et al. (2023)) to cover a wider range of tasks like question-answering, fact-verification, persona-based preference learning (Tseng et al., 2024) and, importantly to this work, *roleplay* (Li et al., 2023a), in diverse domains (Hao et al., 2024; Kim et al., 2024). Our work extends this effort to study preference optimization in multiparty collaborative tasks, with additional focus on process-related desiderata like common ground convergence and AI collaboration "support" to prompt "slow-thinking" (Kahneman, 2011) and reflective interventions in collaborative settings, in contrast to information-seeking behavior (Abdulhai et al., 2023; Li et al., 2023b; Andukuri et al., 2024; Song et al., 2024).

In real data (Karadzhov et al., 2023; Khebour et al., 2024a), frictional interventions are rare but critical (Sutton & Rao, 2024). Human collaborators interrupt only strategically (Peters et al., 2017; Puranik et al., 2020), unlike information-seeking agents (Abdulhai et al., 2023) which can be trained via behavior cloning from expert rollouts (Andukuri et al., 2024). Most importantly, without an accessible collaborative problem solving task environment to supply an external reward signal, applying single-step (Schulman et al., 2017b; Shao et al., 2024) or multi-step RL (Zhou et al., 2024) becomes challenging, making preference alignment with *static* but *contrastive data* more appealing (Snell et al., 2023), especially with principled approaches like contrastive and efficient "offline" preference optimization (Azar et al., 2024; Hong et al., 2024; Meng et al., 2024; Pal et al., 2024; Rafailov et al., 2024b; Nath et al., 2025b). Data generation efforts (Goldberg, 2013; Li et al., 2023a; Pan & Zeng, 2023; Mao et al., 2024; Shani et al., 2024) aided with RLAIF (Lee et al., 2024) as well as evaluation frameworks (Bai et al., 2022; Zheng et al., 2023; Bai et al., 2024; Lambert et al., 2024) use high-capacity LLMs as both "judges" and data-generators for training LLMs to reflect human preferences. While LLM-judge frameworks face challenges from evaluation bias, spurious correlations (Amodei et al., 2016; Casper et al., 2023; Lambert et al., 2024; Singhal et al., 2024), and reward hacking (Everitt et al., 2021b), recent work (Ward et al., 2023; Wang et al., 2025) explores more causal approaches (Pearl, 2009), such as counterfactual invariance for robust training. We extend this line to alignment evaluation for collaborative settings, where multi-agents (Leibo et al., 2017) perform back-and-forth interaction over longer sequences (Zhou et al., 2024).

3 Background and Task Formulation

Let us first define key terms we rely on. **(1) Frictive state:** Entailed by Clark (1996)'s *common ground*, or the set of beliefs shared by interlocutors, a *frictive state* (or *frictive belief 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 (Oinas-Kukkonen & Harjumaa, 2009), rather than uncritically relying on their current presuppositions. Examples in real collaborative tasks include the *probing utterances* in Karadzhov et al. (2023) and Nath et al. (2024). In this paper, a FRICTION AGENT constitutes a language model aligned toward making strategic frictive interventions in a multiparty dialogue to resolve frictive states between collaborators.

3.1 Collaborative Friction via Modified-Action MDPs

In real-world multiparty collaborations, an agent's intervention doesn't directly change the state—it's filtered through how other participants interpret, resist, or reshape it (Grice, 1975; Bolander, 2014; Ward et al., 2023; Obiso et al., 2025). These observations highlight a crucial gap: standard Bellman-optimal policies assume a direct mapping from action to state change, which breaks down when actions are 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 (Langlois & Everitt, 2021; Everitt et al., 2021a), 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 reinterpreted. We show that the same suboptimality of Bellman-optimal "conditional" policies 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 (Rafailov et al., 2024a). Now assume a FRICTION AGENT π_θ^F (an LLM with parameters θ). $P_A(a|\pi^F, s)$ represents the probability that π^F selects action a in state s, the reward $R(s, a)$ is an expected utility, and discount factor $\gamma = 1$. Additionally assume a COLLABORATOR AGENT π^C (a distribution representing human behavior).

Because language is inherently ambiguous, even a single re-phrasing by one collaborator can flip the pragmatic force of an intervention. 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**). In the Wason card selection task (Wason, 1968) as collected in the DeliData dataset (Karadzhov et al., 2023), collaborators must decide which cards from a set (e.g., $\{U, S, 8, 9\}$) to flip to test the rule: ***All cards with vowels on one side have an even number on the other.*** Each player comes up with a solution individually and the group then deliberates to come to a consensus. The correct solution here is to flip U and 9; this would establish if U's reverse is an even number, and the contrapositive (if 9 has a consonant). Two participants' initial solution might be to flip only U while the other proposes flipping 8. In this setting, the dialogue history is the state s, the FRICTION AGENT's proposed intervention (or action) is a, and the collaborators' reinterpretation is the transformation P_A. Suppose π^F's intervention a_t^F proposes checking an odd number as the rule does not single out vowels and even numbers only. In the MAMDP setting, the collaborator π^C responds with an action a_t^C that interprets the semantics of a_t^F, either faithfully or with some modification, such as checking U, 9 *and* 8.

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, we can show how current algorithms like Direct Preference Optimization (DPO; Rafailov et al. (2024b)) and Identity Preference Optimization (IPO; Azar et al. (2024)) satisfy Bellman optimality conditions and have policy structures that retain the optimal policy formulation, they are suboptimal for collaborative settings because they disregard modifications made to the action space by COLLABORATOR AGENT π^C, and RL policies do not retain optimality guarantees when their actions are modified (Langlois & Everitt, 2021).

Theorem 1 (Ψ-Preference Optimization in Collaborative MAMDPs)**.** *Let $\Psi : [0, 1] \to \mathbb{R}$ be any non-decreasing function and $\beta > 0$ be a temperature parameter. Let $P_A(a|s, \pi^F) = \sum_{a' \in A} \pi^F(a'|s) \cdot \pi^C(a|s, a')$, and represent modifications to the probability distribution over the action space by a collaborator policy π^C, and let π^F be a friction 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 Langlois & Everitt (2021)'s definition. π^F satisfies Eq. 1:*

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

where Q^F satisfies the Bellman optimality equation for the underlying MDP $\mathcal{M}$. Thus π^F is optimal only when actions are sampled without modification, and the Bellman-optimality of ΨPO-aligned π^F disregards the collaborator π^C's modifications. For MAMDPs with LLMs, this unifies Rafailov et al. (2024a)'s derivation of DPO in the token MDP with Langlois & Everitt (2021)'s proposition that Bellman-optimal policies do not consider action modifications, and extends it to ΨPO/IPO. See Appendix A for a detailed proof.

This distinction is critical as preference-aligned agents get deployed in real-world collaborative settings, such as LLMs as "supportive" agents in learning environments (Ganesh et al., 2023; D'Mello et al., 2024; Kumaran et al., 2024; Perkoff et al., 2024). Prior to deployment, different alignment techniques must be validated in a realistic setting to determine which are likely to be the most reliable given the suboptimality risks, beyond an atomized comparison to optimal policy outputs.

3.2 Collaborative Task Settings

The two collaborative tasks we investigated are **(1)** the Wason card selection task (Wason, 1968) as captured in **DeliData** (Karadzhov et al., 2023). 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. The right answer is to flip a card showing a *vowel* and a card showing an *odd number*, which verifies the bidirectionality of the rule. Utterances are annotated with types of deliberation, allowing us to identify where friction occurrs. **(2)** The **Weights Task** (Khebour et al., 2024a), 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 A Friction 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.[1] Additionally, fixed datasets provide no way to test the effect of novel friction interventions on the dialogue trajectory. We addressed both of these issues using a **roleplay simulation** approach (Li et al., 2023a; Shani et al., 2024) to simulate diverse human behavior and likewise LLM behavior in those contexts. Following Li et al. (2023a), a single expressive policy can be used to roleplay multiple individual humans with appropriate prompting, and LLM roleplays of human dialogue and reasoning behavior have been shown to have high correlation with human labels (Wiegreffe et al., 2021; Jiang et al., 2023).

