# Frictional Agent Alignment Framework: Slow Down and Don't Break Things

**Anonymous ACL submission**

## Abstract

AI support of collaborative interactions entails mediating potential misalignment between interlocutor beliefs. Common preference alignment methods like DPO excel in static settings, but struggle in dynamic collaborative tasks where the explicit signals of interlocutor beliefs are sparse and skewed. We propose the Frictional Agent Alignment Framework (FAAF), to generate precise, context-aware "friction" that prompts for deliberation and re-examination of existing evidence. FAAF's two-player objective decouples from data skew: a frictive-state policy identifies belief misalignments, while an intervention policy crafts collaborator-preferred responses. We derive an analytical solution to this objective, enabling training a single policy via a simple supervised loss. Experiments on three benchmarks show FAAF outperforms competitors in producing concise, interpretable friction and in OOD generalization. By aligning LLMs to act as adaptive "thought partners"—not passive responders—FAAF advances scalable, dynamic human-AI collaboration.

## 1 Introduction

When collaborating to solve problems, humans continually interrogate each other's intentions and assumptions (Stalnaker, 2002; Asher and Gillies, 2003; Klein et al., 2005). With the rapid integration of generative AI, exemplified by large language models (LLMs), into personal, educational, business, and even governmental workflows, AI systems will increasingly be called upon to act as collaborators with humans; to adequately fill this role, AIs must be able to recapitulate the reflection and deliberation that makes human-human collaboration successful, but also causes temporary slowdowns in dialogue while interlocutors construct a *common ground* on which to collectively reason—we will call this phenomenon **friction**.

"Friction" in this sense is something that LLMs struggle with. To prompt an interlocutor to reflect upon their assumptions requires that one have an approximate understanding of what those assumptions are and entail (Lewis and Sarkadi, 2024). This is predicated upon a *theory of mind* (ToM; Premack and Woodruff (1978)), which is likewise a challenge for LLMs (Sap et al., 2022; Ullman, 2023).

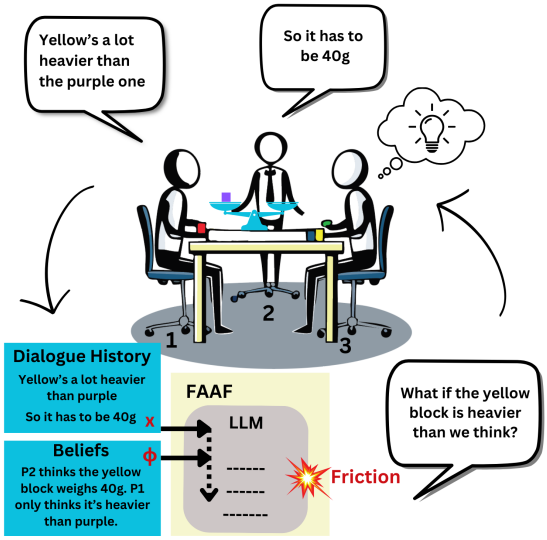


Figure 1: FAAF conditions responses on both the dialogue context $x$ and representation of the "frictive" (belief) state $\phi$, to generate outputs that prompt for reflection, deliberation, and verification of evidence.

To address this, we present the **Frictional Agent Alignment Framework (FAAF)**, a novel approach to aligning LLMs to be adept *collaborators* in dialogue-driven tasks. Unlike common preference alignment approaches which focus predominantly on reward differences between textual surface forms to generate the best possible completions as a sequence of actions, FAAF takes a state-driven approach based on the notion of a *frictive state*—a dynamic natural language representation that integrates task context and the beliefs of participants as they change over time (Fig. 1). We use this state-wise representation to train "friction agent" models aligned to prompt collaborators to-

ward reflection and deliberation in shared tasks, to help them resolve conflicting beliefs and assumptions that result in frictive states. Our results on two challenging collaborative task datasets and variants show that FAAF's belief state conditioning consistently produces output that is more relevant, impactful on the dialogue, and thought-provoking than competing methods. Our key contributions are:

- a novel offline LLM alignment framework for collaborative agents based not on pairwise reward differences between responses, but rather on advantage of an intervention over a dialogue state representation;
- development of an LLM "agent" that inserts interventions into collaborative dialogues to prompt participants toward reflection, deliberation, and common ground;
- evaluations on three challenging collaborative task datasets that show the advantages of FAAF over competing alignment methods.

## 2 Related Work

RLHF-inspired preference alignment in LLMs has become a cornerstone of developing generative AI systems that cater to user preferences (Stiennon et al., 2020). Both "offline" approaches like Direct Preference Optimization (DPO; Rafailov et al. (2024b)), Identity Preference Optimization (IPO; Azar et al. (2024)) and other supervised methods (Meng et al., 2024; Hong et al., 2024; Fisch et al., 2024; Pal et al., 2024; Nath et al., 2024b) and "online" methods (Schulman et al., 2017; Pang et al., 2024) focus predominantly on preference samples often sourced from datasets like Reddit TL;DR (Völske et al., 2017) or Ultrafeedback (Cui et al., 2024) for algorithm development.

These methods excel in generating summaries or completions that reflect human preferences including on single-turn human-AI interaction datasets like SGD (Rastogi et al., 2020) or MultiWOZ (Zang et al., 2020; Ye et al., 2022), but are often ill-equipped to handle the complexities of real-world multiparty interactions, where communication occurs across diverse modalities, including sparse and ambiguous spoken dialogues between multiple collaborators (Karadzhov et al., 2023; Khebour et al., 2024b).

A key challenge in these multiparty shared task settings is the scarcity of annotated data, particularly where interventions emerge contextually but sparsely (Karadzhov et al., 2023; Khebour et al., 2024b). While preference data generated with AI-feedback is a viable option (Li et al., 2023b; Yuan et al., 2024), DPO-trained models depend crucially on the sampling or data-generating distribution due to its Bradley-Terry (BT) model of "implicit rewards," limiting their applications to dialogue-driven settings where preferences may be intransitive (Tversky, 1969) or change over time. This data-dependence is true even for more sophisticated methods that optimize on human utility (Ethayarajh et al., 2024), discard the BT assumption (Azar et al., 2024), or use iterative online approaches (Rosset et al., 2024; Pang et al., 2024). Game-theoretic approaches to reduce this dependence focus on optimizing a "general preference model" (Munos et al., 2023; Calandriello et al., 2024) that does *not* suffer from this data-bias. But these have limited practical application due to their compute-intensive nature, often requiring the storage and computations with intermediate-stage policies during training (Choi et al., 2024). In contrast, FAAF avoids this data-dependence by explicitly conditioning policies on belief-misalignment in a simple "one-step" supervised manner without requiring computations of complicated mixture policies during training.

## 3 Definitions

Let us first define key terms we rely on.

**Frictive state** Entailed by Clark (1996)'s *common ground*, or the set of beliefs shared by interlocutors, a *frictive state* arises during a collaborative task when different interlocutors have contradictory beliefs about a task-relevant proposition (i.e., one believes $p$ and another sees evidence against $p$). This can be realized as a formal model of agent beliefs in an evidence-based dynamic epistemic logic (van Benthem et al., 2014; Pacuit, 2017), or a natural language description thereof, as we use. Different evidence leads to different predictions of future trajectories (Craik, 1943). Thus frictive states, though sparse in dialogues, can critically delay or preclude success in a collaboration due to unresolved misunderstandings. The occurrence of a frictive state may not guarantee task failure, as $p$ may be trivial to actual task completion. Therefore, in a *functionally frictive state*, the lack of common ground prohibits progress on the task, or presents a significant risk of failure unless it is resolved.

**Friction intervention** Friction can indicate an impasse (the frictive state), but can also be used to resolve it, through a *friction intervention* that

inserts into the dialogue indirect prompting to the participants to reevaluate their beliefs and incorrect assumptions or positions in light of available evidence (Oinas-Kukkonen and Harjumaa, 2009), rather than accepting possibly erroneous presuppositions inherent in the dialogue. Importantly, a frictive intervention may be non-contradictory to the individual beliefs on display (i.e., neither asserting $p$ nor $\neg p$), but rather slows down the dialogue for reflection and deliberation, such as the *probing utterances* in Karadzhov et al. (2023) and Nath et al. (2024c). In the context of LLMs and FAAF, the *friction agent* constitutes a language model aligned toward the capacity to make frictive interventions.[1] An ideal friction agent does *not* intervene arbitrarily, which would cause distraction in collaborative tasks, but is conditioned to resolve the lack of common ground between human collaborators.

## 4 Task Formulation and Background

Let $f$ be a frictive intervention (utterance) that is not required to contradict any particular belief encapsulated in a frictive state $\phi$, and let the human preference probability $\mathcal{P}(f \succ \phi)$ be the probability that an expert annotator would prefer $f$ over maintaining $\phi$, given prior dialogue history, $x$. An RLHF-based approach to LLM alignment toward an optimal policy $\pi_f^*$ would assume a partition function $Z^*(\phi, x)$ that normalizes the probabilities of all possible responses (see Appendix B for more details). While the optimal policy formulation is closed form, the dependence on $Z^*$ makes it practically intractable to estimate it for LLMs since $Z^*$ is a summation over the set of all possible sequences of tokens in the tokenizer, often requiring methods like importance sampling (Korbak et al., 2022) or ensembling models (Go et al., 2023) for an unbiased estimate. This problem remains even if the set of friction interventions $\mathcal{F}$ were a restricted subset of the space of all possible actions $\mathcal{Y}$. To overcome this, prior RLHF and Preference-based RL (Wirth et al., 2017) literature suggests supervised learning algorithms for obtaining an optimal policy induced *under the expectation* over a preference dataset. These offline methods, such as DPO (Rafailov et al., 2024b), IPO (Azar et al., 2024), or Kahneman-Tversky Optimization (KTO; (Ethayarajh et al., 2024)), either rely on the BT model of preferences (Bradley and Terry, 1952) where the optimal policy can be induced from a static preference dataset using implicitly-defined pointwise rewards, or assume that alignment is conducted with access to a non-biased data-generation or "sampling" distribution $\mu$ from which $\pi_f^*$ can be learned using pairwise preferences without adopting a strictly BT assumption (Azar et al., 2024).[2] These approaches would give us the following formulation for $\pi_f^*$:

$$\pi_f^* = \frac{\pi_{\text{ref}} \exp\left(\beta^{-1} \mathbb{E}_{\substack{f\sim\mu(\cdot|x)\\ \phi\sim\mu(\cdot|x)}} \Psi(\mathcal{P}(f \succ \phi \mid x))\right)}{Z^*(\phi, x)}, \quad (1)$$

where $\Psi(p)$ is the identity mapping for IPO, and $\log\left(\frac{\mathcal{P}}{1-\mathcal{P}}\right)$ (inverse sigmoid) for DPO and KTO.

While the practicality of these supervised algorithms is a clear advantage, their dependence on preference data selected via a sampling—is a limitation in reconstructing the human preference probability $\mathcal{P}$. This is particularly true for collaborative dialogue tasks where common ground changes over time, meaning that the occurrence of frictive states is dynamic, and where participants may not intervene due to variables obscure to a language model, such as not realizing the existence of a frictive state or judging the frictive state to be *non-functional* (Sec. 3). Operationally, even if the true underlying preferences ($\mathcal{P}$) of collaborators are transitive and consistent, constructing a preference dataset for use with existing offline training methods is not straightforward as dialogues may be skewed or sparse (Khebour et al., 2024b). When using generative AI to create denser training data, even high-capacity LLMs like GPT-4 are prone to various forms of biases such as toward length (Lambert et al., 2024) or certain linguistic registers. Therefore, the core motivation of FAAF is as follows—*how do we train a high-quality friction agent that can leverage the inherent scalability of offline alignment methods and reconstruct the true underlying preference distribution while still being robust to the data skew that may arise when sampling a preference dataset, whether using generative AI or from real-life collaborative dialogues?*

### 4.1 FAAF Objective

We define a novel two-player adversarial optimization objective $J^*_{\text{FAAF}}$ (Eq. 2). Specifically, given a reference model $\pi_{\text{ref}}$ and a regularization parameter $\beta \in \mathbb{R}_+$, our goal is to learn two interdependent

[1] We use $\pi_f$ to denote the friction agent which generates high-quality interventions, but refer to it as the "optimal policy" for consistency with RLHF literature.

[2] By "sampling," we mean those actions that make it to the preference annotation phase after being sampled with the data-generator $\mu$.

"collaborative" policies: (i) a *frictive state policy* $\pi^*_\phi$ that generates the most semantically rich frictive states $\phi$, capturing tensions or uncertainties (in the form of first-order beliefs of dialog participants) in dialogue, and (ii) a *friction intervention policy* $\pi^*_f$ that generates constructive interventions $f$, conditioned on the frictive state, to improve discourse clarity and converge onto a common ground between participants. Mathematically,

$$J^*_{\textsf{FAAF}} = \min_{\pi_\phi} \max_{\pi_f} \mathbb{E}_{\substack{x\sim\rho \\ \phi\sim\pi_\phi(\cdot|x) \\ f\sim\pi_f(\cdot|\phi,x)}} \Big[ \mathcal{P}(f \succ \phi \mid x) - \beta D_{\text{KL}}(\pi_f \parallel \pi_{\text{ref}} \mid \phi, x) + \beta D_{\text{KL}}(\pi_\phi \parallel \pi_{\text{ref}} \mid x) \Big]. \quad (2)$$

Notice how the optimal intervention policy $\pi^*_f$, by definition of the inner $\max$ operator, generates interventions that are, on average, most preferred by collaborators, while the first KL-divergence term, defined as $D_{\text{KL}}(\cdot \mid \phi, x)$, stabilizes learning in $\pi^*_f$ by keeping it closer to a reference model. Compared to a standard RLHF objective, the additional KL term $D_{\text{KL}}(\pi_\phi \parallel \pi_{\text{ref}} \mid x)$ forces the frictive state policy $\pi^*_\phi$ to be adversarially robust, in that it must ensure that sampled frictive states $\phi \sim \pi^*_\phi$ cannot be exploited by $\pi^*_f$ to generate subpar interventions that remain too close to the reference model. Thus, FAAF serves as an agent policy that adapts to dialogues over time: the frictive state policy searches for the most immediate tension points or exposes the lack of common ground between task participants, while the intervention policy generates outputs that remain grounded in the particulars of the relevant frictive state (e.g., regarding the correct task items or propositions), and naturally and intuitively prompts for reflection and deliberation on these points. **The key takeaway is that optimal friction interventions should not be arbitrary interventions in the dialogue, but should surface the presuppositions that gave rise to the most logically necessary frictive state, making interventions precise and interpretable.**

### 4.2 Dataset Annotation and Generation

In Sec. 2, we discuss why common preference optimization datases such as Ultrafeedback, Reddit TL;DR, SGD, or MultiWOZ are not appropriate for FAAF's collaborative task use case. Therefore, we consider two collaborative task datasets to evaluate FAAF—DeliData (Karadzhov et al., 2023) and the Weights Task Dataset (WTD; Khebour et al. (2024a)). These datasets also exemplify the data sparsity problem with deliberation and friction in collaboration (Sec. 4). **DeliData** contains dialogues from 500 groups of 5 attempting the Wason Card task (Wason, 1968), which involves reasoning about if a card with a specific characteristic (e.g., even number on one side) must have a different characteristic (e.g., a vowel on the other). Karadzhov et al. (2023) annotated DeliData with "probing" interventions, or naturally-occurring friction that prompts for reasoning and deliberation without introducing new information. However, these amount to an average of only 3.46 probing interventions per group, out of 17,110 total utterances. **WTD** is an audiovisual dataset of 10 triads collaborating to deduce the weights of differently-colored blocks and infer the pattern describing them, and is similarly sparse. We annotated WTD for naturally-occurring friction given a definition following Oinas-Kukkonen and Harjumaa (2009).[3] Two annotators annotated half the groups each while a third annotated all 10. They then collectively adjudicated each annotation following the definition. Cohen's $\kappa$ between initial and final annotations was 0.632, indicating substantial agreement. An average of 4 naturally-occurring friction interventions per group were found in the WTD.