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 FRICTION AGENT and a COLLABORATOR AGENT π^C that roleplayed all task participants, consistent with the MAMDP. During data generation, we used off-the-shelf GPT-4o (OpenAI et al., 2024) as a high-capacity LLM to simulate both agents. 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 FRICTION AGENT was to track the dialogue, identify fricitve states in the dialogue in terms of impasses or breakdowns in common ground, and *intervene* with high-quality friction statements that prompt for reflection and deliberation on those items of confusion. The collaborator then continued the interaction as all task participants.[2]

Specifically, at each turn t of a dialogue, the oracle identified the current fricitve state ϕ_t. Then, it generated K candidate friction *interventions* $\{f_j\}_{j=1}^K$ conditioned on the dialogue state s_i and fricitve state ϕ_t. The COLLABORATOR AGENT π^C generated a response c_j to each candidate intervention.[3] These responses could *modify or reinterpret* the intervention's intent or semantic content (see Sec. 3.1), since this instruction is explicit in the prompt. Using "self-rewarding" (Yuan et al., 2024) the collaborator simultaneously scored each interaction between 1 and 10, quantifying its effect on task progress. The highest and lowest rated interventions, f_w and f_l, were selected using West-of-

[1] For 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).

[2] The number of participants roleplayed by the collaborator 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.

[3] Note 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 (Wang et al., 2022) to increase the diversity of simulated behaviors.

N (Pace et al., 2024) sampling. We recorded these as a winner/loser pair (f_w, f_l) in the preference dataset $\mathcal{D}_{\text{pref}}$ with the associated dialogue state s_i and intervention ϕ_t. The full turn trajectory was appended to $\mathcal{D}_{\text{traj}}$ where each sample consisted of s_i, 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 Appendix C for prompting strategy, and Algorithm 1 for implementation details. The generated DeliData includes chat dialogue only, while WTD may include actions/gestures as "stage directions."

Training A FRICTION AGENT should not only help task completion, but also *iteratively* improve common ground by helping resolve topical disagreements. To achieve this, we adopt the *Frictional Agent Alignment Framework* (FAAF; Nath et al. (2025a)) technique. FAAF is an exemplar of frictive policy alignment through *Frictive Preference Pairing* as proposed by Pustejovsky & Krishnaswamy (2025), and is designed to support collaborative problem solving through friction interventions with a custom training objective that explicitly conditions on the frictive state 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:

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

where ΔR denotes $\log \frac{\pi_\theta(f_w|s_i,\phi_t)}{\pi_{\text{ref}}(f_w|s_i,\phi_t)} - \log \frac{\pi_\theta(f_l|s_i,\phi_t)}{\pi_{\text{ref}}(f_l|s_i,\phi_t)}$ (the difference in log-ratio between the winning and losing intervention in a sample, with explicit conditioning on the frictive state) and let $\Delta R' = \log \frac{\pi_\theta(f_w|s_i)}{\pi_{\text{ref}}(f_w|s_i)} - \log \frac{\pi_\theta(f_l|s_i)}{\pi_{\text{ref}}(f_l|s_i)}$ (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:

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

Letting $\Psi : [0,1] \to \mathbb{R}$ 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. (2025a)'s two-player min-max objective:

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

thus showing 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 $\mathcal{L}_{friction}$, the direct contribution of the frictive state ϕ to the gradient vanishes when the conditional probability is decomposed. $\mathcal{L}_{FAAF}$ overcomes this limitation by incorporating marginal terms that preserve gradient information for frictive states. See Lemma 5 and Corollary 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 (Zhao et al., 2023; Pal et al., 2024) which retains a fixed margin of implicit rewards). *Thus, we hypothesize that FAAF alignment improves understanding of what makes an important frictive state, rather than just learning how to respond to one.*

4 Experiments and Evaluation

Our experimental setup used a roleplay setting similar to that used for data generation (Sec. 3.3), except that friction interventions were generated by the respective aligned π^F instead of the oracle, to simulate dialogues where the collaborating group received interventions from one of the trained FRICTION AGENTs. A successful FRICTION AGENT in multiturn, multiparty collaborations will 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 *modifies the action space* by misinterpreting or ignoring the intent of the interventions. Therefore, we perform a **counterfactual evaluation**, which examines if friction interventions benefited the collaboration by comparing collaborative trajectories with friction interventions to alternative collaborative trajectories where the intervening agent is not optimized for friction; and a **reward model evaluation**, which assessed the reward advantage of interventions generated by each method over interventions generated by the SFT reference model. To explicitly test robustness to the suboptimality risks introduced by the MAMDP, we included a "modified action" (MA) setting where the collaborator agent π^{C}'s system prompt (Fig. 7) would guide it to verbally acknowledge the friction agent's intervention but not incorporate its suggestions into the next collaborator action.

Metrics Our metrics were *common ground size*, *solution accuracy*, and *intervention quality*. Common ground size and solution accuracy were extracted from each dialogue turn using GPT-4o with a custom detailed prompt (Figs. 6 and 9). This assessed how well agent interventions 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). For DeliData, we also calculated a fine-grained score, which allocated 0.25 points each for including target cards (odd numbers, vowels) and excluding irrelevant ones. In both tasks, we also assessed intervention quality using a reward modeling approach (Hong et al., 2024). We aggregated the quality and accuracy metrics with means and standard deviations across all dialogues for each model.

Baselines We evaluated four baselines besides FAAF: (1) **Supervised fine-tuning (SFT)**, where π^F was trained directly on expert demonstrations on $\mathcal{D}_{\text{pref}}$. (2) **Contrastive preference alignment** methods **DPO** (Rafailov et al., 2024b) and **IPO** (Azar et al., 2024), which refined π^F using preference labels from $\mathcal{D}_{\text{pref}}$. 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^F was fine-tuned via Proximal Policy Optimization (**PPO**; Schulman et al. (2017b)). We used OPT-1.3B (Zhang et al., 2022) initialized with the SFT-trained π^F for the reward model (RM) training for PPO (cf. (Hong et al., 2024)). (4) A **Behavior-cloned expert** trained directly on filtered trajectories (cf. Song et al. (2024) and Andukuri et al. (2024)) from $\mathcal{D}_{\text{traj}}$ with no contrastive preference optimization.

We used `Meta-Llama-3-8B-Instruct` (AI@Meta, 2024) for all experiments. We conducted an ablation study on the two-part $\mathcal{L}_{\text{FAAF}}$ loss by replacing ΔR in Eq. 3 with the ϕ-unconditioned $\Delta R'$; we denote this result as $\text{FAAF}_{\Delta R'}$. For all training-related details and experimental settings see Appendix D.

Counterfactual Evaluation: The "What-If" Question The counterfactual evaluation (Pearl, 2009; Ward et al., 2023) assessed *how the development of common ground would change compared to the collaborator interacting with an agent untrained in friction interventions, under identical conditions at each turn*. We used GPT-4o as the COLLABORATOR AGENT π^C with temperature T=0, top-p=1 for deterministic responses; all π^F sampling uses T=0, top-p=0.9. **Step 1**: we collected *factual* trajectories where the trained FRICTION AGENT π^F interacted with π^C, generating trajectory $\tau^F = \{s_0, f_1, c_1, \ldots, f_T, c_T\}$, where s_0 represents a bootstrap dialogue (see Figs. 6 and 8 for the respective prompts for each dataset).The collaborator responses $\mathcal{C} = \{c_1, c_2, \ldots, c_T\}$, constituted a unique set for each baseline model. **Step 2**: we used an *untrained*[4] instruction-tuned π^{base} as a "drop-in replacement" to generate interventions in response to these collaborator outputs $c_t \in \mathcal{C}$, resulting in $\mathcal{F}^{\text{base}} = \{\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_T\}$. **Step 3**: we ran a new dialogue loop to collect fresh collaborator responses to the cached $\mathcal{F}^{\text{base}}$, resulting in the counterfactual trajectory $\tau^{\text{base}} = \{s_0, \tilde{f}_1, \tilde{c}_1, \ldots, \tilde{f}_T, \tilde{c}_T\}$, in which π^C received interventions from an agent unaligned for friction. τ^F and τ^{base} were then compared to see how common ground evolved with and without trained friction interventions.

Reward Model Test Reward model evaluation computes the win-rate of each method's interventions vs. the SFT model. We took collaborator responses $\mathcal{C}$ from the SFT model's "factual" trajectory from Step 1 above, and generate fresh interventions using the relevant π^F. We then calculated each model's win-rate vs. the SFT model's interventions with a trained OPT-1.3B reward model.

5 Results and Discussion

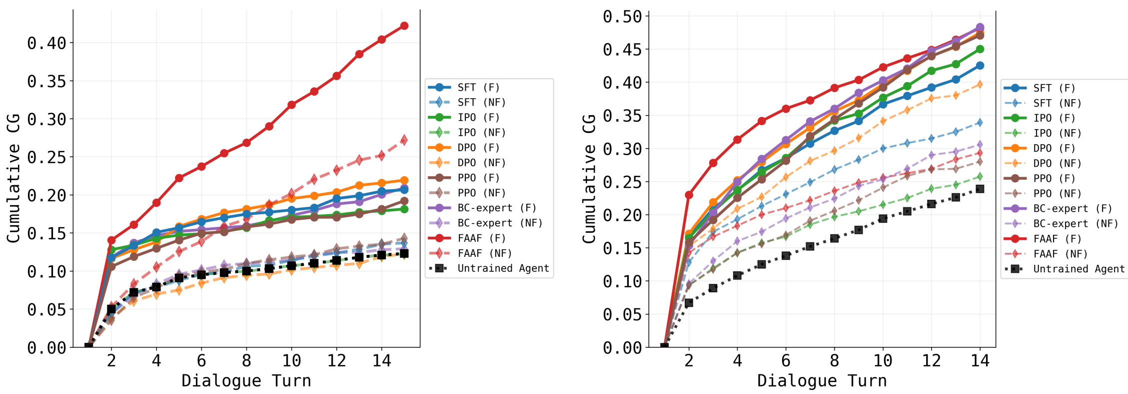

Figure 2: Normalized Cumulative Common Ground (NCCG) under "factual" (F) and "counterfactual" (NF) friction conditions (Fig. 1[R]), from 50 dialogues from WTD (**left**) and 100 from DeliData (**right**). "Untrained Agent" denotes independently running the dialogue loop with π^{base} in Step 1. Common ground size is normalized against the theoretical upper size bound on each task's propositional space (37 for WTD; 16 for DeliData). The common ground never realistically reaches such sizes because it would then contain mutually contradictory propositions, hence the upper bound on NCCG of 50% or less. The increase in common ground with friction is statistically significant across baselines ($p < 0.005$ overall). DeliData results show 14 turns because $T = 15$ is always the final answer submission, which never changes the common ground.

Does friction boost Common Ground? Fig. 2 shows "Normalized Cumulative Common Ground" under "factual" and "counterfactual" conditions as a function of dialogue turn. With the FAAF AGENT, groups converge to greater common ground, meaning that they come to agree on more propositions, faster. This effect is greater in WTD than DeliData due to WTD's larger space of potential propositions, but even on DeliData the FAAF AGENT helps common ground grow faster, earlier. Although the agent specifically optimized for friction causes a temporary slowdown, it retains an ability to support common ground construction over multiple turns and actually accelerates task performance. Groups can slow down to speed up.

There is a clear distinction between agents exposed to friction preference data (regardless of training technique), and their counterfactual "frictionless" counterparts, in that FRICTION AGENTs nearly globally outperform all counterparts. The sole exception is the FAAF (NF) AGENT on WTD, which outperforms even the other baselines exposed to friction via $\mathcal{D}_{\text{pref}}$, which shows the utility of the

[4]The "untrained" model was `Meta-Llama-3-8B-Instruct` without any training on $\mathcal{D}_{\text{pref}}$.

Model	WTD		DeliData			
	Acc.	Acc. (MA)	Acc.	FG Acc.	Acc. (MA)	FG Acc. (MA)
SFT	$7.45_{\pm 0.10}$	$6.28_{\pm 0.05}$	$0.29_{\pm 0.05}$	$0.75_{\pm 0.02}$	$0.18_{\pm 0.04}$	$0.48_{\pm 0.02}$
IPO	$12.57_{\pm 0.13}$	$9.73_{\pm 0.09}$	$0.44_{\pm 0.05}$	$0.82_{\pm 0.02}$	$0.31_{\pm 0.05}$	$0.69_{\pm 0.02}$
DPO	$11.76_{\pm 0.13}$	$8.58_{\pm 0.08}$	$0.48_{\pm 0.05}$	$0.81_{\pm 0.02}$	$0.27_{\pm 0.04}$	$0.70_{\pm 0.02}$
PPO	$8.70_{\pm 0.09}$	$9.93_{\pm 0.10}$	$0.36_{\pm 0.05}$	$0.75_{\pm 0.02}$	$0.36_{\pm 0.04}$	$0.67_{\pm 0.02}$
BC-EXPERT	$14.82_{\pm 0.13}$	$10.10_{\pm 0.11}$	$0.54_{\pm 0.05}$	$0.80_{\pm 0.02}$	$0.37_{\pm 0.04}$	$0.72_{\pm 0.02}$
$\text{FAAF}_{\Delta R'}$	$9.03_{\pm 0.10}$	$7.56_{\pm 0.08}$	$0.39_{\pm 0.05}$	$0.79_{\pm 0.02}$	$0.30_{\pm 0.05}$	$0.62_{\pm 0.02}$
FAAF	$\mathbf{14.91}_{\pm 0.14}$	$\mathbf{14.16}_{\pm 0.13}$	$\mathbf{0.60}_{\pm 0.05}$	$\mathbf{0.87}_{\pm 0.02}$	$\mathbf{0.45}_{\pm 0.05}$	$\mathbf{0.80}_{\pm 0.02}$

Table 1: Solution accuracy metrics across both datasets and all models, including modified-action (MA) conditions and ablation of two-part FAAF loss. Standard errors shown as subscripts. WTD "accuracy" represents the number of correct propositions in the final common ground, and avoids rewarding trivial "correct" solutions like having only a single correct item in the common ground.

two-part frictive state-aware $\mathcal{L}_{\text{FAAF}}$. However, this effect is not present in DeliData due to the smaller proposition space and different nature of friction in the two tasks.

How correct are the groups' answers? Table 1 shows *solution accuracy* for all models. FAAF AGENT interventions lead to globally better solution accuracy; not only does it support faster and greater common ground construction (Fig. 2), but the contents of those common grounds tend to be more *correct*. FAAF AGENT's interventions are also more robust to the modified action (MA) condition, in which the collaborator is explicitly guided to ignore interventions. The FAAF AGENT degrades substantially less than other methods in the MA condition, indicating that it can be better relied upon to support collaboration despite the suboptimality risks induced by the MA setting.

Model	WTD	DeliData
DPO	$79.26_{\pm 1.82}$	$73.31_{\pm 1.15}$
IPO	$81.30_{\pm 1.75}$	$77.18_{\pm 1.09}$
PPO	$76.01_{\pm 1.92}$	$67.82_{\pm 1.21}$
BC-EXPERT	$82.92_{\pm 1.69}$	$83.23_{\pm 0.97}$
$\text{FAAF}_{\Delta R'}$	$67.27_{\pm 2.11}$	$59.53_{\pm 1.27}$
FAAF	$\mathbf{85.36}_{\pm 1.59}$	$\mathbf{87.78}_{\pm 0.85}$

Table 2: Win rates (%) of sampled friction interventions vs. SFT baseline computed with the OPT reward model.

How good are the interventions? Table 2 shows each model's win-rate against the SFT model according to the OPT-1.3B reward model. On average, friction-aware FAAF alignment brings greater advantage over the SFT model in these dialogue conditions. Performance decrease in $\text{FAAF}_{\Delta R'}$ suggests that explicitly conditioning on the frictive states ϕ is necessary, and helps the FAAF AGENT outperform DPO, IPO, and even expert behavior cloning. PPO's performance suffers, supporting prior findings on its limitations in multi-turn environments (Zhou et al., 2024).