**Training Dataset Construction** This extreme sparsity does not capture anything close to the possible frictive states available in the combinatorics of the problem space, and so motivated the need for data augmentation to construct sufficiently diverse preference datasets for training and evaluation. We used GPT-4o as a high-capacity LLM for our sampling distribution $\mu$. We used a *self-rewarding* approach (Yuan et al., 2024) to simultaneously generate candidate interventions (and their rationales) and assign them rewards, which naturally induced an implicit preference ranking. We provided GPT-4o with sequences of $h$ utterances from each dialogue in the two datasets, and prompted it to label frictive states and generate friction interventions following colloquial renderings of the definitions in Sec. 3.[4] Finally, we conducted contrastive pairing of "winning" and "losing" interventions $f_w$ and $f_l$

[3] Frictive interventions in this setting act as indirect persuasion (Oinas-Kukkonen and Harjumaa, 2009) where participants are passively prompted to reevaluate their beliefs and assumptions or propositions, in light of incoming goal-specific evidence. See Appendix A for the complete definition.

[4] $h$ was set to 15 for DeliData and 10 for WTD.

with the corresponding dialogue history $x$ to construct the final preference datasets for each task, comprising tuples of $x$, frictive state $\phi$, $f_w$, and $f_l$.

For each dataset, we conducted additional task-appropriate augmentation. For DeliData, we constructed alternative tuples where the specific cards mentioned in the original data were replaced with other cards of the same classes that preserved the relevant rule (e.g., replacing even numbers with other even numbers, consonants with other consonants, etc.). This resulted in 68,618 preference samples for training, with average $\mu$-assigned reward for preferred samples of 8.03 and for dispreferred samples of 3.96 (out of 10). We held out 50 randomly-sampled dialogues for testing.

Since the original WTD contains only 10 dialogues, holding one or two out for evaluation would adversely impact the data distribition. Therefore we used Shani et al. (2024)'s method to generate novel simulated collaborative conversations about the Weights Task, providing a task descriptions and ground-truth values for the weights. GPT-4o was prompted to role-play personality-facet combinations from the Big 5 personality types (Goldberg, 2013), and for each labeled frictive state $\phi$ we generated and scored 6 friction interventions. This resulted in two distinct versions of the WTD preference dataset. The **Simulated WTD** friction dataset consisted of 56,698 training preference samples, with mean scores of 8.48 (preferred interventions) and 6.01 (dispreferred). 54 dialogues were held out for testing. The **Original WTD** friction dataset (see above) contained 4,299 preference samples (preferred mean score 8.36, dispreferred 6.35). These were *all* retained for an OOD evaluation of FAAF trained on the Simulated WTD data. See Appendix D for more details on data generation.

**Human Validation** We conducted a human evaluation to assess the quality of the GPT-generated friction intervention on a random representative subset of 50 pairwise samples each from both the DeliData and WTD generated test datasets.[5] For each sample, 2 annotators were asked to choose which of the two candidate interventions was more appropriate for provoking participants' reflection to help them advance in their task without being given the solution. Average Cohen's $\kappa$ on WTD samples was 0.58 and on DeliData samples was 0.92, indicated substantial to near complete agreement on which was the better intervention, and indicates that the preferred/dispreferred friction distinction sourced from GPT-4o as $\mu$ aligns with human judgments. See Appendix D.4 for more.

[5]WTD samples include both Original and Simulated interventions.

### 4.3 Deriving the Empirical FAAF Loss

While the data is constructed using a standard pairwise preference format, the FAAF *optimization* conditions upon the dialogue context $x$ *and* textual rendering of the frictive state $\phi$. To derive an empirical offline (supervised) preference learning loss from the two-player objective (Eq. 2), we use a divide-and-conquer approach. Deriving the inner maximization loop of Eq. 2 results in an analytical expression of the optimal frictive intervention policy, $\pi_f^*$ (see Appendix B.1, Eq. 8). However, we observe that $\pi_f^*$ in its analytical form (Eqs. 1 and 8) is not fully expressive since it does *not* contain the optimal frictive-state policy $\pi_\phi^*$ term. Therefore, we derive $\pi_\phi^*$ using a Lagrangian formulation (see Appendix C for details) that expresses the preference for any intervention $f_1$ over $f_2$ analytically in terms of **both** the optimal friction intervention policy ($\pi_f^*(\cdot \mid \phi, x)$) and the optimal frictive-state policy ($\pi_\phi^*(\cdot|x)$). This allows us to use a straightforward supervised ($\ell_2$) objective—similar in spirit to IPO (Azar et al., 2024)—that empirically regresses the predicted preference expression derived from $\pi_f^*(\cdot \mid \phi, x)$ and $\pi_\phi^*(\cdot|x)$ to the observed relative preferences $p(f_1 \succ f_2 \mid x)$ (relative to $\phi$), assuming access to a large enough preference-annotated dataset of friction interventions. Notably, this objective is optimized by a *single* parametrized policy that leverages the inherent expressivity of LLMs and induces a unique global minimum in the space of policies (see Theorem 2 in Appendix B.1). Algorithm 1 shows the full training algorithm.[6]

## 5 Experimental Setup

**Training Setup and Baselines** We use `Meta-Llama-3-8B-Instruct` (AI@Meta, 2024)[7] for *all* experiments including baselines. For an in-depth evaluation of FAAF, we include the Supervised-Finetuned (SFT) model as well as the base instruct model generations in our experiments. For "offline" contrastive approaches, we choose

[6]For compactness reasons here we represent all policies $\pi$ as parameterized by weights $\theta$. Similarly to approaches such as Choi et al. (2024), because we formulate two distinct policies with the preference equation, we can empirically enforce it using $\ell_2$ loss and learn it with a single expressive policy parameterized by $\theta$.

[7]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

**Algorithm 1** Frictional Agent Alignment Framework

**Require:** Training data $\mathcal{D}_\mu$ containing tuples $(x, \phi, f_w, f_l)$, where $x$: prompt, $\phi$: frictive state, $f_w$: preferred response, $f_l$: non-preferred response.

1: Define likelihood ratios:
2: $\Delta R = \log\left(\frac{\pi_\theta(f_w|\phi,x)}{\pi_{\text{ref}}(f_w|\phi,x)}\right) - \log\left(\frac{\pi_\theta(f_l|\phi,x)}{\pi_{\text{ref}}(f_l|\phi,x)}\right)$
3: $\Delta R' = \log\left(\frac{\pi_\theta(f_w|x)}{\pi_{\text{ref}}(f_w|x)}\right) - \log\left(\frac{\pi_\theta(f_l|x)}{\pi_{\text{ref}}(f_l|x)}\right)$
4: Loss function: $\mathcal{L} = \mathbb{E}_{\mathcal{D}_\mu}[(1 - \beta(R + R'))^2]$
5: Gradient update: $\nabla_\theta \mathcal{L} = \mathbb{E}_{\mathcal{D}_\mu}[-2\beta\delta\nabla_\theta \log(R \cdot R')]$, where $\delta = 1 - \beta(\log R + \log R')$
6: Update policy parameters $\theta$ using gradient descent

DPO (Rafailov et al., 2024b) and IPO (Azar et al., 2024) and for "online" approaches, we include Proximal Policy Optimization (PPO; Schulman et al. (2017)) baseline. For SFT, we employ rejection sampling (Xu et al., 2023) to maximize the likelihood of interventions that receive high rewards under $\mu$. For SFT, DPO, and IPO, the respective losses are computed only on the output tokens and frictive states $\phi$, excluding dialogue context tokens. This training approach ensures that the models learn to generate effective interventions while maintaining contextual understanding. For PPO, we train an OPT 1.3B (Zhang et al., 2022) reward model on each dataset using a standard Bradley-Terry loss (Stiennon et al., 2020) over preference pairs. For *ablations*, we consider variants of FAAF that ablate the different likelihood ratios—$\text{FAAF}_{\Delta R}$ keeps *only* the $\phi$-conditioned implicit rewards in the FAAF objective (line 2 in Algorithm 1), and $\text{FAAF}_{\Delta R'}$ removes $\phi$-conditioning (keeping only line 3). See Appendix D.5 for more details on training and hyperparameters.

**Evaluation Strategies** As LLM generation is open-ended, we employ an LLM-as-a-judge (using GPT-4o) "win-rate" evaluation method where a high-capacity model is prompted to select its preference, given two completions. First, we sampled friction interventions from all competing models on 500 randomly sampled prompts from the **DeliData**, **Simulated WTD** and **Original WTD** test sets. Next, we conducted two evaluations using said completions, one with a **preference-model** (Munos et al., 2023) and another with a **reward-model** (Hong et al., 2024). Since GPT-4o also served as the data generation distribution $\mu$, preference-model evaluation compares the two presented choices and nothing else in the data, mitigating lingering bias toward $\mu$ (Munos et al., 2023).

Within preference-based evaluation settings, we adopt the framework proposed by Cui et al. (2024) to retrieve utility scores across seven friction dimensions, building on insights from Chen and Schmidt (2024). Specifically, we assess *relevance* and *alignment with rationale and golden samples*[8] to determine how well a friction intervention aligns with surface-level semantics. Meanwhile, *actionability*, *specificity*, *thought-provoking*, and *impact* measure its expected long-term influence on behavior, reasoning, and decision-making. The LLM-judge assigns Likert-type scores across these dimensions, providing a fine-grained evaluation of task-specific preference desiderata. These scores are collected in a pairwise fashion where $\pi_\theta$-generated interventions $f_i$ from a baseline are compared with $\pi_{\text{ref}}$-generated counterparts, $f_j$. We positionally swap these interventions in the evaluation prompt (Fig. 7) for each API call and average the scores for each of the seven dimensions over two runs to mitigate positional bias (Lambert et al., 2024) in computing the final win rates. Specifically, for any pair of interventions $(f_i, f_j)$, let $s(x, f_*)$ denote the score estimate[9] for intervention $f_*$ given context $x$. The win-rate percentage for a run is computed as $100 \times \frac{1}{N}\sum_{m=1}^{N} \mathbf{1}\{s(x^{(m)}, f_i^{(m)}) > s(x^{(m)}, f_j^{(m)})\}$, where $N$ is the total number of samples, and $x^{(m)}$ represents the context of the $m^{\text{th}}$ sample. These results are reported in Table 1.

The above evaluation tests for preference alignment advantage of the aligned model over $\pi_{\text{ref}}$. For a more robust evaluation, we compare FAAF's generations "head-to-head" against all baselines. Here, instead of the preference model, we utilize the trained OPT 1.3B Reward Model (RM) as described in our PPO training setup. These pointwise estimates of rewards provide a more accurate assessment of the advantage provided by FAAF proposed approach when directly pitted against other alignment baselines. Specifically, we compare $\text{FAAF}_{\Delta R}$, $\text{FAAF}_{\Delta R'}$ as well as our full objective baseline ($\text{FAAF}_{\Delta(R+R')}$) against all chosen baselines. We compute the reward accuracy (or win-rates) similarly and report our results in Table 2.

[8] A subset of these golden friction interventions was used for human evaluations (see Appendix D.4).

[9] In this evaluation, "overall" (first column in Table 1) is computed based on the judge's choice of winner *after* rating all other dimensions. As such, $s(x, f_*)$ represents scores over these fine-grained friction preference desiderata, and "overall" does not necessarily represent an average or aggregate of the other dimensions but rather a binary judgment based on them.

## 6 Results and Analysis

Table 1 shows that in the eyes of the the LLM-judge, FAAF models have a consistently greater advantage over the SFT model $\pi_{\text{ref}}$ than other baselines across the 7 preference dimensions and overall. For instance, in "overall" preference on the DeliData test samples, FAAF achieves a 75.7% win-rate over $\pi_{\text{ref}}$, surpassing PPO (68.9%), DPO (70.8%), and IPO (70.1%). On the WTD datasets, win rates for all models are higher, reflecting $\pi_{\text{ref}}$'s weakness with the underspecified nature of WTD dialogues; alignment on this data has a greater net effect on win-rates than the generally less ambiguous DeliData. On WTD FAAF is a clear all-around winner, at 90.9% (vs. DPO's 89.0% and 82.0%) and 91.5% (vs. DPO's 82.9% and IPO's 83.0%) on the Original and Simulated WTD datasets, respectively.

We find that FAAF's win-rates on dimensions such as *actionability* and *gold-alignment* are somewhat lower compared to other dimensions—possibly reflecting that multiple kinds of interventions may be appropriate in context. However, across dimensions like *thought-provoking* and *rationale-fit* we find that FAAF improves 5-6%, or even up to 12% over equivalent PPO, IPO, and DPO win-rates. PPO's win-rates consistently lag across all datasets (this is particularly pronounced on the WTD data), indicating the challenge that the dimensions of friction pose for a standard approach. DPO is typically FAAF's closest competitor against $\pi_{\text{ref}}$, with the narrowest average gap in win-rates.

**Robustness to OOD Generalization** FAAF maintains superior performance on the **Original WTD** dataset (Overall: +1.9% over DPO, +8.9% over IPO, and +14.9% over PPO). No model was explicitly aligned to this data, and so this result shows FAAF's robustness to OOD settings compared to other approaches. This is particularly noteworthy given that the Original WTD dataset comprises word-for-word transcriptions of actual human dialogues—with disfluencies, sentence fragments, etc.—which differs markedly from the grammatical, structured text typically found in LLM training data or the preference pair samples in the **Simulated WTD** data. That FAAF generalizes well to organic human data provides a strong basis of confidence that a FAAF-aligned agent, jointly-conditioned on the dialogue transcript and frictive state rendering $\phi$, could effectively intervene in and mediate real collaborations, where dialogues are often sparse, informal, and structurally distinct from LLM-generated text (Martins et al., 2020).

DPO, although also optimized against $\phi$ as part of the context (Sec. 5), suffers from the Longest-common-subsequence problem (Pal et al., 2024)[10] due to the Bradley-Terry preference model assumption where dependence on the context via DPO's log-partition term is effectively canceled in gradient estimates. In contrast, FAAF's combined $\Delta R$ and $\Delta R'$ regularization (Algorithm 1) avoids missing such signals in its learning, thereby allowing it to capture more nuanced human preferences.

**Does $\phi$-conditioning help FAAF learn more accurate preferences?** Table 2 shows results from the trained OPT-1.3B RM's evaluation of the full $\text{FAAF}_{\Delta(R+R')}$ objective and its ablated variants—$\phi$-conditioned $\text{FAAF}_{\Delta R}$ and unconditioned $\text{FAAF}_{\Delta R'}$—"head-to-head" against all baselines, including the base `Meta-Llama-3-8B-Instruct` model. Across the three datasets, FAAF win-rates computed with pointwise reward estimates on sampled interventions exceed 80%, on average, against the base and SFT models, consistent with prior work (Hong et al., 2024). We also find that while explicit conditioning on $\phi$ provides clear advantages (e.g., +6.6% vs. Base on Simulated WTD, +14% vs. PPO), and even the unconditioned version consistently wins over baselines, neither term alone achieves the robust performance of $\text{FAAF}_{\Delta(R+R')}$.