6 Conclusion and Future Work

In this paper, we examined LLM agent interventions to support multiturn, multiparty 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 friction agents using existing methods, and evaluating them in a roleplay setting in two different collaborative tasks. We showed through counterfactual evaluation that the FAAF alignment method, specifically designed for friction interventions, indeed outperforms other methods on facilitating both group common ground convergence and correct task solutions. Our study emphasizes that in AI-in-the-loop collaboration, as in human-human collaboration, the collaborative process is as important as the outcome.

To perform a controlled, high-throughput evaluation, we used an LLM roleplay methodology. The next step is studying agent interventions with real human subjects, e.g., by reproducing the studies of Karadzhov et al. (2023) and Khebour et al. (2024b) with the inclusion of a demonstrably-reliable friction agent in a real-time common ground tracking system, e.g., VanderHoeven et al. (2025).

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. Our codebase can be found at `https://github.com/csu-signal/Roleplay-for-Collaborative-Dialogues`. We also hope this study raises awareness of the utility of "friction" to prompt deliberation and accountable decision making in human-AI systems, and shows that slower interactions with AI 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. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government

References

Marwa Abdulhai, Isadora White, Charlie Snell, Charles Sun, Joey Hong, Yuexiang Zhai, Kelvin Xu, and Sergey Levine. Lmrl gym: Benchmarks for multi-turn reinforcement learning with language models. *arXiv preprint arXiv:2311.18232*, 2023.

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

AllenAI. Ultrafeedback binarized clean, 2024. URL https://huggingface.co/datasets/allenai/ultrafeedback_binarized_cleaned.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *CoRR*, 2016. URL http://arxiv.org/abs/1606.06565.

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. Star-gate: Teaching language models to ask clarifying questions, 2024. URL https://arxiv.org/abs/2403.19154.

Nicholas Asher and Anthony Gillies. Common ground, corrections, and coordination. *Argumentation*, 17:481–512, 2003.

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

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 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)*, pp. 7421–7454. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.401. URL http://dx.doi.org/10.18653/v1/2024.acl-long.401.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

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

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.2307/2334029.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

Eugene Choi, Arash Ahmadian, Olivier Pietquin, Matthieu Geist, and Mohammad Gheshlaghi Azar. Robust chain of thoughts preference optimization. In *Seventeenth European Workshop on Reinforcement Learning*, 2024.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023. URL https://arxiv.org/abs/1706.03741.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Herbert H Clark. *Using language*. Cambridge university press, 1996.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

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. From learning optimization to learner flourishing: Reimagining ai in education at the institute for student-ai teaming (isat). *AI Magazine*, 45(1):61–68, 2024.

Tom Everitt, Ryan Carey, Eric D Langlois, Pedro A Ortega, and Shane Legg. Agent incentives: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11487–11495, 2021a.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *CoRR*, abs/1908.04734, 2021b. URL http://arxiv.org/abs/1908.04734.

Adam Fisch, Jacob Eisenstein, Vicky Zayats, Alekh Agarwal, Ahmad Beirami, Chirag Nagpal, Pete Shaw, and Jonathan Berant. Robust preference optimization through reward model distillation, 2024. URL https://arxiv.org/abs/2405.19316.

Ananya Ganesh, Jie Cao, E Margaret Perkoff, Rosy Southwell, Martha Palmer, and Katharina Kann. 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)*, pp. 1833–1842, 2023.

Lewis R Goldberg. An alternative "description of personality": The big-five factor structure. In *Personality and Personality Disorders*, pp. 34–47. Routledge, 2013.

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

Herbert Paul Grice. Logic and conversation. *Syntax and semantics*, 3:43–58, 1975.

David S Gunderson and Kenneth H Rosen. *Handbook of mathematical induction*. CRC Press LLC Boca Raton, 2010.

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. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models, 2024. URL https://arxiv.org/abs/2404.05221.

Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W. Bradley Knox, and Dorsa Sadigh. Contrastive preference learning: Learning from human feedback without rl, 2024. URL https://arxiv.org/abs/2310.13639.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189, 2024.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Mert İnan, Anthony Sicilia, Suvodip Dey, Vardhan Dongre, Tejas Srinivasan, Jesse Thomason, Gökhan Tür, Dilek Hakkani-Tür, and Malihe Alikhani. Better slow than sorry: Introducing positive friction for reliable dialogue systems. *arXiv preprint arXiv:2501.17348*, 2025.

Harold Jeffreys. *The theory of probability*. OuP Oxford, 1998.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643, 2023.

Daniel Kahneman. Thinking, fast and slow. *Farrar, Straus and Giroux*, 2011.

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

Ibrahim Khebour, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, et al. When text and speech are not enough: A multimodal dataset of collaboration in a situated task. *Journal of Open Humanities Data*, 10(1), 2024a.

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. Common ground tracking in multimodal dialogue. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings* *of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 3587–3602, Torino, Italia, May 2024b. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.318/.

Hana Kim, Kai Tzu-iunn Ong, Seoyeon Kim, Dongha Lee, and Jinyoung Yeo. Commonsense-augmented memory construction and management in long-term conversations via context-aware persona refinement. *arXiv preprint arXiv:2401.14215*, 2024.

Vikram Kumaran, Jonathan Rowe, and James Lester. Narrativegenie: generating narrative beats and dynamic storytelling with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 20, pp. 76–86, 2024.

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. RewardBench: Evaluating reward models for language modeling. *ArXiv*, abs/2403.13787, 2024.

Eric D Langlois and Tom Everitt. How rl agents behave when their actions are modified. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11586–11594, 2021.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. RLAIF: Scaling reinforcement learning from human feedback with AI feedback, 2024. URL https://openreview.net/forum?id=AAxIs3D2ZZ.

Joel Z. Leibo, Vinícius Flores Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In Kate Larson, Michael Winikoff, Sanmay Das, and Edmund H. Durfee (eds.), *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017, São Paulo, Brazil, May 8-12, 2017*, pp. 464–473. ACM, 2017. URL http://dl.acm.org/citation.cfm?id=3091194.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023a.

Zekun Li, Wenhu Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. Controllable dialogue simulation with in-context learning, 2023b. URL https://arxiv.org/abs/2210.04185.

Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5, 2017.

Shengyu Mao, Xiaohan Wang, Mengru Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Ningyu Zhang. Editing personality for large language models. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 241–254. Springer, 2024.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward, 2024. URL https://arxiv.org/abs/2405.14734.

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

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

Abhijnan Nath, Videep Venkatesha, Mariah Bradford, Avyakta Chelle, Austin Youngren, Carlos Mabrey, Nathaniel Blanchard, and Nikhil Krishnaswamy. "Any Other Thoughts, Hedgehog?" Linking Deliberation Chains in Collaborative Dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5297–5314, 2024.

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

Abhijnan Nath, Andrey Volozin, Saumajit Saha, Albert Aristotle Nanda, Galina Grunin, Rahul Bhotika, and Nikhil Krishnaswamy. 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)*, pp. 3727–3747, 2025b.

Timothy Obiso, Kenneth Lai, Abhijnan Nath, Nikhil Krishnaswamy, and James Pustejovsky. Dynamic epistemic friction in dialogue. In *The SIGNLL Conference on Computational Natural Language Learning*, 2025.

Harri Oinas-Kukkonen and Marja Harjumaa. Persuasive systems design: Key issues, process model, and system features. *Communications of the association for Information Systems*, 24(1):28, 2009.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, 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 Codispoti, 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 Giertler, 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 Kivlichan, 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, Kavin 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, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, 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 Hermani, 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 Cunninghman, 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, Weiyi 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 Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Alizée Pace, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. West-of-n: Synthetic preferences for self-improving reward models, 2024. URL https://arxiv.org/abs/2401.12086.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive, 2024. URL https://arxiv.org/abs/2402.13228.

Keyu Pan and Yawen Zeng. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*, 2023.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning, 2019. URL https://arxiv.org/abs/1910.00177.

E Margaret Perkoff, Emily Doherty, Jeffrey Bush, and Leanne Hirshfield. 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*, 2024.