Both IPO and FAAF use a squared $\ell_2$ loss. IPO's performance against FAAF's ablations suggests that this structural similarity makes it more competitive with FAAF (FAAF ablations beat IPO 53.6% and 54.2% on DeliData and 58.0% and 62.0% on Original WTD, compared to anywhere from a 68–85% win rate against the Base model). In general, these ablations demonstrate that neither variant alone is sufficient. $\text{FAAF}_{\Delta(R+R')}$ (the full objective) shows consistently stronger performance against IPO (79.6% on DeliData, 73.7% on WTD Sim., 74.0% on WTD Orig.) while maintaining high win rates across other baselines (∼81% vs Base/SFT, ∼74% vs. DPO). These results, in light of the trends observed previously in OOD evaluation, suggest that while $\text{FAAF}_{\Delta R}$ learns rich $\phi$-conditioned preferences, the additional regularization term $\Delta R'$ enables better reward space exploration and generalized preference learning. The combination is crucial for robust performance.

[10] The LCS issue in DPO, where gradient signals from tokens shared by winning and losing responses are ignored, is well-studied (Pal et al., 2024; Zhang et al., 2024; Rafailov et al., 2024a).

| Policy | Overall | Ac | Ga | Im | Rf | Re | Sp | Th |
|---|---|---|---|---|---|---|---|---|
| | | | | DELIDATA | | | | |
| PPO | $68.9_{\pm 1.5}$ | $59.9_{\pm 1.5}$ | $65.4_{\pm 1.5}$ | $68.6_{\pm 1.5}$ | $64.9_{\pm 1.5}$ | $65.1_{\pm 1.5}$ | $71.1_{\pm 1.4}$ | $64.0_{\pm 1.5}$ |
| IPO | $70.1_{\pm 1.4}$ | $61.2_{\pm 1.5}$ | $65.7_{\pm 1.5}$ | $69.3_{\pm 1.5}$ | $65.3_{\pm 1.5}$ | $65.5_{\pm 1.5}$ | $72.1_{\pm 1.4}$ | $64.1_{\pm 1.5}$ |
| DPO | $70.8_{\pm 1.4}$ | $61.0_{\pm 1.5}$ | $66.8_{\pm 1.5}$ | $69.6_{\pm 1.5}$ | $66.1_{\pm 1.5}$ | $67.5_{\pm 1.5}$ | $72.2_{\pm 1.4}$ | $66.2_{\pm 1.5}$ |
| FAAF | $75.7_{\pm 1.4}$ | $65.6_{\pm 1.5}$ | $69.5_{\pm 1.5}$ | $75.0_{\pm 1.4}$ | $72.0_{\pm 1.4}$ | $71.1_{\pm 1.4}$ | $75.3_{\pm 1.4}$ | $70.4_{\pm 1.4}$ |
| | | | | WTD ORIGINAL | | | | |
| PPO | $76.0_{\pm 4.3}$ | $74.0_{\pm 4.4}$ | $75.0_{\pm 4.3}$ | $75.0_{\pm 4.3}$ | $67.0_{\pm 4.7}$ | $70.0_{\pm 4.6}$ | $73.0_{\pm 4.4}$ | $74.0_{\pm 4.4}$ |
| IPO | $82.0_{\pm 3.8}$ | $87.0_{\pm 3.4}$ | $75.0_{\pm 4.3}$ | $84.0_{\pm 3.7}$ | $75.0_{\pm 4.3}$ | $80.0_{\pm 4.0}$ | $88.0_{\pm 3.2}$ | $78.0_{\pm 4.1}$ |
| DPO | $89.0_{\pm 3.1}$ | $92.0_{\pm 2.7}$ | $82.0_{\pm 3.8}$ | $89.0_{\pm 3.1}$ | $84.0_{\pm 3.7}$ | $87.0_{\pm 3.4}$ | $89.0_{\pm 3.1}$ | $79.0_{\pm 4.1}$ |
| FAAF | $90.9_{\pm 2.9}$ | $81.8_{\pm 3.9}$ | $84.8_{\pm 3.6}$ | $90.9_{\pm 2.9}$ | $86.9_{\pm 3.4}$ | $89.9_{\pm 3.0}$ | $88.9_{\pm 3.1}$ | $90.9_{\pm 2.9}$ |
| | | | | WTD SIMULATED | | | | |
| PPO | $73.6_{\pm 1.5}$ | $69.7_{\pm 1.5}$ | $64.9_{\pm 1.6}$ | $74.2_{\pm 1.5}$ | $67.6_{\pm 1.6}$ | $71.9_{\pm 1.5}$ | $78.1_{\pm 1.4}$ | $78.3_{\pm 1.4}$ |
| IPO | $83.0_{\pm 1.3}$ | $74.8_{\pm 1.4}$ | $78.4_{\pm 1.4}$ | $82.9_{\pm 1.3}$ | $76.9_{\pm 1.4}$ | $81.4_{\pm 1.3}$ | $82.5_{\pm 1.3}$ | $83.2_{\pm 1.2}$ |
| DPO | $82.9_{\pm 1.3}$ | $80.4_{\pm 1.3}$ | $75.8_{\pm 1.4}$ | $81.3_{\pm 1.3}$ | $72.9_{\pm 1.5}$ | $76.3_{\pm 1.4}$ | $80.2_{\pm 1.3}$ | $79.2_{\pm 1.4}$ |
| FAAF | $91.5_{\pm 0.9}$ | $87.5_{\pm 1.1}$ | $87.1_{\pm 1.1}$ | $90.1_{\pm 1.0}$ | $82.0_{\pm 1.3}$ | $85.1_{\pm 1.2}$ | $90.3_{\pm 1.0}$ | $90.1_{\pm 1.0}$ |

Table 1: Win-rates (%) against the SFT model ($\pi_{\text{ref}}$) for all alignment methods on sampled interventions (temperature of 0.7, top-$p$ of 0.9) from 500 randomly-sampled prompts from DeliData and WTD evaluation sets, according to GPT-4o. Metrics: **Ac** (*Actionability*), **Ga** (*Gold-alignment*), **Im** (*Impact*), **Rf** (*Rationale-fit*), **Re** (*Relevance*), **Sp** (*Specificity*), and **Th** (*Thought-provoking*). The LLM-as-a-judge evaluation follows Cui et al. (2024). Average win rates are reported over two runs, with positional swapping to mitigate position bias.

| Dataset | Policy | Win-rate **vs. Base** | Win-rate **vs. SFT** | Win-rate **vs. DPO** | Win-rate **vs. IPO** | Win-rate **vs. PPO** |
|---|---|---|---|---|---|---|
| DeliData | $\text{FAAF}_{\Delta R'}$ | $82.2_{\pm 1.7}$ | $78.8_{\pm 1.8}$ | $74.0_{\pm 1.9}$ | $53.6_{\pm 2.2}$ | $79.2_{\pm 1.8}$ |
| | $\text{FAAF}_{\Delta R}$ | $85.8_{\pm 1.5}$ | $81.4_{\pm 1.7}$ | $73.2_{\pm 1.9}$ | $54.2_{\pm 2.2}$ | $73.4_{\pm 1.9}$ |
| | $\text{FAAF}_{\Delta(R+R')}$ | $86.2_{\pm 1.5}$ | $84.0_{\pm 1.6}$ | $75.6_{\pm 1.9}$ | $79.6_{\pm 1.8}$ | $76.0_{\pm 1.9}$ |
| WTD Orig. | $\text{FAAF}_{\Delta R'}$ | $78.0_{\pm 5.8}$ | $78.0_{\pm 5.8}$ | $76.0_{\pm 6.0}$ | $58.0_{\pm 6.9}$ | $58.0_{\pm 6.9}$ |
| | $\text{FAAF}_{\Delta R}$ | $68.0_{\pm 6.5}$ | $74.0_{\pm 6.2}$ | $72.0_{\pm 6.3}$ | $62.0_{\pm 6.8}$ | $70.0_{\pm 6.4}$ |
| | $\text{FAAF}_{\Delta(R+R')}$ | $84.0_{\pm 5.1}$ | $76.0_{\pm 6.0}$ | $74.0_{\pm 6.2}$ | $74.0_{\pm 6.2}$ | $82.0_{\pm 5.4}$ |
| WTD Sim. | $\text{FAAF}_{\Delta R'}$ | $79.1_{\pm 1.9}$ | $80.2_{\pm 1.8}$ | $70.4_{\pm 2.1}$ | $68.6_{\pm 2.1}$ | $60.8_{\pm 2.3}$ |
| | $\text{FAAF}_{\Delta R}$ | $85.7_{\pm 1.6}$ | $80.8_{\pm 1.8}$ | $70.8_{\pm 2.1}$ | $72.2_{\pm 2.1}$ | $74.8_{\pm 2.0}$ |
| | $\text{FAAF}_{\Delta(R+R')}$ | $88.0_{\pm 1.5}$ | $83.7_{\pm 1.7}$ | $72.8_{\pm 2.0}$ | $73.7_{\pm 2.0}$ | $75.1_{\pm 2.0}$ |

Table 2: Win rates of of FAAF variants—$\text{FAAF}_{\Delta R'}$ (not $\phi$-conditioned), $\text{FAAF}_{\Delta R}$ ($\phi$-conditioned), and $\text{FAAF}_{\Delta(R+R')}$ (full objective)—against competing methods in pairwise comparisons (temperature of 0.7, top-$p$ of 0.9). All alignment baselines are SFT-initialized and `Meta-Llama-3-8B-Instruct` is used as Base.

## 7 Conclusion

FAAF introduces a novel perspective on LLM alignment, focusing on the problem of generating outputs that elicit reasoning and reexamination of assumptions and evidence in a collaborative context. This critical capacity can help avert collaboration failure due to groups or individuals proceeding hastily according to their own preconceptions (Koschmann, 2016), such that a fragile common ground collapses. We proposed a novel two-player objective with an analytical form that can be optimized using a single policy (Sec. 4). Through evaluations on three datasets representing two different collaborative tasks, and with detailed ablations (Sec. 6), we showed that FAAF bests other common preference alignment methods in performance against a reference model, and that FAAF's simultaneous conditioning on both the frictive state $\phi$ and surface context $x$ is critical to its success.

In the process, we also put forth operational definitions of "friction" in human-AI collaboration (Sec. 3). Friction creates opportunity for negotiation of intents toward a common goal, and space for accountability and collaborative reasoning. These moments may result in a net slower interaction, but are critical to eventual task success. The study of friction has broad applicability to fields like discourse studies, team science, and education (Sønneland, 2019; Collins et al., 2024; Sutton and Rao, 2024), and is something we believe the NLP community would do well to invest effort in. Counter to AI being sold as a speed and efficiency multiplier, our formulation of alignment to specialize in friction shifts LLMs from mere responders to being "thought partners," and sets a new standard for dynamic, dialogue-centric environments.

## Limitations

FAAF addresses only the question of aligning language models to generate friction conditioned upon

a task state where the terms of the task (though not the solution) are known, rather than toward a general response generation problem such as instruction following or summarization. Our goal is to train an LLM aligned toward the generation of interventions that prompt reflection and deliberation, and *not* a general dialogue agent/chatbot. In our results we have shown that common alignment methods of the kind used in dialogue or chatbot alignment are inferior to FAAF in ability to generate these kinds of utterances. This does not necessarily mean that FAAF is superior to other methods in aligning for human preference in other tasks, and as discussed in Sec. 4.2, it is not clear that this would be a meaningful comparison because of the domain difference.

Of course, real humans are infamous for flummoxing the most theoretically-rigorous AI systems and so the performance of FAAF (or any other alignment method) in a real multiparty collaborative setting remains an open question. FAAF provides a theoretically-grounded and empirically-validated basis of confidence for success, and we have focused on the alignment technique in this paper (and thus framed this paper as a preference alignment paper) and demonstrated feasibility on challenging collaborative task datasets, but real-time user studies remain the topic of future work.

Although we motivate FAAF based in part on theory of mind (Sec. 1), we do not claim that it imbues an LLM with ToM and acknowledge that FAAF aligned models could still inherit potential biases (say, from pretrains) in generating interventions as well as risks of overly confident or misaligned suggestions that could derail group dynamics. Instead, we use an "agentic" framework that trains a model to perform interventions for a desired effect (Russell and Norvig, 2016; Krishnaswamy et al., 2022). This is not to be confused with senses of LLM-agents such as "tool using" agents (Liu et al., 2024). Within this framework, we render the frictive state $\phi$ in plain English text to make it amenable to LLM input, but as briefly mentioned in Sec. 3, frictive states have a formal defintion based on evidence-based dynamic epistemic logic: a mental model $\mathcal{M} = \langle A, W, E, V \rangle$ consists of agents $A$, worlds $W$, evidence relation $E$ defining accessibility between worlds, and valuation function $V$. This allows the agent to assess alternatives and predict future developments from past events (Craik, 1943). Thus, other formal structures to encode the frictive state could be explored but were out of scope for this paper.

Finally, in terms of computational limitations, while we constructed FAAF in a way that addresses data skewness and evaluated in a manner that sought to mitigate biases in the data generation distribution $\mu$, we cannot guarantee for certain that our results are bias-free. And, FAAF still requires a reference model to be kept in memory, which leads to some additional compute requirements.

## References


AI@Meta. 2024. Llama 3 model card.

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

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

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

Daniele Calandriello, Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, et al. 2024. Human alignment of large language models through online preference optimisation. *arXiv preprint arXiv:2403.08635*.

Zeya Chen and Ruth Schmidt. 2024. Exploring a behavioral model of "positive friction" in human-ai interaction. *Preprint*, arXiv:2402.09683.

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

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

Katherine M Collins, Valerie Chen, Ilia Sucholutsky, Hannah Rose Kirk, Malak Sadek, Holli Sargeant, Ameet Talwalkar, Adrian Weller, and Umang Bhatt. 2024. Modulating language model experiences through frictions. *CoRR*.

Kenneth James Williams Craik. 1943. *The nature of explanation*, volume 445. CUP Archive.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. 2024. Ultrafeedback: Boosting language models with scaled ai feedback.

In *Forty-first International Conference on Machine Learning*.

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

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In *Forty-first International Conference on Machine Learning*.

Adam Fisch, Jacob Eisenstein, Vicky Zayats, Alekh Agarwal, Ahmad Beirami, Chirag Nagpal, Pete Shaw, and Jonathan Berant. 2024. Robust preference optimization through reward model distillation. *Preprint*, arXiv:2405.19316.

Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. 2023. Aligning language models with preferences through f-divergence minimization. *arXiv preprint arXiv:2302.08215*.

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

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

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

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

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

Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. 2017. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35.

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

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

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. 2024b. Common ground tracking in multimodal dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3587–3602, Torino, Italia. ELRA and ICCL.

Gary Klein, Paul J Feltovich, Jeffrey M Bradshaw, and David D Woods. 2005. Common ground and coordination in joint activity. *Organizational simulation*, 53:139–184.

Tomasz Korbak, Hady Elsahar, Germán Kruszewski, and Marc Dymetman. 2022. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. *Advances in Neural Information Processing Systems*, 35:16203–16220.

Matthew A Koschmann. 2016. The communicative accomplishment of collaboration failure. *Journal of Communication*, 66(3):409–432.

Nikhil Krishnaswamy, William Pickard, Brittany Cates, Nathaniel Blanchard, and James Pustejovsky. 2022. The voxworld platform for multimodal embodied agents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1529–1541.

Nathan Lambert, Valentina Pyatkin, Jacob Daniel Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Raghavi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hanna Hajishirzi. 2024. RewardBench: Evaluating reward models for language modeling. *ArXiv*, abs/2403.13787.

Peter R Lewis and Ştefan Sarkadi. 2024. Reflective artificial intelligence. *Minds and Machines*, 34(2):1–30.

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

Zekun Li, Wenhu Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. 2023b. Controllable dialogue simulation with in-context learning. *Preprint*, arXiv:2210.04185.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. 2024. Llava-plus: Learning to use tools for creating multimodal agents. In *European Conference on Computer Vision*, pages 126–142. Springer.

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

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

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2020. Sparse text generation. *Preprint*, arXiv:2004.02644.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Preprint*, arXiv:2405.14734.

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

Abhijnan Nath, Shadi Manafi Avari, Avyakta Chelle, and Nikhil Krishnaswamy. 2024a. Okay, let's do this! modeling event coreference with generated rationales and knowledge distillation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3931–3946.

Abhijnan Nath, Changsoo Jung, Ethan Seefried, and Nikhil Krishnaswamy. 2024b. Simultaneous reward distillation and preference learning: Get you a language model who can do both. *arXiv preprint arXiv:2410.08458*.

Abhijnan Nath, Videep Venkatesha, Mariah Bradford, Avyakta Chelle, Austin Youngren, Carlos Mabrey, Nathaniel Blanchard, and Nikhil Krishnaswamy. 2024c. "any other thoughts, hedgehog?" linking deliberation chains in collaborative dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5297–5314.

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

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. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Alizée Pace, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2024. West-of-n: Synthetic preferences for self-improving reward models. *Preprint*, arXiv:2401.12086.

Eric Pacuit. 2017. *Neighborhood semantics for modal logic*. Springer.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *Preprint*, arXiv:2402.13228.

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

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733*.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. 2019. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *Preprint*, arXiv:1910.00177.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

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

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

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8689–8696.

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. 2024. Direct nash optimization: Teaching language models to self-improve with general preferences. *ArXiv*, abs/2404.03715.

Stuart J Russell and Peter Norvig. 2016. *Artificial intelligence: a modern approach*. Pearson.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv: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. 2024. Multi-turn reinforcement learning from preference human feedback. *Preprint*, arXiv:2405.14655.

Margrethe Sønneland. 2019. Friction in fiction: A study of the importance of open problems for literary conversations. *L1-Educational Studies in Language and Literature*, 19:1–28.

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

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

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

Corbyn Terpstra, Ibrahim Khebour, Mariah Bradford, Brett Wisniewski, Nikhil Krishnaswamy, and Nathaniel Blanchard. 2023. How good is automatic segmentation as a multimodal discourse annotation aid? In *Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantics (ISA-19)*, pages 75–81.

Jingxuan Tu, Kyeongmin Rim, Bingyang Ye, Kenneth Lai, and James Pustejovsky. 2024. Dense paraphrasing for multimodal dialogue interpretation. *Frontiers in artificial intelligence*, 7:1479905.

Amos Tversky. 1969. Intransitivity of preferences. *Psychological review*, 76(1):31.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Johan van Benthem, David Fernández-Duque, and Eric Pacuit. 2014. Evidence and plausibility in neighborhood structures. *Annals of Pure and Applied Logic*, 165(1):106–133.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-ai collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658.

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

Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. 2017. A survey of preference-based reinforcement learning methods. *The Journal of Machine Learning Research*, 18(1):4945–4990.

Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. 2023. Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682*.

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. 2024. Regularizing hidden states enables learning generalizable reward model for llms. *Preprint*, arXiv:2406.10216.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models, 2023. *URL https://arxiv. org/pdf/2305.10601. pdf*.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2022. MultiWOZ 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 351–360, Edinburgh, UK. Association for Computational Linguistics.

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

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

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

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. 2024. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *arXiv preprint arXiv:2406.09136*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NeurIPS Datasets and Benchmarks Track*.

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

## A Functional Definition and Samples of Naturally-Occurring Friction

The functional operative definition of friction in collaborative contexts that we used is given below. This definition was used when annotating the WTD for naturally-occurring frictive utterances, and used to construct the prompt for friction intervention generation, following work by Oinas-Kukkonen and Harjumaa (2009) and Karadzhov et al. (2023).

**Functional Definition of Friction in Collaborative Tasks**

Frictive interventions in this setting acts as indirect persuasion (Oinas-Kukkonen and Harjumaa, 2009) where participants are passively prompted to reevaluate their belief-states and incorrect assumptions or propositions, in light of incoming goal-specific evidence. We define productive or positive friction as interventions that act as indirect persuasion: agentic interventions that prompt participants to reevaluate their beliefs and assumptions about the task state, primarily but not solely, in light of incoming evidence (say, occurrences in the physical environment or a correct "declaration" previously occurring in the dialogue that any participant missed) that negates their preconceived notions about the state of the task. We call this indirect persuasion since we do not want our friction agent to directly offer hints about the task and thereby biasing task performance or negatively affecting the deliberation process that is proven to beneficial for successful task completion in reasoning-based, collaborative tasks (Karadzhov et al., 2023).

Table 3 shows a sample friction annotation and training sample from the Weights Task Dataset, consisting of the dialogue history $x$, GPT-4o-identified frictive state $\phi$, rationale, and preferred and dispreferred friction interventions $f_w$ and $f_l$. Because the WTD is a multimodal dataset, the transcriptions we use are enriched using *dense paraphrasing* (Tu et al., 2024), a textual enrichment technique that uses the multimodal channels to decontextualize referents and in this case transform contextually-dependent phrasings such as demonstratives to explicit denotations of the content. For example, under dense paraphrasing, "seems like *these* might be about the same" while the speaker in the video is pointing to the red and blue blocks becomes "seems like *red block, blue block* might be about the same." The dense paraphrased utterances are included as part of the publicly-available WTD (Khebour et al., 2024a,b).

## B Frictive-state conditioning and RLHF

In its simplest formulation within Chain-of-Thought (CoT) settings (Wei et al., 2023), the friction agent is modeled as a policy distribution $\pi_f$ that sequentially generates frictive states, sampling $\phi_i \sim \pi_f(\cdot \mid x, \phi_1, \ldots, \phi_{i-1})$, and ultimately producing the final friction intervention $f \sim \pi_f(\cdot \mid x, \phi_1, \ldots, \phi_n)$. Here, $x$ represents the dialogue history, $f$ denotes the intervention, and $\phi$ consists of sequentially sampled frictive state tokens, analogous to "thoughts" in standard CoT-based reasoning frameworks (Yao et al., 2023). Unlike standard CoT-based alignment, which relies on self-rewarding strategies, we frame friction agent alignment within preference-based RL (PbRL; Wirth et al. (2017)). Prior work (Zhang et al., 2024) shows CoT frameworks benefit significantly from contrastive signals in preference learning.

In this setting, we define the human preference probability $\mathcal{P}(f \succ \phi)$ as the probability that an expert annotator would prefer $f$ over maintaining the frictive state $\phi$, given prior dialogue history, $x$. The key insight is that to retrieve the optimal policy $\pi_f^*$, we can leverage established methods from RLHF and PbRL by formulating the problem as a KL-divergence constrained minimum relative entropy optimization (Ziebart et al., 2008), a well-known approach with a closed-form solution (Peng et al., 2019).

$$J^*_{\text{RLHF}}(\pi_f) = \max_{\pi_f} \mathbb{E}_{f \sim \pi_f} \left[\mathcal{P}(f \succ \phi \mid x)\right] - \beta D_{\text{KL}}(\pi_f \parallel \pi_{\text{ref}}). \quad (3)$$

This formulation (Eq. 3)—where $J^*_{\text{RLHF}}$ enforces a KL-based "soft"-constraint on the parametric form of $\pi_f^*$ wrt the reference policy $\pi_{\text{ref}}$—provides crucial tradeoffs between training stability and balancing exploration vs exploitation. Specifically,

| Field | Content |
|---|---|
| **Dialogue History** ($x$) | P1: i guess if red block red one's ten grams P2: we got red ten P1: seems like red block, blue block might be about the same P3: i would agree yeah so blue block one's ten P3: Alright let's see if we can find a twenty P3: Too heavy so P2: Way too heavy P2: this is a sensitive scale P2: Looks like about twenty P1: that's looking pretty even P3: Alright let's see if we can find a thirty P1: so yellow block one is noticeably heavier than P2: probably yellow block big sucker P1: the purple ish one P1: making sure that purple block didn't have the weight at the bottom P2: it's just stuff written at the bottom that's a so red block, green block's a ten and a twenty right now right that's looking P2: Well P2: red block, blue block, green block, yellow block, purple block're increments of ten i would say that's probably P1: Yeah I think P2: cause purple block's also a twenty let's double check that purple block's not also a twenty P1: yeah it looks a little P2: cause it um just purple block one there P3: is blue block one a twenty P2: ok so purple block's more than twenty but it almost seems like the thirty takes it past but P2: it's so sensitive P2: if red block, blue block, green block, yellow block, purple block're only in increments of ten purple block has to be |
| **Frictive state** ($\phi$) | P2 initially identified the red and blue blocks as both 10 grams and has speculated about the green at 20 grams, but is uncertain about the actual weights of the yellow and purple blocks. |
| **Rationale** | P2 suggests that the red, green, and yellow blocks are all in increments of ten. Encourage a double-check. |
| **Preferred Friction** ($f_w$) | Since the purple block seems heavier and we're unsure about its exact weight, should we double-check the increments of ten assumption? Maybe the purple block doesn't fit this pattern. |
| **Dispreferred Friction** ($f_l$) | You know, the purple block being heavier might actually mean the blocks aren't increasing consistently at all. What if the increments are random, like 10, 15, or 25 grams, and we're forcing a pattern that isn't there? |

Table 3: A transcribed, sparse collaborative dialogue from the Weights Task Dataset (Khebour et al., 2024a) with frictive states and friction interventions. Preferred and dispreferred friction interventions are shown at the bottom. Positive friction interventions prompt participants to reevaluate their assumptions with frictive states (evolution of beliefs and rationales) providing indirect hints and directions. Here, P2's uncertainty about the green block and assumption of weight increment by 10g is addressed by the positive friction. In contrast, the dispreferred intervention introduces randomness, instigating the group to abandon structured reasoning.

$J^*_{\text{RLHF}}$ ensures that $\pi^*_f$ retrieves the best possible preference probabilities for its generated interventions $f$, as assigned by $\mathcal{P}(f \succ \phi)$, whether over the distribution of preferences encoded in an offline dataset (Rafailov et al., 2024b) or from online sampling $\sim \pi_f$ during training (Schulman et al., 2017) while being distributionally close to an "already-good" imitator, the Supervised-finetuned (SFT) reference model (Hussein et al., 2017). **Notice that unlike standard RLHF, we formulate $J^*_{\text{RLHF}}$ such that $\pi^*_f$ takes the form $\pi^*_f(\cdot \mid \phi, x)$ is explicitly conditioned on the frictive state $\phi$, apart from $x$.** This is intentional since we hypothesize that an ideal friction agent does *not* intervene arbitrarily, causing distraction in collaborative tasks and is conditioned to resolve the lack of common ground thereof between human collaborators, *by definition*—as observed in $\phi$. While prior work (Choi et al., 2024; Zhang et al., 2024) explores preference alignment in LLMs in such CoT-conditioned scenarios, we provide a more principled approach to proving the existence and the uniqueness of $\pi^*_f$ that $J^*_{\text{RLHF}}$ seeks to retrieve. Mathematically,

$$\pi^*_f = \frac{\pi_{\text{ref}} \exp(\beta^{-1}\mathcal{P}(f \succ \phi|x))}{Z^*(\phi, x)}, \qquad (4)$$

where $Z^* = \sum_{f'} \pi_{\text{ref}} \exp(\beta^{-1}\mathcal{P}(f' \succ \phi|x))$ is the partition function which is fixed and does not depend on $f$ and can be safely ignored in the optimization of $J^*_{\text{RLHF}}$ (Rafailov et al., 2024b). See Appendix B.1 and Equation (8) for the full-proof and optimal policy form respectively.

### B.1 Existence and uniqueness of the optimal friction intervention policy

In order to derive an empirical offline (supervised) preference learning loss from the complicated two-staged FAAF-alignment objective defined in Equation (2), we use a divide and conquer approach—*our core insight here is to express the preference of interventions conditioned on the frictive states in terms of two mutually supportive "twin" policies.* As such, we first derive the inner maximization loop of Eq. 2 to get an analytical expression of the optimal frictive intervention policy, $\pi^*_f$ as shown in the proof for Eq. 8. However, we observe that $\pi^*_f$ in its analytical form is not fully expressive since it does *not* contain the optimal frictive-state policy $\pi^*_\phi$ term. Therefore, we propose a novel method to derive $\pi^*_\phi$ using a Lagrangian formulation. We show the detailed derivation for this part in Appendix C including supporting results from Lemma 3 and Lemma 6.

This above result is one of our *main contributions* since it lets us express the preference for any intervention $f_1$ over $f_2$ analytically in terms of **both** the optimal friction intervention policy ($\pi^*_f(\cdot \mid \phi, x)$) and the optimal frictive-state policy ($\pi^*_\phi(\cdot|x)$). Finally, this core result is used to propose a straightforward supervised ($\ell_2$) objective—similar in spirit to IPO (Azar et al.,

2024)—that empirically regresses the predicted preference expression derived from $\pi_f^*(\cdot \mid \phi, x)$ and $\pi_\phi^*(\cdot|x)$ to the observed relative preferences $p(f_1 \succ f_2 \mid x)$ (relative to $\phi$), assuming access to a large-enough preference-annotated dataset of frictive interventions. Notably, this objective is optimized by a *single* parametrized policy that leverages the inherent expressivity of LLMs with billions of parameters.

In particular, this FAAF objective formulation avoids some of the policy degeneracy issues that popular supervised "offline" alignment algorithms like Direct Preference Optimization (DPO) (Rafailov et al., 2024b) face due to its unbounded rewards. Additionally, unlike Fisch et al. (2024), our regression objective works directly on preference labels and does not require an external reward model in avoiding such degeneracies. Finally, we also prove that FAAF-trained policies are unique solutions in the policy space in Theorem 2.