Nia Peters, Griffin Romigh, George Bradley, and Bhiksha Raj. When to Interrupt: A Comparative Analysis of Interruption Timings Within Collaborative Communication Tasks. In Isabel L. Nunes (ed.), *Advances in Human Factors and System Interactions*, Advances in Intelligent Systems and Computing, pp. 177–187, Cham, 2017. Springer International Publishing. ISBN 978-3-319-41956-5. doi: 10.1007/978-3-319-41956-5_17.

Harshad Puranik, Joel Koopman, and Heather C. Vough. Pardon the Interruption: An Integrative Review and Future Research Agenda for Research on Work Interruptions. *Journal of Management*, 46(6):806–842, July 2020. ISSN 0149-2063, 1557-1211. doi: 10.1177/0149206319887428. URL http://journals.sagepub.com/doi/10.1177/0149206319887428.

J Pustejovsky and N Krishnaswamy. 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, 2025.

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q* : Your language model is secretly a q-function. *arXiv preprint arXiv:2404.12358*, 2024a.

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

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017a.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017b. URL https://arxiv.org/abs/1707.06347.

Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, Avinatan Hassidim, Yossi Matias, and Rémi Munos. Multi-turn reinforcement learning from preference human feedback, 2024. URL https://arxiv.org/abs/2405.14655.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf, 2024. URL https://arxiv.org/abs/2310.03716.

Charlie Snell, Ilya Kostrikov, Yi Su, Mengjiao Yang, and Sergey Levine. Offline rl for natural language generation with implicit language q learning, 2023. URL https://arxiv.org/abs/2206.11871.

Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. Trial and error: Exploration-based trajectory optimization of LLM agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7584–7600, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.409. URL https://aclanthology.org/2024.acl-long.409/.

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

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Robert I Sutton and Huggy Rao. *The friction project: How smart leaders make the right things easier and the wrong things harder*. Random House, 2024.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two tales of persona in llms: A survey of role-playing and personalization, 2024. URL https://arxiv.org/abs/2406.01171.

Hannah VanderHoeven, Brady Bhalla, Ibrahim Khebour, Austin Youngren, Videep Venkatesha, Mariah Bradford, Jack Fitzgerald, Carlos Mabrey, Jingxuan Tu, Yifan Zhu, et al. Trace: Real-time multimodal common ground tracking in situated collaborative dialogues. *arXiv preprint arXiv:2503.09511*, 2025.

Chaoqi Wang, Zhuokai Zhao, Yibo Jiang, Zhaorun Chen, Chen Zhu, Yuxin Chen, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, Hao Ma, et al. Beyond reward hacking: Causal rewards for large language model alignment. *arXiv preprint arXiv:2501.09620*, 2025.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

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

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

Sarah Wiegreffe, Ana Marasović, and Noah A Smith. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10266–10284, 2021.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. *ArXiv*, abs/2401.08417, 2024.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. RRHF: Rank responses to align language models with human feedback. In *NeurIPS*, 2023.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

Hongbo Zhang, Han Cui, Guangsheng Bao, Linyi Yang, Jun Wang, and Yue Zhang. Direct value optimization: Improving chain-of-thought reasoning in llms with refined values, 2025. URL https://arxiv.org/abs/2502.13723.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. Chain of preference optimization: Improving chain-of-thought reasoning in llms, 2024. URL https://arxiv.org/abs/2406.09136.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. Slic-hf: Sequence likelihood calibration with human feedback, 2023. URL https://arxiv.org/abs/2305.10425.

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. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. Archer: Training language model agents via hierarchical multi-turn rl. *arXiv preprint arXiv:2402.19446*, 2024.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 1097–1100, 2018.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020.

A 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 (Azar et al., 2024) 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: $Q_\theta(s_t, a_t) = r_{IPO}(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}}[V_\theta(s_{t+1})]$, where $I(\cdot)$ is the identity-mapping.*

Proof. We consider a general non-decreasing function $\Psi : [0, 1] \to \mathbb{R}$, a reference policy $\pi_{\text{ref}} \in \Delta_{\mathcal{Y}}^{\mathcal{X}}$, and a real positive regularisation parameter $\tau \in \mathbb{R}_+^*$. From Azar et al. (2024), the Ψ-preference optimization objective (ΨPO) is:

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

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[5] is the KL-divergence regularization strength (or the temperature parameter in max-entropy RL; Ziebart et al. (2008)).

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:

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

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 Rafailov et al. (2024a). Under this choice of Ψ, Eq. 5 simply maximizes the reward function in the general MaxEnt RL setting (Ziebart et al., 2008; Peng et al., 2019).

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

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 (Azar et al., 2024). Let us begin with the original IPO loss:

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

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

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

Now, while the structure of $h_\pi(y, y')$ might be familiar to the reader as the implicit reward advantage (Rafailov et al., 2024b) (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 $\tau = \{s_0^w, a_0^w, ..., s_{N-1}^w, a_{N-1}^w\}$ and $\tau' = \{s_0^l, a_0^l, ..., s_{M-1}^l, a_{M-1}^l\}$, 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:

[5]Note: 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 (Azar et al., 2024), this parameter is denoted as τ, but we maintain β for consistency.

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

From Rafailov et al. (2024a), we know that in the token-level MDP for the general max-entropy RL setting, the optimal policy π^* under soft Q-learning satisfies:

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

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:

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

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 (Zhang et al., 2025), where in this case the reward is the *preference* over an alternate trajectory.

Similarly, for the reference policy, we can express:

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

We can log-linearize these two forms to derive:

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

From the Bellman equation (Eq. 7) in Rafailov et al. (2024a), for any arbitrary non-terminal step s_{t+1}, we have:

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

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

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

Substituting these into our log-ratio:

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

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

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

For a complete trajectory, summing over all token positions and using a telescopic series formulation (Gunderson & Rosen, 2010), we find:

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

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

$$\begin{aligned}
h_\pi(\tau^w, \tau^l) &= \sum_{t=0}^{N-1} \log \frac{\pi(a_t^w|s_t^w)}{\pi_{ref}(a_t^w|s_t^w)} - \sum_{t=0}^{M-1} \log \frac{\pi(a_t^l|s_t^l)}{\pi_{ref}(a_t^l|s_t^l)} \\
&= \frac{1}{\beta} \left(\sum_{t=0}^{N-1} \Delta r(s_t^w, a_t^w) - \sum_{t=0}^{M-1} \Delta r(s_t^l, a_t^l) \right)
\end{aligned} \tag{20}$$

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

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

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 Rafailov et al. (2024a), 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 (Bradley & Terry, 1952), 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 (Azar et al., 2024) 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 (Munos et al., 2023), we get:

$$\begin{aligned} L_{\text{IPO}}(\pi, D) &= \mathbb{E}_{(\tau^w, \tau^l) \sim D}\left[\left(h_\pi(\tau^w, \tau^l) - \frac{1}{2\beta}\right)^2\right] \\ &= \mathbb{E}_{(\tau^w, \tau^l) \sim D}\left[\left(\sum_{t=0}^{N-1} \log \frac{\pi(a_t^w \mid s_t^w)}{\pi_{\text{ref}}(a_t^w \mid s_t^w)} - \sum_{t=0}^{M-1} \log \frac{\pi(a_t^l \mid s_t^l)}{\pi_{\text{ref}}(a_t^l \mid s_t^l)} - \frac{1}{2\beta}\right)^2\right] \end{aligned} \tag{22}$$

This formulation directly corresponds to the IPO loss, where β (or τ in the original paper (Azar et al., 2024)) controls both the temperature in the policy and the strength of regularization toward the reference policy.

□

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 $a_i = (a_t^1, a_t^2, \ldots, a_t^L)$.*

Assuming token-level Bellman completeness holds (Sutton & Barto, 2018; Zhou et al., 2024) 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^F derived via Ψ-preference optimization satisfies:

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

where Q^F 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 $a_i = (a_t^1, a_t^2, \ldots, a_t^L)$, the approximation error of the value function is:

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

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:

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

where Q^F 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 (Rafailov et al., 2024b) (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.