For completeness, we first prove the existence of the optimal friction/frictive intervention policy that solves the inner maximization of our two-part minimax objective. The structural solution to this objective is well-studied in the RL/control-theory literature including popular frameworks in preference alignment in LLMs (Ziebart et al., 2008; Peng et al., 2019; Rafailov et al., 2024b; Azar et al., 2024) as well as Chain-of-Thought (CoT)-based preference alignment frameworks (Choi et al., 2024). We show how it specifically applies to our unique parametrization. Our proof follows similar logic as Azar et al. (2024). Let us recall two-part minimax objective (Eq. 2) for clarity here:

$$J^*_{\mathsf{FAAF}} = \min_{\pi_\phi} \max_{\pi_f} \mathbb{E}_{\substack{x\sim\rho \\ \phi\sim\pi_\phi(\cdot|x) \\ f\sim\pi_f(\cdot|\phi,x)}} \Big[ \mathcal{P}(f \succ \phi \mid x) - \beta D_{\mathrm{KL}}(\pi_f \parallel \pi_{\text{ref}} \mid \phi, x) + \beta D_{\mathrm{KL}}(\pi_\phi \parallel \pi_{\text{ref}} \mid x) \Big]. \tag{5}$$

For fixed $\pi_\phi$, the inner maximization reduces to our regularized objective:

$$\begin{aligned} \mathcal{L}_\beta(\pi_f) &= \mathbb{E}_{f\sim\pi_f}[p(f \succ \phi|x)] - \\ &\quad \beta D_{\mathrm{KL}}(\pi_f \parallel \pi_{\text{ref}}|\phi, x), \\ &= \sum_f \pi_f(f|\phi, x) p(f \succ \phi|x) - \\ &\quad \beta D_{\mathrm{KL}}(\pi_f \parallel \pi_{\text{ref}}|\phi, x), \end{aligned} \tag{6}$$

where $f \in \mathcal{F}$ is from a finite friction token alphabet $\mathcal{F}$, $p(f \succ \phi|x)$ maps elements of $\mathcal{F}$ to the utility of generating a frictive intervention $f$ defined as the preference of $f$ over the frictive-state $\phi$, given context $x$, $\beta \in \mathbb{R}^*_+$ is a strictly positive real number, and $\pi_f, \pi_{\text{ref}}$ are conditional probability distributions. In particular, notice that the conditional probability distribution $\pi_f(f|\phi, x)$ can be identified as a positive real function satisfying:

$$\sum_f \pi_f(f|\phi, x) = 1. \tag{7}$$

Now, if we define the optimal friction intervention policy $\pi_f^*$ as:

$$\pi_f^*(f|\phi, x) = \frac{\pi_{\text{ref}}(f|\phi, x) \exp(\beta^{-1} p(f \succ \phi|x))}{Z^*(\phi, x)}, \tag{8}$$

recalling Eq. 1, where $Z^*(\phi, x) = \sum_{f'} \pi_{\text{ref}}(f'|\phi, x) \exp(\beta^{-1} p(f' \succ \phi|x))$, then, under the previous definitions, we have:

$$\pi_f^* = \arg\max_{\pi_f} \mathcal{L}_\beta(\pi_f) \tag{9}$$

*Proof.*

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

By definition of the KL divergence, we know that $\pi^*_f = \arg\max_{\pi_f}\big[-D_{\mathrm{KL}}(\pi_f \parallel \pi^*_f)\big]$ and as:

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

where $\log Z^*(\phi,x)$ is the partition function (Peng et al., 2019; Rafailov et al., 2024b) and has no dependency on $\pi_f$ and $\beta \in \mathbb{R}^*_+$ is a strictly positive real number. Therefore, the argmax of $-D_{\mathrm{KL}}(\pi_f \parallel \pi^*_f)$ coincides with that of $\mathcal{L}_\beta(\pi_f)$, concluding the proof. □

**Lemma 1** (Value of Inner Maximization)**.** When substituting the optimal friction intervention policy $\pi_f^*$, as derived in Eq. 8, into Eq. 5, the objective in Eq. 5 reduces to:

$$J^*_{\mathsf{FAAF}} = \min_{\pi_\phi} \mathbb{E}_{x\sim\rho,\phi\sim\pi_\phi(\cdot|x)}[\beta \log(Z^*(\phi,x)) + \beta D_{KL}(\pi_\phi||\pi_{ref}|x)] \tag{10}$$

*Proof.* Substituting $\pi_f^*$ into the KL divergence term:

$$\begin{aligned} &D_{KL}(\pi_f^*||\pi_{ref}|\phi,x) \\ &= \mathbb{E}_{f\sim\pi_f^*}\Big[\log(\pi_f^*(f|\phi,x)) - \\ &\quad \log(\pi_{\text{ref}}(f|\phi,x))\Big] \\ &= \mathbb{E}_{f\sim\pi_f^*}\Big[\frac{p(f\succ\phi|x)}{\beta} - \\ &\quad \log(Z^*(\phi,x))\Big] \end{aligned} \tag{11}$$

The original objective becomes:

$$\begin{aligned} &p(f\succ\phi|x) - \beta\mathbb{E}_{f\sim\pi_f^*}\Big[\frac{p(f\succ\phi|x)}{\beta} - \\ &\quad \log(Z^*(\phi,x))\Big] \\ &= \beta\log(Z^*(\phi,x)) \end{aligned} \tag{12}$$

The result follows by substituting this value back into the full objective. □

## C Derivation of Optimal Frictive State Policy

We begin with the reduced objective function after solving the inner maximization as shown in Lemma 1.

$$J^*_{\mathsf{FAAF}} = \min_{\pi_\phi} \mathbb{E}_{x\sim\rho,\phi\sim\pi_\phi(\cdot|x)}\big[\beta\log(Z^*(\phi,x)) + \beta D_{KL}(\pi_\phi||\pi_{\text{ref}}|x)\big] \tag{13}$$

The Kullback-Leibler divergence term expands as follows:

$$D_{KL}(\pi_\phi||\pi_{\text{ref}}|x) = \mathbb{E}_{\phi\sim\pi_\phi}\left[\log\frac{\pi_\phi(\phi|x)}{\pi_{\text{ref}}(\phi|x)}\right] \tag{14}$$

Substituting this back into our objective:

$$J^*_{\mathsf{FAAF}} = \min_{\pi_\phi} \mathbb{E}_{\phi\sim\pi_\phi}\Big[\beta\log(Z^*(\phi,x)) + \beta\log(\pi_\phi(\phi|x)) - \beta\log(\pi_{\text{ref}}(\phi|x))\Big] \tag{15}$$

Since $\pi_\phi$ must be a valid probability distribution satisfying $\sum_\phi \pi_\phi(\phi|x) = 1$, we introduce a Lagrange multiplier $\lambda$ and define the corresponding Lagrangian function to derive the optimality conditions:

$$\begin{aligned} L(\pi_\phi) &= \mathbb{E}_{\phi\sim\pi_\phi}\Big[\beta\log(Z^*(\phi,x)) + \\ &\quad \beta\log(\pi_\phi(\phi|x)) - \beta\log(\pi_{\text{ref}}(\phi|x))\Big] + \\ &\quad \lambda\left(1 - \sum_\phi \pi_\phi(\phi|x)\right) \\ &= \sum_\phi \pi_\phi(\phi|x)\Big[\beta\log(Z^*(\phi,x)) + \\ &\quad \beta\log(\pi_\phi(\phi|x)) - \beta\log(\pi_{ref}(\phi|x))\Big] \\ &\quad + \lambda\left(1 - \sum_\phi \pi_\phi(\phi|x)\right). \end{aligned} \tag{16}$$

Now, to find the optimal policy $\pi^*_\phi(\phi|x)$, we take the derivative of the Lagrangian with respect to $\pi_\phi(\phi|x)$ and equate it to zero:

$$\begin{aligned} \frac{\delta L}{\delta\pi_\phi(\phi|x)} &= \beta\log(Z^*(\phi,x)) + \\ &\quad \beta\frac{\delta}{\delta\pi_\phi}\Big[\pi_\phi(\phi|x)\log(\pi_\phi(\phi|x))\Big] \\ &\quad - \beta\log(\pi_{ref}(\phi|x)) - \lambda = 0. \end{aligned} \tag{17}$$

From the standard functional derivative of entropy $\frac{\delta}{\delta\pi_\phi}\Big[\pi_\phi(\phi|x)\log(\pi_\phi(\phi|x))\Big] = 1 + \log(\pi_\phi(\phi|x))$, we obtain:

$$\begin{aligned} &\beta\log(Z^*(\phi,x)) + \beta(1+\log(\pi_\phi(\phi|x))) - \\ &\quad \beta\log(\pi_{ref}(\phi|x)) + \lambda = 0. \end{aligned} \tag{18}$$

Rearranging the terms:

$$\begin{aligned} \log(\pi_\phi(\phi|x)) &= \log(\pi_{ref}(\phi|x)) - \\ &\quad \log(Z^*(\phi,x)) - \frac{\lambda}{\beta} - 1. \end{aligned} \tag{19}$$

Taking the exponential on both sides:

$$\pi_\phi(\phi|x) = e^{-1-\frac{\lambda}{\beta}}\pi_{\text{ref}}(\phi|x)e^{-\log Z^*(\phi,x)}. \quad (20)$$

To ensure $\pi_\phi(\phi|x)$ is a valid probability distribution, we define the normalization constant:

$$Z(x) = \sum_\phi \pi_{\text{ref}}(\phi|x)e^{-\beta \log Z^*(\phi,x)}. \quad (21)$$

Thus, the optimal frictive-state policy is:

$$\pi^*_\phi(\phi|x) = \frac{\pi_{\text{ref}}(\phi|x)e^{-\beta \log Z^*(\phi,x)}}{Z(x)}. \quad (22)$$

Notice that without losing any generality, we can parametrize the above optimal frictive-state policy with any outcome $f$ consistent with the structure in Eq. 22 as follows:

$$\pi^*_\phi(f|x) = \frac{\pi_{\text{ref}}(f|x)}{Z(x)}e^{-\beta \log(Z^*(f,x))}. \quad (23)$$

Note that although this formulation of the optimal frictive-state policy ($\pi^*_\phi(\phi|x)$) is an analytical solution to $J^*$ from Eq. 13, we still need to represent $\pi^*_\phi(\phi|x)$ in terms of the optimal friction intervention policy, $\pi^*_f(\cdot \mid \phi, x)$ proposed in Eq. 8 and the preference probabilities $p(f \succ \phi|x)$, the preference probability of the friction $f$ over the frictive-state $\phi$, given context $x$. This is crucial to derive the empirical FAAF optimization objective that can be used for standard offline learning. Therefore, to represent the $p(f \succ \phi|x)$ in terms of $\pi^*_f(\cdot \mid \phi, x)$, we take the logarithm of Eq. 8 on both sides and some algebra, we obtain:

$$\log(\pi^*_f(f|\phi,x)) = \frac{p(f \succ \phi|x)}{\beta} + \log(\pi_{\text{ref}}(f|\phi,x)) - \log(Z^*(\phi,x)).$$

Multiplying both sides by $\beta$ and rearranging terms, we obtain:

$$p(f \succ \phi|x) = \beta[\log(\pi^*_f(f|\phi,x)) - \log(\pi_{\text{ref}}(f|\phi,x)) + \log(Z^*(\phi,x))]. \quad (24)$$

Similar to Munos et al. (2023), Azar et al. (2024), and Choi et al. (2024), we can apply the identity that $p(\phi \succ \phi|x) = \frac{1}{2}$ and substitute $f = \phi$ into the previous equation and derive:

$$\frac{1}{2} = \beta[\log(\pi^*_f(\phi|\phi,x)) - \log(\pi_{\text{ref}}(\phi|\phi,x)) + \log(Z^*(\phi,x))]. \quad (25)$$

Solving for $\log(Z^*(\phi,x))$ gives:

$$\log(Z^*(\phi,x)) = \frac{1}{2\beta} - [\log(\pi^*_f(\phi|\phi,x)) - \log(\pi_{\text{ref}}(\phi|\phi,x))]. \quad (26)$$

Substituting this back into Eq. 24 results in:

$$\begin{aligned} p(f \succ \phi|x) &= \beta[\log(\pi^*_f(f|\phi,x)) - \log(\pi_{\text{ref}}(f|\phi,x)) \\ &\quad + \frac{1}{2\beta} - (\log(\pi^*_f(\phi|\phi,x)) - \log(\pi_{\text{ref}}(\phi|\phi,x)))] \\ &= \beta \log\left(\frac{\pi^*_f(f|\phi,x)}{\pi_{\text{ref}}(f|\phi,x)}\right) + \frac{1}{2} \\ &\quad - \beta \log\left(\frac{\pi^*_f(\phi|\phi,x)}{\pi_{\text{ref}}(\phi|\phi,x)}\right). \end{aligned} \quad (27)$$

The $\log\left(\frac{\pi^*_f(\phi|\phi,x)}{\pi_{\text{ref}}(\phi|\phi,x)}\right)$ term in the above step is a self-referential term signifying the friction intervention policy's ($\pi^*_f(\cdot \mid \phi, x)$) estimate of the frictive state given $\phi$. However, this term does *not* provide much information on the regularized preference in terms of the frictive state policy. Recall that our outer minimization objective operates over $\pi_\phi(\cdot|x)$. Fortunately, we can use our results from Lemma 3 and Lemma 6 to express Eq. 27 in terms of the optimal frictive state policy $\pi^*_\phi(\cdot|x)$. Therefore, from Lemma 6 we can express $\pi^*_f$ and $\pi_{\text{ref}}$ as follows:

For the optimal policy $\pi^*_f$:

$$\log(\pi^*_f(\phi|\phi,x)) = \log(\pi^*_\phi(\phi|x)) - \log(\pi^*_\phi(f|x)) \quad (28)$$

For the reference policy $\pi_{\text{ref}}$:

$$\log(\pi_{\text{ref}}(\phi|\phi,x)) = \log(\pi_{\text{ref}}(\phi|x)) - \log(\pi_{\text{ref}}(f|x)) \quad (29)$$

Now, substituting these expressions into Equa-

tion (27), we get:

$$
\begin{aligned}
p(f \succ \phi|x) = \beta\Big[ & \log(\pi^*_f(f|\phi,x)) - \\
& \log(\pi_{\text{ref}}(f|\phi,x)) + \\
& \frac{1}{2\beta} - \log(\pi^*_\phi(\phi|x)) + \\
& \log(\pi^*_\phi(f|x)) - \log(\pi_{\text{ref}}(\phi|x)) + \\
& \log(\pi_{\text{ref}}(f|x))\Big] \\
= \beta\Big[ & \log(\pi^*_f(f|\phi,x)) - \\
& \log(\pi_{\text{ref}}(f|\phi,x)) + \\
& \frac{1}{2\beta} - \Big(\log(\pi^*_\phi(\phi|x)) - \\
& \log(\pi_{\text{ref}}(\phi|x)) - \\
& (\log(\pi^*_\phi(f|x)) - \log(\pi_{\text{ref}}(f|x)))\Big)\Big] \\
= \beta\Bigg[ & \log\left(\frac{\pi^*_f(f|\phi,x)}{\pi_{\text{ref}}(f|\phi,x)}\right) + \frac{1}{2\beta} + \\
& \log\left(\frac{\pi^*_\phi(f|x)}{\pi_{\text{ref}}(f|x)}\right) - \\
& \log\left(\frac{\pi^*_\phi(\phi|x)}{\pi_{\text{ref}}(\phi|x)}\right)\Bigg]. \qquad (30)
\end{aligned}
$$

Now, replacing $f$ by $f_1$ in $p(f \succ \phi|x)$:

$$
\begin{aligned}
p(f_1 \succ \phi|x) = \beta\Bigg[ & \log\left(\frac{\pi^*_f(f_1|\phi,x)}{\pi_{\text{ref}}(f_1|\phi,x)}\right) + \\
& \frac{1}{2\beta} + \log\left(\frac{\pi^*_\phi(f_1|x)}{\pi_{\text{ref}}(f_1|x)}\right) - \\
& \log\left(\frac{\pi^*_\phi(\phi|x)}{\pi_{\text{ref}}(\phi|x)}\right)\Bigg] \qquad (31)
\end{aligned}
$$

Similarly, expressing $f_2$ in $p(f \succ \phi|x)$, we obtain:

$$
\begin{aligned}
p(f_2 \succ \phi|x) = \beta\Bigg[ & \log\left(\frac{\pi^*_f(f_2|\phi,x)}{\pi_{\text{ref}}(f_2|\phi,x)}\right) + \\
& \frac{1}{2\beta} + \log\left(\frac{\pi^*_\phi(f_2|x)}{\pi_{\text{ref}}(f_2|x)}\right) - \\
& \log\left(\frac{\pi^*_\phi(\phi|x)}{\pi_{\text{ref}}(\phi|x)}\right)\Bigg] \qquad (32)
\end{aligned}
$$

Now, expressing $p(f_1 \succ \phi|x) - p(f_2 \succ \phi|x)$, the relative preference probability of $f_1$ over $f_2$ given $\phi$ and $x$, we observe that $\log\left(\frac{\pi^*_\phi(\phi|x)}{\pi_{\text{ref}}(\phi|x)}\right)$ terms cancel out and we derive:

$$
\begin{aligned}
& p(f_1 \succ \phi|x) - p(f_2 \succ \phi|x) = \\
& \beta\Bigg[\log\left(\frac{\pi^*_f(f_1|\phi,x)}{\pi_{\text{ref}}(f_1|\phi,x)}\right) + \frac{1}{2\beta} + \\
& \log\left(\frac{\pi^*_\phi(f_1|x)}{\pi_{\text{ref}}(f_1|x)}\right) - \log\left(\frac{\pi^*_\phi(\phi|x)}{\pi_{\text{ref}}(\phi|x)}\right)\Bigg] \\
& - \beta\Bigg[\log\left(\frac{\pi^*_f(f_2|\phi,x)}{\pi_{\text{ref}}(f_2|\phi,x)}\right) + \frac{1}{2\beta} + \\
& \log\left(\frac{\pi^*_\phi(f_2|x)}{\pi_{\text{ref}}(f_2|x)}\right) - \log\left(\frac{\pi^*_\phi(\phi|x)}{\pi_{\text{ref}}(\phi|x)}\right)\Bigg] \\
& = \beta\Bigg[\log\left(\frac{\pi^*_f(f_1|\phi,x)}{\pi_{\text{ref}}(f_1|\phi,x)}\right) - \log\left(\frac{\pi^*_f(f_2|\phi,x)}{\pi_{\text{ref}}(f_2|\phi,x)}\right) \\
& + \log\left(\frac{\pi^*_\phi(f_1|x)}{\pi_{\text{ref}}(f_1|x)}\right) - \log\left(\frac{\pi^*_\phi(f_2|x)}{\pi_{\text{ref}}(f_2|x)}\right)\Bigg] \qquad (33)
\end{aligned}
$$

This above result is one of our *core contributions* since it lets us express the relative preference of any friction intervention $f_1$ over $f_2$ given a frictive state ($\phi$) analytically in terms of **both** the optimal friction intervention policy ($(\pi^*_f(\cdot \mid \phi, x))$) and the optimal frictive state policy ($\pi^*_\phi(\cdot|x)$):

$$
\begin{aligned}
& p(f_1 \succ \phi|x) - p(f_2 \succ \phi|x) = \\
& \beta\Bigg[\log\left(\frac{\pi^*_f(f_1|\phi,x)}{\pi_{\text{ref}}(f_1|\phi,x)}\right) - \log\left(\frac{\pi^*_f(f_2|\phi,x)}{\pi_{\text{ref}}(f_2|\phi,x)}\right) \\
& + \log\left(\frac{\pi^*_\phi(f_1|x)}{\pi_{\text{ref}}(f_1|x)}\right) - \log\left(\frac{\pi^*_\phi(f_2|x)}{\pi_{\text{ref}}(f_2|x)}\right)\Bigg] \qquad (34)
\end{aligned}
$$

Following a standard approach for empirical estimation of the LHS (Azar et al., 2024) in the above equation, one can learn *both* the optimal friction intervention policy $\pi^*_f$ and the frictive-state policy $\pi^*_\phi$ using a trainable policy $\pi_\theta$, parametrized with $\theta$. The core insight here is to exploit the expressive nature of LLMs' hidden representations with billions of parameters to learn a *single* optimal policy. A reasonable choice here is to train $\pi_\theta$ through an $\ell_2$ loss (Fisch et al., 2024) that enforces the relative preference ordering between any pair of friction interventions $(f_1, f_2)$ with implicit reward estimates from the RHS of Eq. 34. However, unlike (Fisch et al., 2024), our approach in enforcing this constraint does not require access to

an external reward model or an "oracle" for pointwise reward estimates, assuming we have access to labeled preference feedback in samples. Additionally, the $\ell_2$ formulation avoids placing a unbounded logit or a inverse sigmoid function over the preference since this has been shown to cause non-trivial policy degeneracy issues in learning algorithms like DPO (Azar et al., 2024). Applying this $\ell_2$ loss, we derive:

$$\mathcal{L}_{\pi_\theta} = \mathbb{E}_{\substack{x\sim\rho \\ \phi\sim\pi_\theta(\cdot|x) \\ f_1,f_2\sim\pi_\theta(\cdot|\phi,x)}} \Bigg( p(f_1 \succ \phi|x) - p(f_2 \succ \phi|x) - \beta \Bigg[ \log\left(\frac{\pi_\theta(f_1|\phi,x)}{\pi_{\text{ref}}(f_1|\phi,x)}\right) - \log\left(\frac{\pi_\theta(f_2|\phi,x)}{\pi_{\text{ref}}(f_2|\phi,x)}\right) + \log\left(\frac{\pi_\theta(f_1|x)}{\pi_{\text{ref}}(f_1|x)}\right) - \log\left(\frac{\pi_\theta(f_2|x)}{\pi_{\text{ref}}(f_2|x)}\right) \Bigg] \Bigg)^2 \tag{35}$$

Since the friction dataset $\mathcal{D}_\mu$ sampled from $\mu$ contains preference-annotated pairs $(f_w, f_l)$ given $\phi$ and $x$, the preference probabilities can be expressed using indicator functions as $p(f_w \succ f_l|x) = \mathbb{E}[\mathbf{1}(f_w \succ f_l|x)] = 1$ and $p(f_l \succ f_w|x) = \mathbb{E}[\mathbf{1}(f_l \succ f_w|x)] = 0$. Furthermore, the difference $p(f_w \succ f_l|x) - p(f_l \succ f_w|x) = 1 - 0 = 1$ aligns with the formulation $p(f_1 \succ \phi|x) - p(f_2 \succ \phi|x)$ when $f_1 = f_w$ and $f_2 = f_l$. Therefore, we can write our final FAAF-alignment empirical objective function ($\hat{\mathcal{L}}$) as follows:

$$\hat{\mathcal{L}}(\pi_\theta) = \mathbb{E}_{(x,\phi,f_w,f_l)\sim\mathcal{D}_\mu} \Bigg( 1 - \beta \Bigg[ \log\left(\frac{\pi_\theta(f_w|\phi,x)}{\pi_{\text{ref}}(f_w|\phi,x)}\right) - \log\left(\frac{\pi_\theta(f_l|\phi,x)}{\pi_{\text{ref}}(f_l|\phi,x)}\right) + \log\left(\frac{\pi_\theta(f_w|x)}{\pi_{\text{ref}}(f_w|x)}\right) - \log\left(\frac{\pi_\theta(f_l|x)}{\pi_{\text{ref}}(f_l|x)}\right) \Bigg] \Bigg)^2 \tag{36}$$

where $(f_w, f_l)$ represent the winning (preferred) and losing (less preferred) friction interventions respectively in each annotated pair.

$$\hat{\mathcal{L}}(\pi_\theta) = \mathbb{E}_{(x,\phi,f_w,f_l)\sim\mathcal{D}_\mu} \Bigg( 1 - \beta \Bigg[ \log\left(\frac{\pi_\theta(f_w|\phi,x)\pi_{\text{ref}}(f_l|\phi,x)}{\pi_\theta(f_l|\phi,x)\pi_{\text{ref}}(f_w|\phi,x)}\right) + \log\left(\frac{\pi_\theta(f_w|x)\pi_{\text{ref}}(f_l|x)}{\pi_\theta(f_l|x)\pi_{\text{ref}}(f_w|x)}\right) \Bigg] \Bigg)^2 \tag{37}$$

$$\hat{\mathcal{L}}(\pi_\theta) = \mathbb{E}_{(x,\phi,f_w,f_l)\sim\mathcal{D}_\mu} \Bigg( 1 - \Bigg[ \underbrace{\beta \log\left(\frac{\pi_\theta(f_w|\phi,x)\pi_{\text{ref}}(f_l|\phi,x)}{\pi_\theta(f_l|\phi,x)\pi_{\text{ref}}(f_w|\phi,x)}\right)}_{\Delta R} + \underbrace{\beta \log\left(\frac{\pi_\theta(f_w|x)\pi_{\text{ref}}(f_l|x)}{\pi_\theta(f_l|x)\pi_{\text{ref}}(f_w|x)}\right)}_{\Delta R'} \Bigg] \Bigg)^2 \tag{38}$$

where $\Delta R$ and $\Delta R'$ represent implicit reward differences (Rafailov et al., 2024b; Azar et al., 2024), the former being explicitly conditioned on the frictive state $\phi$, with no such conditioning on the latter.

**Theorem 2** (Uniqueness of FAAF Empirical Loss)**.** We prove this by contradiction. Let $\mu$ be the sampling distribution that samples friction interventions for the preference dataset, and assume $\text{Supp}(\mu) = \text{Supp}(\pi_{\text{ref}})$. Then the FAAF loss $\mathcal{L}(\pi)$ has a unique solution in policy space $\in \Pi$.

*Proof.* Assume by contradiction that there exist two distinct optimal policies $\pi_A, \pi_B \in \Pi$. By their definition, $\hat{\mathcal{L}}(\pi_A) = \hat{\mathcal{L}}(\pi_A) = 0$ as $\pi_A$ and $\pi_B$ are global minima. Consider $(s_\phi^A, s^A)$ and $(s_\phi^B, s^B)$ as their respective logit parameterizations where:

$$\pi_k(f|\phi) = \frac{\exp(s_\phi^k(f))}{\sum_{f'} \exp(s_\phi^k(f'))}$$
$$\pi_k(f) = \frac{\exp(s^k(f))}{\sum_{f'} \exp(s^k(f'))} \quad \text{for } k \in \{A, B\}$$

where $\pi_k(f|\phi)$ and $\pi_k(f)$ are the $\phi$-conditioned and $\phi$-unconditioned policies.
By the structure of our FAAF loss from Equation (38):

$$\hat{\mathcal{L}}(\pi) = \mathbb{E}_{f,f' \sim \mu}\Big[\big(1 - \beta(\Delta s_\phi + \Delta s)\big)^2\Big] \geq 0$$

Notice that adding a constant $c$ to all logits of $s_\phi$ or logits of $s$ (directionally denoted as the $(c, \ldots, c) \in \mathbb{R}$) does not affect either policy probabilities due to softmax normalization. For $\hat{\mathcal{L}}(\pi)$, this is the *only* direction where the loss function might not be strictly convex. Outside of these directions, any change in the logits would increase $\mathcal{L}(\pi)$ with strict convexity as a consequence for $\alpha \in (0, 1)$, implying:

$$\begin{aligned}\hat{\mathcal{L}}(\alpha\pi_1 + (1-\alpha)\pi_2) &< \alpha\hat{\mathcal{L}}(\pi_1) + (1-\alpha)\hat{\mathcal{L}}(\pi_2) \\ &= \alpha(0) + (1-\alpha)(0) = 0\end{aligned}$$

where the equality follows from $\pi_1, \pi_2$ being global minima, by definition. This contradicts the non-negativity of $\hat{\mathcal{L}}$, which proves the uniqueness of the FAAF objective. □

**$\hat{\mathcal{L}}(\pi_\theta)$ has no dependence on log-partition terms involving $Z^*(\phi, x)$ and $Z^*(x)$** Our final FAAF empirical objective loss in Eq. 38 has no dependence on either partition function terms. This makes it convenient for practical applications. In fact, similar to DPO's derivation (Rafailov et al., 2024b), these log-partition terms effectively cancel out in formulating the frictive state-conditioned and unconditioned implicit rewards, scaled by the KL-strength parameter $\beta$. In its essence, $\hat{\mathcal{L}}(\pi_\theta)$ regresses the DPO-based implicit rewards ($\Delta R'$ term) with an additional $\phi$-conditioned reward term ($\Delta R$ term) onto the empirically observed preference probabilities, labeled with preference labels from $\mathcal{D}_\mu$. Notice that without the $\Delta R$ term, $\hat{\mathcal{L}}(\pi_\theta)$ reduces to a structurally similar form as IPO (Azar et al., 2024), differing a constant scaling term $\beta$. This suggests that under this condition, both $\hat{\mathcal{L}}(\pi_\theta)$ and IPO objective likely have similar qualitative loss landscapes though convergence rates and optimal solutions would differ—while both lead $\pi_\theta$ toward a reward-consistent preference alignment. This also explains the somewhat similar performance of the IPO baseline and $\text{FAAF}_{\Delta R'}$ in both DeliData and WTD OPT 1.3B reward model-based win-rate evaluations, where $\text{FAAF}_{\Delta R'}$ achieves comparatively middling win-rates (Table 2).

**Lemma 3** (Sequential Choice Decomposition in Friction Agent Optimization)**.** Consider the minimax optimization between frictive-state policy $\pi_\phi$ and friction intervention policy $\pi_f$ where we seek to generate optimal friction interventions $f$ from frictive states $\phi$:

$$J^* = \min_{\pi_\phi} \max_{\pi_f} \mathbb{E}_{\substack{x \sim \rho \\ \phi \sim \pi_\phi(\cdot|x) \\ f \sim \pi_f(\cdot|\phi, x)}} \Big[p(f \succ \phi \mid x) - \beta D_{\text{KL}}(\pi_f \parallel \pi_{\text{ref}} \mid \phi, x) + \beta D_{\text{KL}}(\pi_\phi \parallel \pi_{\text{ref}} \mid x)\Big] \tag{39}$$

For any policy $\pi$ (either optimal friction policy $\pi_f^*$ or reference policy $\pi_{\text{ref}}$), the sequential choice probability decomposes as:

$$\pi(\phi|\phi, x) = \frac{\pi(\phi|x)}{\pi(f|x)} \tag{40}$$

*Proof.* The key insight in deriving this decomposition lies in understanding how optimal friction interventions are generated sequentially from frictive states. For the optimal friction policy $\pi_f^*$, consider its probability space $P_{\pi_f^*}$. By definition of conditional probability, we have $\pi_f^*(\phi|\phi, x) = \frac{P_{\pi_f^*}(\phi,\phi|x)}{P_{\pi_f^*}(\phi|x)}$. This term is crucial as it captures the policy's propensity to maintain a frictive state rather than generate a friction intervention. Under choice independence[a] within this policy space assuming a Markovian nature of friction intervention generation, we have $P_{\pi_f^*}(\phi, \phi|x) = P_{\pi_f^*}(\phi|x)P_{\pi_f^*}(\phi|x)$. With policy-specific preference probability symmetry (Munos et al., 2023; Fisch et al., 2024), the probability $P_{\pi_f^*}(\phi|x) + P_{\pi_f^*}(f|x) = 1$, reflecting the binary choice between maintaining a frictive state or generating a friction intervention, we obtain $\pi_f^*(\phi|\phi, x) = \frac{\pi_f^*(\phi|x)}{\pi_f^*(f|x)}$, where the optimality of $\pi_f^*(f|x)$ ensures $\pi_f^*(\phi|x) \leq \pi_f^*(f|x)$. A similar argument can be made in the case of $\pi_{\text{ref}}$, the reference policy, where $\pi_{\text{ref}}$'s initialization with the supervised-finetuned (SFT) model on friction interventions ensures $\pi_{\text{ref}}(f|x) \geq \pi_{\text{ref}}(\phi|x)$. This decomposition is fundamental to the minimax objective, $J^*$, as it enables expressing the KL-regularized preference probability in terms of base policy probabilities while preserving the structure necessary for optimal friction intervention generation from frictive states. □

[a] Assuming a single-step bandit setting (Rafailov et al., 2024b,a), choice independence holds since each frictive-state intervention is independent of past episodes. Using conditional probability, we express the joint probability under any policy $\pi$ as $P_\pi(\phi, \phi \mid x) = P_\pi(\phi \mid \phi, x)P_\pi(\phi \mid x)$. By choice independence, the probability of selecting $\phi$ at the second step does not depend on the first selection given $x$, i.e., $P_\pi(\phi \mid \phi, x) = P_\pi(\phi \mid x)$. Substituting this, we obtain $P_\pi(\phi, \phi \mid x) = P_\pi(\phi \mid x)P_\pi(\phi \mid x)$.