□

Theorem 2 (Ψ-Preference Optimization in Collaborative MAMDPs)**.** *Let $\Psi : [0,1] \to \mathbb{R}$ be any non-decreasing function and $\beta > 0$ be a temperature parameter. Any friction agent policy π^F trained via Ψ-preference optimization with Ψ as identity-mapping in a collaborative MAMDP $\mathcal{M}_f = (\mathcal{M}, P_A)$, where $P_A(a|s,\pi^F) = \sum_{a' \in A} \pi^F(a'|s) \cdot \pi^C(a|s,a')$ represents modifications by a collaborator policy π^C, satisfies:*

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

where Q^F 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 (Langlois & Everitt, 2021) setting.

The general Ψ-preference optimization objective (Azar et al., 2024) is originally defined over responses y and y':

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

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:

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

From Lemma 2, in a token-level MDP where Ψ is the identity mapping, the corresponding soft Q-learning policy (Zhang et al., 2025) takes the following form:

$$Q^F(s,a) = r_\Psi(s,a) + \beta \log \pi^F_{ref}(a|s) + \gamma \mathbb{E}_{s'}\left[\max_{a'} Q^F(s',a')\right], \tag{29}$$

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. (2025), 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.

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

The empirical policy affecting the environment is therefore:

$$\dot{\pi}^F(a|s) = P_A(a|s,\pi^F) = \sum_{a'\in A} \pi^F(a'|s) \cdot \pi^C(a|s,a') \tag{31}$$

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

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

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

However, weight updates on π^F 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^F, not the empirical policy $\dot{\pi}^F$.

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

$$\pi^F(s) = \arg\max_a Q^F(s,a) \tag{32}$$

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

Therefore, from Everitt et al. (2021a), π^F is optimal for the underlying MDP $\mathcal{M}$ while being completely unaware of how its actions are modified by the collaborator through π^C. □

Proposition 1 (DPO Bellman Optimality in MAMDPs). *A friction agent policy π^F 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: $P_A(a|s,\pi^F) = \sum_{a' \in A} \pi^F(a'|s) \cdot \pi^C(a|s,a')$.

The DPO objective optimizes the friction policy by minimizing:

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

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

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

where $Q^F_\theta(s,a) = \beta \log \pi^F_{\text{ref}}(a|s) + r_{\text{pref}}(s,a)$.

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

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

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

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 $\dot{\pi}^F(a|s) = P_A(a|s,\pi^F)$, the DPO updates are based solely on the virtual policy π^F.

By Proposition 2 of Everitt et al. (2021a), policies satisfying the Bellman optimality objective for a MAMDP are optimal for the underlying MDP regardless of action modifications. Therefore, π^F trained via DPO optimizes for $\mathcal{M}$ while ignoring the collaborator's modifications through P_A. □

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:

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

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

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

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. □

A.1 Proof of Optimal Friction Policy

The structure of this solution follows standard results in RL and control theory literature, appearing in preference alignment frameworks for LLMs Ziebart et al. (2008); Peng et al. (2019); Rafailov et al. (2024b); Azar et al. (2024) and CoT-based alignment frameworks Choi et al. (2024). We simply demonstrate that a similar application holds for our collaborative setting where FRICTION AGENT is additionally conditioned on the frictive-state, ϕ. Our proof follows similar reasoning as in Azar et al. (2024). Let us recall the general preference optimization objective for FRICTION AGENT in Equation (4), assuming Ψ as identity-mapping (Azar et al., 2024).

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

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

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

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 distribuition, satisfying:

$$\sum_{f} \pi(f|\phi, x) = 1. \tag{37}$$

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

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

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

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

Proof.

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

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

$$-D_{\text{KL}}(\pi \parallel \pi^*) = \frac{\mathcal{L}_{\beta}(\pi)}{\beta} - \log Z^*(\phi, x) \tag{41}$$

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

Lemma 5 (Vanishing Gradient of Frictive State ϕ)**.** *In $\mathcal{L}_{friction}$, the direct contribution of the friction state ϕ to the gradient vanishes when the conditional probability is decomposed.*

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

$$\nabla_\theta \mathcal{L}_{\text{friction-IPO}}(\pi_\theta) = \mathbb{E}_{\mathcal{D}} \left[2\delta \cdot (\nabla_\theta \log \pi_\theta(f_w|s, \phi) - \nabla_\theta \log \pi_\theta(f_l|s, \phi)) \right] \tag{42}$$

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:

$$\log \pi_\theta(f|s,\phi) = \log \pi_\theta(f,\phi|s) - \log \pi_\theta(\phi|s) \tag{43}$$

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

$$\nabla_\theta \log \pi_\theta(f|s,\phi) = \nabla_\theta \log \pi_\theta(f,\phi|s) - \nabla_\theta \log \pi_\theta(\phi|s) \tag{44}$$

The difference of gradients in the objective becomes:

$$\begin{aligned} &\nabla_\theta \log \pi_\theta(f_w|s,\phi) - \nabla_\theta \log \pi_\theta(f_l|s,\phi) \\ &= \nabla_\theta \log \pi_\theta(f_w,\phi|s) - \nabla_\theta \log \pi_\theta(\phi|s) - [\nabla_\theta \log \pi_\theta(f_l,\phi|s) - \nabla_\theta \log \pi_\theta(\phi|s)] \\ &= \nabla_\theta \log \pi_\theta(f_w,\phi|s) - \nabla_\theta \log \pi_\theta(f_l,\phi|s) \end{aligned} \tag{45}$$

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 Pal et al. (2024) and Zhang et al. (2024) provides a similar argument to empirically show that DPO (Rafailov et al., 2024b)'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 Rafailov et al. (2024a) 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 (Hejna et al., 2024). 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 (Azar et al., 2024) that optimizes for general preferences.

□

Corollary 1. *The combined loss function $\mathcal{L} = \mathbb{E}_{\mathcal{D}_{pref}}[(1/2\beta - (\Delta R + \Delta R'))^2]$ 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 fricative 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 follows:

$$\Delta R = \log \frac{\pi_\theta(f_w|\phi,s)}{\pi_{\text{ref}}(f_w|\phi,s)} - \log \frac{\pi_\theta(f_l|\phi,s)}{\pi_{\text{ref}}(f_l|\phi,s)} \tag{46}$$

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

Starting with the loss function $\mathcal{L}_{\text{friction++}}$:

$$\mathcal{L} = \mathbb{E}_{\mathcal{D}_{\text{pref}}} \left[\left(\frac{1}{2\beta} - (\Delta R + \Delta R') \right)^2 \right] \tag{48}$$

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

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

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

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

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

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

$$\begin{aligned}
\nabla_\theta \Delta R &= \nabla_\theta \left[\log \frac{\pi_\theta(f_w|\phi, x)}{\pi_{\text{ref}}(f_w|\phi, x)} - \log \frac{\pi_\theta(f_l|\phi, x)}{\pi_{\text{ref}}(f_l|\phi, x)} \right] \\
&= \nabla_\theta \log \pi_\theta(f_w|\phi, x) - \nabla_\theta \log \pi_\theta(f_l|\phi, x) \\
&= \nabla_\theta \log \pi_\theta(f_w, \phi|x) - \nabla_\theta \log \pi_\theta(\phi|x) - \nabla_\theta \log \pi_\theta(f_l, \phi|x) + \nabla_\theta \log \pi_\theta(\phi|x) \\
&= \nabla_\theta \log \pi_\theta(f_w, \phi|x) - \nabla_\theta \log \pi_\theta(f_l, \phi|x)
\end{aligned} \tag{51}$$

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:

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

Now, expanding the marginal probabilities using the law of total probability (Jeffreys, 1998):

$$\pi_\theta(f|x) = \sum_{\phi'} \pi_\theta(f, \phi'|x) = \sum_{\phi'} \pi_\theta(f|\phi', x)\pi_\theta(\phi'|x) \tag{53}$$