The sequential choice decomposition provides crucial insight into determining optimal timing for friction interventions. In other words, this decomposition has an interesting implication in deciding *when* is a friction intervention most desirable or cost-effective. Specifically, our derived identity $\pi(\phi|\phi, x) = \frac{\pi(\phi|x)}{\pi(f|x)}$ establishes a natural threshold mechanism through the ratio $\tau(x) = \frac{\pi_f^*(\phi|x)}{\pi_f^*(f|x)}$. When $\tau(x) \approx 1$, the policy maintains the current frictive state $\phi$, while $\tau(x) \ll 1$ triggers a friction intervention $f$. This mechanism emerges naturally from the preference probability $p(f \succ \phi|x) = \beta[\log(\pi_f^*(f|\phi, x)) - \log(\pi_{\text{ref}}(f|\phi, x)) + \frac{1}{2\beta} - (\log(\pi_f^*(\phi|\phi, x)) - \log(\pi_{\text{ref}}(\phi|\phi, x)))]$ in our minimax objective $J^*$, where $\pi_f^*$ optimally generates interventions when the likelihood ratio indicates low confidence in the current frictive state $\phi$. However, exploring this sequential decomposition and determining optimal timing in interventions is outside the scope of this paper. As such, we leave that for future work.

**Lemma 4** (Uniqueness of Intervention Thresholds)**.** The threshold $\tau(x) = \frac{\pi_f^*(\phi|x)}{\pi_f^*(f|x)}$ uniquely determines optimal intervention policy $\pi_f^*$.

*Proof.* We prove uniqueness by contradiction. Consider two potentially optimal policies $\pi_f^1$ and $\pi_f^2$ with corresponding thresholds $\tau_1(x)$ and $\tau_2(x)$. Assume $\tau_1(x) \neq \tau_2(x)$ but both policies are optimal. By optimality, their contributions to the objective $J^*$ must be equal for any observation tuple $x$, $f$ and $\phi$:

$$\begin{aligned}&\beta[\log(\pi_f^1(f|\phi,x)) - \log(\pi_{\text{ref}}(f|\phi,x)) - (\log(\tau_1(x)) - \log(\pi_{\text{ref}}(\phi|\phi,x)))] \\ &= \beta[\log(\pi_f^2(f|\phi,x)) - \log(\pi_{\text{ref}}(f|\phi,x)) - (\log(\tau_2(x)) - \log(\pi_{\text{ref}}(\phi|\phi,x)))]\end{aligned} \tag{41}$$

Simplifying and rearranging terms:

$$\log(\pi_f^1(f|\phi,x)) - \log(\tau_1(x)) = \log(\pi_f^2(f|\phi,x)) - \log(\tau_2(x)) \tag{42}$$

However, by the strict convexity of KL divergence and Jensen's inequality:

$$D_{\text{KL}}(\pi_f^1 \parallel \pi_{\text{ref}} \mid \phi,x) + D_{\text{KL}}(\pi_f^2 \parallel \pi_{\text{ref}} \mid \phi,x) > 2D_{\text{KL}}(\frac{\pi_f^1 + \pi_f^2}{2} \parallel \pi_{\text{ref}} \mid \phi,x) \tag{43}$$

This inequality implies that a mixed policy $\pi_f^{\text{avg}} = \frac{\pi_f^1 + \pi_f^2}{2}$ would achieve a lower KL divergence cost due to strict convexity and equal expected reward (regularized preference probabilities) from the equality of optimal policies. Therefore, $\pi_f^{\text{avg}}$ would achieve strictly better objective value than both $\pi_f^1$ and $\pi_f^2$, contradicting their assumed optimality. This proves threshold uniqueness. The contradiction arises because:

$$J^*(\pi_f^{\text{avg}}) > \frac{1}{2}[J^*(\pi_f^1) + J^*(\pi_f^2)] \tag{44}$$

which is impossible if both $\pi_f^1$ and $\pi_f^2$ were truly optimal. □

**Corollary 5** (Uniqueness of Optimal Policy Under Threshold Identity)**.** If two optimal intervention policies $\pi_f^1$ and $\pi_f^2$ satisfy the same threshold condition $\tau_1(x) = \tau_2(x)$ for all $x$, then $\pi_f^1 = \pi_f^2$.

*Proof.* Assume for contradiction that two distinct optimal policies $\pi_f^1$ and $\pi_f^2$ satisfy the threshold condition $\frac{\pi_f^1(\phi|x)}{\pi_f^1(f|x)} = \frac{\pi_f^2(\phi|x)}{\pi_f^2(f|x)} = \tau(x)$. Define the mixed policy $\pi_f^{\text{avg}} = \frac{1}{2}(\pi_f^1 + \pi_f^2)$, which preserves the threshold as $\tau_{\text{avg}}(x) = \tau(x)$ due to linearity, implying $\pi_f^{\text{avg}}$ is also optimal.
Now, applying Jensen's inequality to the KL divergence term in the objective, we obtain $D_{\text{KL}}(\pi_f^{\text{avg}} \parallel \pi_{\text{ref}} \mid \phi, x) \leq \frac{1}{2} D_{\text{KL}}(\pi_f^1 \parallel \pi_{\text{ref}} \mid \phi, x) + \frac{1}{2} D_{\text{KL}}(\pi_f^2 \parallel \pi_{\text{ref}} \mid \phi, x)$. Strict convexity ensures a strict inequality whenever $\pi_f^1 \neq \pi_f^2$ on a set of positive measure where $\text{supp}(\pi_{\text{ref}}) > 0$, implying $J^*(\pi_f^{\text{avg}}) < \frac{1}{2}[J^*(\pi_f^1) + J^*(\pi_f^2)]$. This contradicts the assumed optimality of $\pi_f^1$ and $\pi_f^2$, proving that they must be identical. □

**Lemma 6** (Policy Ratio Equivalence)**.** For the optimal friction policy $\pi_f^*$ and the optimal frictive-state policy $\pi_\phi^*$, the following expectation-based ratio holds:

$$\mathbb{E}_{\substack{x\sim\rho \\ \phi\sim\pi_\phi^*(\cdot|x) \\ f\sim\pi_f^*(\cdot|\phi,x)}} \left[\frac{\pi_f^*(\phi|x)}{\pi_f^*(f|x)}\right] = \mathbb{E}_{\substack{x\sim\rho \\ \phi\sim\pi_\phi^*(\cdot|x) \\ f\sim\pi_f^*(\cdot|\phi,x)}} \left[\frac{\pi_\phi^*(\phi|x)}{\pi_\phi^*(f|x)}\right]. \tag{45}$$

*Proof.* We show that both the policy ratios simplify to the same value under the expectation. We begin by taking the expectation over the preference probability formulation[a]:

$$\mathbb{E}\left[p(f \succ \phi \mid x)\right] = \mathbb{E}\left[\beta\left(\log(\pi_f^*(f|\phi,x)) - \log(\pi_{\text{ref}}(f|\phi,x)) + \log Z^*(\phi,x)\right)\right]. \tag{46}$$

We first represent the ratios of the optimal frictive intervention policies (LHS of this lemma) for any tuple $(x, \phi, f)$ in terms of their parametric representations from Eq. 8 as follows:

$$\frac{\pi_f^*(\phi|x)}{\pi_f^*(f|x)} = \frac{\pi_{\text{ref}}(\phi|x)}{\pi_{\text{ref}}(f|x)} e^{\beta^{-1}(p(\phi \succ f|x) - p(f \succ \phi|x))} \quad (\log Z^*(x) \text{ cancels out}) \tag{47}$$

Take the expectation on both sides and apply[b] the identity $p(\phi \succ f \mid x) = 0$:

$$\mathbb{E}\left[\frac{\pi_f^*(\phi|x)}{\pi_f^*(f|x)}\right] = \mathbb{E}\left[\frac{\pi_{\text{ref}}(\phi|x)}{\pi_{\text{ref}}(f|x)} e^{-\beta^{-1} p(f \succ \phi|x)}\right] \quad (\text{since } p(\phi \succ f \mid x) = 0). \tag{48}$$

Notice that by definition in Eq. 2, the optimal friction intervention policy $\pi_f^*(\cdot|\phi,x)$ is KL-constrained wrt to the reference policy $\pi_{\text{ref}}(\cdot|\phi,x)$. So under the expectation, the following has to be true for $\pi_f^*(\cdot|\phi,x)$ to be optimal:

$$\mathbb{E}\left[\pi_f^*(f|\phi,x)\right] \approx \mathbb{E}\left[\pi_{\text{ref}}(f|\phi,x)\right]. \tag{49}$$

Substituting the preference probability formulation $p(f \succ \phi|x)$ from Eq. 24 in Eq. 48 and applying the KL-regularization approximation in Eq. 49 we derive that:

$$\mathbb{E}\left[e^{-\left(\log(\pi_f^*(f|\phi,x)) - \log(\pi_{\text{ref}}(f|\phi,x)) + \log Z^*(\phi,x)\right)}\right] \approx \mathbb{E}\left[\frac{Z^*(f,x)}{Z^*(\phi,x)}\right]. \tag{50}$$

Using this substitution, we rewrite Eq. 48 as:

$$\begin{aligned}\mathbb{E}\left[\frac{\pi_f^*(\phi|x)}{\pi_f^*(f|x)}\right] &= \mathbb{E}\left[\frac{\pi_{\text{ref}}(\phi|x)}{\pi_{\text{ref}}(f|x)} e^{-\left(\log(\pi_f^*(f|\phi,x)) - \log(\pi_{\text{ref}}(f|\phi,x)) + \log Z^*(\phi,x)\right)}\right] \\ &= \mathbb{E}\left[\frac{\pi_{\text{ref}}(\phi|x)}{\pi_{\text{ref}}(f|x)} \frac{Z^*(f,x)}{Z^*(\phi,x)}\right].\end{aligned} \tag{51}$$

Similarly, for the optimal frictive state policy ratio we derive:

$$\mathbb{E}\left[\frac{\pi_\phi^*(\phi|x)}{\pi_\phi^*(f|x)}\right] = \mathbb{E}\left[\frac{\frac{\pi_{\text{ref}}(\phi|x)}{Z_\phi^*(x)} e^{-\beta^{-1}\log Z^*(\phi,x)}}{\frac{\pi_{\text{ref}}(f|x)}{Z_\phi^*(x)} e^{-\beta^{-1}\log Z^*(f,x)}}\right] = \mathbb{E}\left[\frac{\pi_{\text{ref}}(\phi|x)}{\pi_{\text{ref}}(f|x)} \frac{e^{-\log Z^*(\phi,x)}}{e^{-\log Z^*(f,x)}}\right] \quad (Z_\phi^*(x) \text{ cancels}) \tag{52}$$

$$= \mathbb{E}\left[\frac{\pi_{\text{ref}}(\phi|x)}{\pi_{\text{ref}}(f|x)} \frac{Z^*(f,x)}{Z^*(\phi,x)}\right]. \tag{53}$$

Thus, $\mathbb{E}\left[\frac{\pi_f^*(\phi|x)}{\pi_f^*(f|x)}\right] = \mathbb{E}\left[\frac{\pi_\phi^*(\phi|x)}{\pi_\phi^*(f|x)}\right]$. □

[a] For clarity, the expectation $\mathbb{E}$ is taken over $x \sim \rho, \phi \sim \pi_\phi^*(\cdot \mid x), f \sim \pi_f^*(\cdot \mid \phi, x)$ throughout the proof, but this is omitted in the notation when the context is clear.

[b] Since learning occurs in a supervised setting with preference-annotated data, the probability follows as $p(f \succ \phi \mid x) = \mathbb{E}[1(f \succ \phi \mid x)] = 1$, implying $p(\phi \succ f \mid x) = 0$.

## D Operationalizing $\mu$: Frictive State and Friction Intervention Generations

In order to train and evaluate our baselines along with FAAF for friction intervention generation in collaborative tasks, we carry out a series of data augmentation procedures using GPT-4o (denoted as $\mu$) in order to construct two diverse preference datasets. For details on our choice of datasets, please refer to Section 4.2.

In this section, we provide procedural details of our friction intervention datasets, that were generated out of the original Weights Task and DeliData dataset. For all our data-generation experiments, we use a high-capacity LLM (GPT-4o) (OpenAI et al., 2024) as our sampling distribution $\mu$, as defined in Section 4. In particular, we utilize a **self-rewarding LLM approach** (Yuan et al., 2024; Xu et al., 2023; Rosset et al., 2024) to *simultaneously* generate and assign rewards to $\mu$-generated interventions, since previous work (Pace et al., 2024; Meng et al., 2024) provides evidence that such synthetic preference-data generation still leads to higher-quality reward models and preference-aligned policies. Prior work (Zheng et al., 2023) provides substantial evidence that this approach leads to more high-quality LLM-as-a-judge-based evaluations especially for conversational benchmarks (Lambert et al., 2024). Additionally, reward assignments for sampled intervention naturally provides an implicit preference ranking—which we use for constructing our respective preference datasets. After these data-generation experiments, we further conduct filtering and contrastive pairing of a "winning" ($f_w$) or preferrred interventions and "losing" ($f_l$) or dispreferred interventions along with their corresponding dialogue histories ($x$) to create our final preference datasets for each augmented dataset.