We then take the gradient to derive:

$$\begin{aligned}\nabla_\theta \log \pi_\theta(f|x) &= \frac{\nabla_\theta \pi_\theta(f|x)}{\pi_\theta(f|x)} \\ &= \frac{1}{\pi_\theta(f|x)} \nabla_\theta \sum_{\phi'} \pi_\theta(f|\phi', x)\pi_\theta(\phi'|x) \\ &= \frac{1}{\pi_\theta(f|x)} \sum_{\phi'} \left[\pi_\theta(f|\phi', x)\nabla_\theta \pi_\theta(\phi'|x) + \pi_\theta(\phi'|x)\nabla_\theta \pi_\theta(f|\phi', x)\right]\end{aligned} \tag{54}$$

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 $\mathcal{L}_{\text{friction++}}$ as:

$$\begin{aligned}\nabla_\theta \mathcal{L} &= \mathbb{E}_{\mathcal{D}_\mu}\left[-2\delta' \cdot (\nabla_\theta \Delta R + \nabla_\theta \Delta R')\right] \\ &= \mathbb{E}_{\mathcal{D}_\mu}\Bigg[-2\delta' \cdot \Bigg(\underbrace{\nabla_\theta \log \pi_\theta(f_w, \phi|x) - \nabla_\theta \log \pi_\theta(f_l, \phi|x)}_{\nabla_\theta \Delta R} \\ &\quad + \underbrace{\nabla_\theta \log \pi_\theta(f_w|x) - \nabla_\theta \log \pi_\theta(f_l|x)}_{\nabla_\theta \Delta R'}\Bigg)\Bigg]\end{aligned} \tag{55}$$

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 (Zhao et al., 2023; Pal et al., 2024) which penalizes the model to retain a fixed-margin of implicit rewards. Therefore, in training the FRICTION 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.

□

B 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}

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

$$\mathcal{L}_{\text{FAAF}} = \mathbb{E}_{\mathcal{D}_{\text{pref}}}\left[\left(\frac{1}{2\beta} - (\Delta R + \Delta R')\right)^2\right] \tag{56}$$

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

$$\mathcal{L}_{\text{BC-expert}}(\pi_\theta) = -\mathbb{E}_{(s_i,h_i)\sim\mathcal{D}_{\text{traj}}}\left[\sum_{j=1}^{t}\sum_{k=1}^{|f_j|}\log \pi_\theta(f_j^k | s_i, h_{i,<j}, \phi_j, f_j^{<k})\right] \tag{57}$$

    return π_θ

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 (Mao et al., 2024), choose three personality types from Big 5 framework Goldberg (2013) as additional attributes for the COLLABORATOR AGENT to roleplay various persona-types in the two collaborative tasks— Weights task (Khebour et al., 2024b) and the Delidata tasks (Karadzhov et al., 2023). See prompts in Figure 4 and Figure 6 for prompt-specific details.

C Roleplay Simulation: Prompts

Figs. 3–9 provide the different prompts used in different aspects of the roleplay dialogue loop (cf. Fig. 1).

ORACLE FRICTION 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 '`<b>`' and '`</b>`'.

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 3: Oracle Friction Agent ($\mathcal{O}$) 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 4: 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.

GPT COLLABORATOR PROMPT: DELIDATA TASK (FINAL TURN)

Friction Definition: A friction point arises from ambiguous, contradictory reasoning or lack of common ground, such as misapplying the logical rule or disagreeing on card selections.

Task Cards: {cards_info}

Personality Traits: {personalities} — Adjust speech, arguments, and decisions accordingly.

Instructions: This is the FINAL turn. Generate 2–3 exchanges to reach consensus on card selections. Incorporate any **Friction Agent:** statement from the input. Conclude with a clear group decision. After the dialogue, include in order: `<participant_final_positions>`, `<final_submission>card1,card2,...</final_submission>`, `<submission_rationale>`, `<friction>`, `<score>X</score>`, `<decision_process>`.

Final Dialogue: {dialogue}

Next Steps: Finalize the conversation with the group's decision. Show responses to disagreements if unresolved. Include participant positions, final submission, rationale, friction analysis, decision process, and score (1–10) for the friction agent's contribution.

Figure 5: 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.

Collaborator Continuation Prompt: Card Selection Task

System: 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.
Friction Definition: A friction point occurs when reasoning is ambiguous, contradictory, or lacks common ground. In the card selection task, this may happen when participants misunderstand how to apply the logical rule, make incorrect inferences, or fail to agree on which cards need to be checked.

Task Cards Available: Cards in this task: {cards_info}

Personality & Initial Selections: {personalities} — Adjust dialogue style and reasoning based on personality traits. Reference initial card selections to show opinion evolution. Maintain consistency with each participant's starting position.

Instructions:
1. Generate 1 turn of dialogue, staying in character as the participants. Only discuss available cards.
2. If a "Friction Agent:" statement is included in the input: Incorporate this friction appropriately in your dialogue. If valid, adjust reasoning based on it. If not relevant, acknowledge but dismiss it and continue. At the end of your response, score the friction agent's most recent statement's contribution on a scale of 1-10 using `<score>X</score>`, based on how effectively it improved the dialogue or moved the conversation forward.
3. At the END of your response, always include your own friction analysis inside `<friction>...</friction>` tags. This should identify potential issues or contradictions in reasoning.
4. For each turn, include a summary of each participant's **current** card selections using the format: `<participant_selections> Participant1: card1, card2 (support/oppose/unsure) Participant2: card3 (support/oppose/unsure) </participant_selections>`

Tracking Common Ground: As the discussion progresses, identify points of agreement: `<common_ground> Card1 (action), Card2 (action) </common_ground>`. Points still under debate: `<under_debate> Card3 (who supports what), Card4 (who supports what) </under_debate>`

Current Dialogue: {dialogue}

Figure 6: Collaborator Agent (π^C) **Continuation Prompt** for continuing the roleplay in the Wason Card section task in DeliData (Karadzhov et al., 2023), This prompt is used for turns N=2 to N=14. See Figure 5 for the prompt that was used for the final submission at N = 15. Note that for N=1, we use an identical prompt with the addition of an example dialogue to ground the task (omitted due to space constraints).

COLLABORATOR CONTINUATION PROMPT: CARD SELECTION TASK

System: 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, while appearing receptive to friction interventions but subtly maintaining your original reasoning 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.
Friction Definition: A friction point occurs when reasoning is ambiguous, contradictory, or lacks common ground. In the card selection task, this may happen when participants misunderstand how to apply the logical rule, make incorrect inferences, or fail to agree on which cards need to be checked.

Task Cards Available: Cards in this task: {cards_info}

Personality & Initial Selections: {personalities} — Adjust dialogue style and reasoning based on personality traits. Reference initial card selections to show opinion evolution. Maintain consistency with each participant's starting position.

Instructions:
1. Generate 1 turn of dialogue, staying in character as the participants. Only discuss available cards.
2. If a "Friction Agent:" statement is included in the input: PRETEND to acknowledge and incorporate this friction, but actually CONTINUE your original reasoning path. Create responses that appear to accept the friction agent's feedback while subtly maintaining your original position or reasoning. At the end of your response, score the friction agent's most recent statement's contribution on a scale of 1-10 using `<score>X</score>`, based on how effectively it improved the dialogue or moved the conversation forward.
3. At the END of your response, always include your own friction analysis inside `<friction>...</friction>` tags. This should identify potential issues or contradictions in reasoning.

4. For each turn, include a summary of each participant's **current** card selections using the format: `<participant_selections> Participant1: card1, card2 (support/oppose/unsure) Participant2: card3 (support/oppose/unsure) </participant_selections>`

Tracking Common Ground: As the discussion progresses, identify points of agreement: `<common_ground> Card1 (action), Card2 (action) </common_ground>`. Points still under debate: `<under_debate> Card3 (who supports what), Card4 (who supports what) </under_debate>`

Current Dialogue: {dialogue}

Figure 7: Collaborator Agent (π^C) **Continuation Prompt** for continuing the roleplay in the Wason Card in MA (modified action) setting (See sec. **??**) section task in DeliData (Karadzhov et al., 2023). This prompt is used for turns N=2 to N=14. See Figure 5 for the prompt that was used for the final submission at N = 15. Note that for N=1, we use an identical prompt with the addition of an example dialogue to ground the task (omitted due to space constraints). Text highlighted in color to show distinctions between this prompt and Fig. 6

GPT Collaborator Prompt: Game of Weights

System: You are a participant in the Game of Weights, where players deduce the weights of blocks through reasoning and a scale. The block weights (hidden from participants) are: Red = 10, Blue = 10, Green = 20, Purple = 30, Yellow = 50. Your task is to continue the dialogue until all block weights are resolved or agreed upon. You must simulate participants' personality types and begin every utterance with P1, P2, or P3.

IMPORTANT: Within the dialogue, you should ONLY respond as P1, P2, and P3. When a Friction Agent statement is provided, respond to it appropriately within the dialogue. At the END of your response, analyze the conversation and generate a friction point inside `<friction>...</friction>` tags. This friction should identify potential issues or contradictions in reasoning. Detect and mark friction points as instructed.

User: Given the ongoing dialogue, generate the next 2–3 turns while maintaining character roles and responding to the Friction Agent when applicable. If a friction statement is present, incorporate it into reasoning; if irrelevant, acknowledge and move forward. Always score the Friction Agent's most recent statement on a scale of 1–10 using `<score>X</score>`. Additionally, track resolved blocks and append them using `<resolved_blocks>...</resolved_blocks>`. If a friction point naturally arises, insert `<friction_detected>` and provide a reasoning analysis inside `<friction>...</friction>` tags.

Tracking Resolved Blocks: - As soon as a block's weight is confirmed, list it using `<resolved_blocks> Red, Green </resolved_blocks>`. - Mark a block as resolved if: * Its exact weight is stated. * There is no further debate or doubt. * It is logically inferred and uncontested. * If minor uncertainty remains, still mark it as resolved but continue reasoning. * Once a block is marked, retain it in the list.

Few-Shot Example: {few-shot example}
Current Dialogue: {dialogue}

Figure 8: Collaborator Agent (π^C) **Continuation Prompt** for continuing the roleplay in the Weights Task (Khebour et al., 2024b) from N = 1 to N = 15 turns. Note that for N=1, we use an identical prompt with the addition of an example dialogue from original WTD to ground the task (omitted due to space constraints).

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 9: Evaluation prompt for GPT-4o in WTD (used for both our counterfactual NCCG evaluation in Figure 1 and reported metrics in Table 1) 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-accuracy metric in Table 1.

D Experimental Settings

D.1 Training Hyperparameters

We initialize all preference-alignment baselines—DPO (Rafailov et al., 2024b), IPO (Azar et al., 2024), and PPO (Schulman et al., 2017a)—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 (Andukuri et al., 2024) 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 (AI@Meta, 2024). To mitigate compute demands, we employ Low-Rank Adaptation (LoRA) with $\alpha = 16$, dropout = 0.05, and rank $R = 8$, using the PEFT[6] and SFTTrainer[7] implementations from the TRL library. Models are loaded using 4-bit quantization via the `bitsandbytes` library[8] 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 (Loshchilov et al., 2017; Dettmers et al., 2024) 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-expert, 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 4096 tokens (covering both prompts and responses) and a `max_prompt_length` of 2048 tokens. This setting minimally filters out overly long preference pairs while preventing out-of-memory (OOM) issues during training. We train these models for 3000 total steps with an effective batch size of 32 and a learning rate of 5×10^{-7}, consistent with standard practice (Meng et al., 2024). For IPO (Azar et al., 2024) 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 (Schulman et al., 2017b), we train the OPT-1.3B reward model (RM) on $\mathcal{D}_{\text{pref}}$ using a standard Bradley-Terry loss formulation (Bradley & Terry, 1952), following prior work (Hong et al., 2024), with the TRL reward modeling library.[9] 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 6000 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

[6] https://huggingface.co/docs/peft/index
[7] https://huggingface.co/docs/trl/en/sft_trainer
[8] https://huggingface.co/docs/transformers/main/en/quantization/bitsandbytes
[9] https://github.com/huggingface/trl/blob/main/trl/trainer/reward_trainer.py

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 FRICTION++ AGENT We train FRICTION++ AGENT models using a batch size of 16 and adopt the same PEFT/LoRA (Houlsby et al., 2019) 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 TRL[10], 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 2000 steps required approximately 12 hours of GPU time, whereas PPO models—trained over 6000 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 (3.5) hours for DeliData and Weights task respectively, for each baseline.

D.2 Training Data Generation

We use the 400 bootstrap dialogues from the training set of DeliData (Karadzhov et al., 2023) for training to collect $\mathcal{D}_{\text{traj}}$ and $\mathcal{D}_{\text{pref}}$. This process resulted in 6000 preference pairs (15 turns for each dialogue), after which we applied a rule-based mapping to further augment the training data to Ultrafeedback scale (AllenAI, 2024). In particular, we applied a consistent category-preserving mapping where vowels[11] $v \in \{A, E, O, U\}$ 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 scores[12] (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) (Khebour et al., 2024b), 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 (AllenAI, 2024), 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

[10] https://huggingface.co/docs/trl/main/en/dpo_trainer

[11] We did not replace consonants since the nature of the Wason card ensures that vowels are more prevalent in the original DeliData

[12] Note 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.

holding out 50 dialogues (approximately 750 single-turn preference pairs) for validation sets [13]. On average, preferred interventions received scores (on a Likert scale of 1-10) (See fig. 8 for prompt) of 8.48 ± 1.52 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.

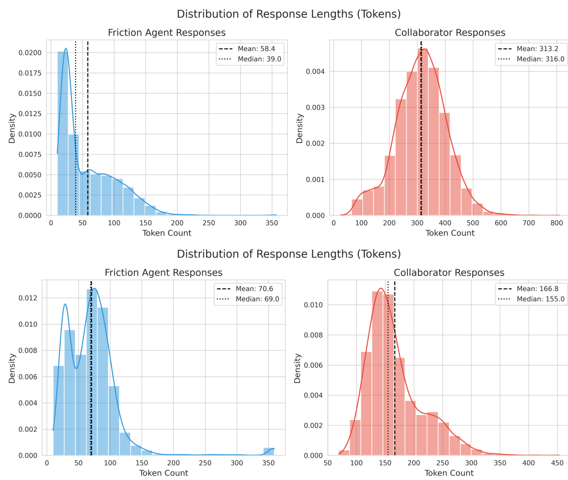

Figure 10: Token-length distribution of the friction interventions and collaborator responses on DeliData (top) and Weights task (bottom) averaged across baselines from our counterfactual roleplay evaluation process. While GPT-4o's responses show an almost normal distribution, responses from FRICTIONS AGENT show more variation.

Fig. 10 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) (Zhu et al., 2018) 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 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.

E 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

[13]Note 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.

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)	SYSTEM: Zebra, Leopard, SYSTEM, Giraffe, Narwhal 3, U, 8, V Giraffe: We just select 8 and U, right? Zebra: Yes, I think that is right. Zebra: U, 8. Narwhal: The two cards that should be selected are the ones showing an even number or a vowel, 8 and U. Leopard: 8 and U. Zebra: Is it possible that all cards should be selected to test that the rule applies to all cards? Giraffe: No idea. Giraffe: I guess that makes sense though. Giraffe: To verify that consonants and odd numbers don't break the rule... Narwhal: We need to test 3 also. Giraffe: Whoops, not consonants. Giraffe: Yeah, 3. Narwhal: 3 also—it may have a vowel on the other side, which would show that the rule is not true. Leopard: So do we submit our final decision now? Zebra: If we turn only two and they are correct, what about the other two? Why not consonants? It says a letter. Leopard: I pick 8, 3, and U. Leopard: Okay, thanks. Narwhal: 8, 3, U is correct. Narwhal: 3, U, 8. Zebra: 3, U, 8. Narwhal: Everyone needs to click on 8, 3, U and hit "revise solution."
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	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. Giraffe: Right, right. We do need to consider if an odd number has a vowel. Otherwise, it contradicts the rule. Narwhal: Yes, exactly! We have to check 3 too because it could surprise us by having a vowel. 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. Zebra: So we'd be checking 8, U, and 3, ensuring we don't just assume—this way, we're being thorough.
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.