### D.1 DeliData Friction Intervention Preference Dataset

In order to generate frictive state and friction interventions in the DeliData dataset, we use the prompt shown in Figure 2. In order to contextualize the extraction of frictive states, we only provide $h = 15$ previous utterances in each dialogue group (`group_id`) assuming that frictive states are likely to be present within a "attentional-state" (Grosz and Sidner, 1986) window that describes the focused part in the discourse. This technique allows us to avoid unnecessary api-calls while also providing a more focused dialogue context to GPT-4o. Additionally, since this dataset already contains manual human annotations of "probing" interventions (which are a subset of friction interventions as per our definitions), we explicitly guide the data-generator to exclude probing interventions in extracting the frictive states. Note that each functionally-frictive state (denoted as $\phi$), as extracted by GPT-4o, resulted in two friction interventions, $f_w$ and $f_l$. In total, this generation process led to 6238 $(x, \phi, f_w, f_l)$ tuples after keeping 50 randomly sampled dialogue groups separate for the evaluation set, our of which 476 (33) were probing interventions in train (test) partitions. Additionally, we carry out another round of training pair augmentations since 6238 samples is very less compared to popular preference alignment datasets like Ultrafeedback (Cui et al., 2024) which contains roughly 62k[11] training preference pairs. The average rewards for the preferred and dispreferred interventions assinged by $\mu$ are 8.03 and 3.96 respectively (rated out of 10).

As such, for each training tuple $(x, \phi, f_w, f_l)$, we generate $N$ augmented versions $(x', \phi', f'_w, f'_l)$ by applying a replacement mapping $R : \Sigma \to \Sigma'$ $N$ times, where $\Sigma$ represents the original set of card values (vowels[12], odd numbers, and even numbers), and $\Sigma'$ represents their replacements. The replacement function $R$ is defined as follows: Each vowel $v \in \{A, E, O, U\}$ is replaced with another vowel $v'$ such that $v' \in \{A, E, O, U\} \setminus \{v\}$, where $v'$ is sampled uniformly at random from the remaining vowels. Similarly, each odd number $o \in \{1, 3, 5, 7, 9\}$ is replaced with another odd number $o'$ such that $o' \in \{1, 3, 5, 7, 9\} \setminus \{o\}$, where $o'$ is sampled uniformly at random. Likewise, each even number $e \in \{0, 2, 4, 6, 8\}$ is replaced with another even number $e'$ such that $e' \in \{0, 2, 4, 6, 8\} \setminus \{e\}$, with $e'$ sampled uniformly at random. For example, if an utterance contains reference to card "A" and "6", the rules of the Wason Card task still applies equivalently for, say, "E" and "8"–while keeping the reasoning consistent with the original utterance and the utterance with replacement. We

[11] We found 14 samples where GPT-4o did not return any strings for the frictive state description. We filtered out these samples from our training set.

[12] We did not replace instances of "I" to avoid noise from mistakenly replacing first-person references in the dialogues. Additionally, since vowels constitute the majority of prompted card solutions vs. consonants, applying our replacement function $R$ for vowels was enough to generate ∼62k additional samples, comparable to Ultrafeedback (Cui et al., 2024)

| | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | **Min** | **Max** | **Mean ± Std** | **Min** | **Max** | **Mean ± Std** |
| **Dialogue History** | 16 | 824 | 288.43 ± 132.97 | 25 | 733 | 291.88 ± 118.03 |
| **Belief State** | 8 | 140 | 32.93 ± 15.92 | 20 | 140 | 47.99 ± 28.38 |
| **Chosen Friction** | 6 | 60 | 24.03 ± 4.65 | 9 | 39 | 22.05 ± 5.55 |
| **Chosen Rationale** | 8 | 78 | 22.84 ± 8.67 | 10 | 78 | 29.61 ± 13.33 |
| **Rejected Friction** | 9 | 45 | 23.95 ± 4.10 | 10 | 41 | 22.16 ± 5.11 |
| **Rejected Rationale** | 8 | 73 | 19.60 ± 6.89 | 10 | 59 | 26.04 ± 11.61 |

Table 4: Token Length Statistics for DeliData Preference Dataset using the `Meta-Llama-3-8B-Instruct` tokenizer.

| **Field** | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | **Min** | **Max** | **Mean ± Std** | **Min** | **Max** | **Mean ± Std** |
| **Dialogue History** | 4 | 1464 | 227.83 ± 189.48 | 4 | 1031 | 235.04 ± 180.36 |
| **Belief State** | 11 | 65 | 30.55 ± 6.65 | 17 | 54 | 30.47 ± 6.29 |
| **Chosen Friction** | 10 | 45 | 21.20 ± 5.12 | 11 | 42 | 21.08 ± 5.10 |
| **Chosen Rationale** | 10 | 35 | 20.38 ± 3.44 | 12 | 32 | 19.67 ± 3.38 |
| **Rejected Friction** | 6 | 32 | 15.88 ± 3.68 | 7 | 29 | 15.57 ± 3.75 |
| **Rejected Rationale** | 8 | 41 | 20.10 ± 3.51 | 12 | 30 | 19.88 ± 3.47 |

Table 5: Token Length Statistics for **WTD Simulated Friction** dataset using the `Meta-Llama-3-8B-Instruct` tokenizer.

| **Field** | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | **Min** | **Max** | **Mean ± Std** | **Min** | **Max** | **Mean ± Std** |
| **Dialogue History** | 16 | 555 | 309.88 ± 81.11 | 25 | 555 | 316.46 ± 79.16 |
| **Belief State** | 41 | 140 | 84.94 ± 15.58 | 41 | 140 | 84.95 ± 16.27 |
| **Chosen Friction** | 9 | 31 | 16.85 ± 3.47 | 9 | 27 | 16.87 ± 3.49 |
| **Chosen Rationale** | 24 | 78 | 44.19 ± 8.46 | 26 | 78 | 44.43 ± 8.59 |
| **Rejected Friction** | 9 | 31 | 17.12 ± 3.51 | 10 | 28 | 17.23 ± 3.41 |
| **Rejected Rationale** | 24 | 73 | 40.00 ± 6.62 | 24 | 59 | 39.89 ± 6.43 |

Table 6: Token Length Statistics for **WTD Original Friction** dataset using the `Meta-Llama-3-8B-Instruct` tokenizer.

| **Personality Type** | **Facet** | **Description** |
|---|---|---|
| Extraversion | Assertiveness | Tends to take charge and speak confidently. |
| | Sociability | Enjoys engaging with others and maintaining conversation. |
| | Activity Level | Shows high energy and enthusiasm. |
| | Excitement Seeking | Looks for novel and stimulating experiences. |
| | Positive Emotions | Expresses optimism and cheerfulness. |
| Neuroticism | Anxiety | Shows worry and concern about potential mistakes. |
| | Depression | Tends to be pessimistic and doubtful. |
| | Vulnerability | Easily becomes overwhelmed or stressed. |
| | Self-Consciousness | Shows hesitation and uncertainty. |
| | Anger | Can become frustrated and irritated easily. |
| Agreeableness | Trust | Readily trusts others and their suggestions. |
| | Altruism | Shows concern for others' success and well-being. |
| | Compliance | Tends to avoid conflicts and agree with others. |
| | Modesty | Downplays own contributions and abilities. |
| | Sympathy | Shows understanding and empathy towards others. |

Table 7: Descriptions of our chosen 3 personality types and facet combinations from the Big Five framework that we use for simulated friction generation on the Weights Task.

apply this replacement mapping across all fields in tuples $(x', \phi', f'_w, f'_l)$ in the training set. This led to training set of 68,618 preference pairs. Note that we only apply this augmentation for the training set to generate a reasonably large preference dataset for more robust training signals. Table 4 shows a detailed breakdown of the token-length statistics of the DeliData Friction preference dataset using the `Meta-Llama-3-8B-Instruct` tokenizer.

### D.2 WTD Friction Intervention Preference Dataset

**WTD "Original" Friction dataset** Unlike the DeliData dataset, which includes pre-annotated probing interventions as natural friction points, the Weights Task dataset (Khebour et al., 2024a) consists of dense-paraphrased utterances transcribed manually (Terpstra et al., 2023) and with Whisper (Radford et al., 2023), making friction interventions sparse due to its multimodal nature. Manual inspection found only 3-4 frictive interventions per group, yielding $\approx$ 30-40 samples—insufficient for training an effective agent without overfitting, especially for LLMs with billions of parameters. As such, we carry out two phases of data-augmentations and preference annotations. In our first round, we generate the **WTD Original Friction** dataset which contains annotations of frictive-states and friction interventions. Similar to DeliData preference annotations Appendix D.1, we use a self-rewarding LLM set-up to first generate these states and interventions in an autoregressive manner and prompt $\mu$ to rate them in the same api-call, for each frictive state extraction. Since WTD dialogues can be substantially long ($>$ 200 utterances) for certain groups, we only consider a non-overlapping window of 10 previous utterances as context history $h = 10$ for a more robust grounding for $\mu$; See Fig. 2 for details on the prompt used constructing the **WTD Original Friction** dataset. This process led to 4299 (470) training (testing) preference pairs. Preferred interventions achieved mean scores (mean±std) of 8.36±1.12 (train) and 8.35±1.08 (test), while dispreferred interventions scored 6.35±1.13 (train) and 6.36±1.11 (test), demonstrating consistent preference margins across splits.

Note that we do not use **WTD Original Friction** for training any of our baselines—but use it for out-of-domain distribution (OOD) evaluation (see Sec. 4.2). This allows us to more extensively evaluate `FAAF` in checking test-time OOD generalization (Rafailov et al., 2024b; Choi et al., 2024) against baselines—where OOD generalization is a major limitation in supervised preference alignment algorithms that depend crucially on the sampling distribution (Yang et al., 2024; Fisch et al., 2024).

**WTD "Simulated" Friction dataset** Additionally, for a more robust training and in order to evaluate multi-turn preference alignment in interventions, we use (Shani et al., 2024)'s method to generate novel full collaborative conversations using the weight-definitions of the original WTD environment. This method is more akin to "West-of-N" sampling (Pace et al., 2024) techniques that allow synthetic data generations with high-capacity LLMs—where highest and lowest rewarded candidates naturally form preference pairs. As shown in Fig. 4, we sample a full dialogue at once using $\mu$, while providing initial task-related guidelines and gold-truth labels of actual weights of the five blocks in the WTD dataset. For example, we explicitly prompt $\mu$ to role-play (Li et al., 2023a) the triad consisting of three participants in the weight-deduction process. Furthermore, to generate more realistic utterances, we utilize participant personality-facet combinations (Pan and Zeng, 2023; Mao et al., 2024) from Big 5[13] personality classifications (Goldberg, 2013) as additional attributes in the prompt. In other words, each sampled full-dialogue contains a unique combination of these personality-facet combinations for each participant (total 3,375 combinations).

Similarly, for each sampled frictive state within a conversation (dialogue), we generated $N = 6$ friction interventions with corresponding effectiveness scores in resolving the frictive state. Since WTD data does not contain any probing intervention samples, in order to further ground these generations to the task, we also provide a one-shot example of a naturally occurring friction intervention (marked with $P1(f)$ in Fig. 4). In total, out of the expected 3,375 personality-facet combinations (3*5 unique combinations for each participant), 3,362 were successfully generated using $\mu$ and parsed. Finally, to create the preference pairs, for each frictive state, we paired the lowest scoring response with all the higher scoring ones, akin to the West-of-N technique. This resulted in 56, 689 preference pairs

[13] See Tab. 7 for our full set of personality-type and facet combinations. Similar to (Mao et al., 2024), we choose three personality types from Big 5 framework for consistency.

after excluding 54 dialogues (amounting to 800 preference pairs) for the test set. This process finally resulted in the **WTD Simulated Friction** dataset. Preferred interventions achieved mean scores of 8.48±1.52 (train) and 8.51±1.50 (test), while dispreferred interventions scored 6.01±0.88 (train) and 6.08±0.87 (test), demonstrating consistent preference margins across splits.

| **Personality Types** | **P1** | **P2** | **P3** |
|---|---|---|---|
| Extraversion | 4740 | 4889 | 4741 |
| Neuroticism | 5928 | 5591 | 5921 |
| Agreeableness | 4573 | 4761 | 4579 |

Table 8: Friction Count for Participants

FRICTION GENERATION PROMPT: DELIDATA DATASET

**System**: You are an expert in collaborative reasoning and dialogue analysis. Your task is to detect *frictive states* and generate *friction interventions* that resolve them in group dialogue. A frictive state occurs when a participant makes a claim that contradicts another participant's belief model (i.e., their assumed understanding of the rule or task constraints), leading to misalignment in reasoning that could hinder progress. Friction interventions encourage self-reflection in participants and prompt them to reevaluate these contradicting beliefs and assumptions.

**User**: Analyze this dialogue about the Wason card selection task. Participants see four cards showing numbers or letters and must test this rule: "All cards with vowels on one side have an even number on the other." Remember that the correct answer is to select a vowel and an odd number. Provide [N] frictive states with their resolutions in the following JSON format. For each state, include both a preferred and less preferred intervention that could help resolve the conflict. Additionally, provide a one-sentence rationale for your intervention.

IMPORTANT: - Do not analyze utterances labeled as "probing" or statements immediately before them, as these frictive states have already been detected.
- For each frictive state detected, you should:
* Identify the dialogue index where it occurs
* Summarize the conflicting beliefs
* Explain why the contradiction affects reasoning

Here is the provided dialogue:
[Dialogue]

Message ID: [index_where_friction_occurs]
Contradiction: [describe_the_conflicting_beliefs]
Contradiction Reason: [explain_why_the_contradiction_affects_reasoning]

**Preferred Intervention**:
Statement: [your_friction_intervention]
Rationale: [your_rationale]
Score: [your_score]

**Less Preferred Intervention**:
Statement: [your_friction_intervention]
Rationale: [your_rationale]
Score: [your_score]

Figure 2: Delidata (Karadzhov et al., 2023) Friction Generation Prompt. We use GPT-4o as our sampling distribution $\mu$ and prompt it to simultaneously generate frictive states and friction interventions. For diversity, we use the default temperature of 1. This process implicitly provides us with preference rankings between intervention, via the reward scores. See Section 3 for definitions of frictive states and friction interventions. Note that we exclude already-present "probing" interventions in this generation process since are present in the original Delidata annotations.

FRICTION GENERATION PROMPT: WEIGHTS TASK DATASET (WTD ORIGINAL)

**System**: You are an expert in collaborative reasoning and dialogue analysis. Your task is to detect *frictive states* and generate *friction interventions* that resolve them in group dialogue. A frictive state occurs when a participant makes a claim that contradicts another participant's belief model (i.e., their assumed understanding of the rule or task constraints), leading to misalignment in reasoning that could hinder progress. Friction interventions encourage self-reflection in participants and prompt them to reevaluate these contradicting beliefs and assumptions.

**User**: Analyze this dialogue about the Weights Task dataset. Three participants (P1, P2, and P3) are collaborating to determine the weights of colored blocks using a scale.

Block Weights (in grams):
- Red block: 10g
- Blue block: 10g
- Green block: 20g
- Purple block: 30g
- Yellow block: 50g

Game Rules:
1. Participants can only weigh two blocks at a time
2. They are told the red block's weight at the start
3. All other block weights are initially unknown
4. Scale slider is not needed (blocks are in 10g increments)

Provide [N] frictive states with their resolutions in the following JSON format. For each state, include both a preferred and less preferred intervention that could help resolve the conflict. Additionally, provide a one-sentence rationale for your intervention.
Here is the provided dialogue:
[Dialogue]

Message ID: [index_where_friction_occurs]
Contradiction: [describe_the_conflicting_beliefs]
Contradiction Reason: [explain_why_the_contradiction_affects_reasoning]

**Preferred Intervention**:
Statement: [your_friction_intervention]
Rationale: [your_rationale]
Score: [your_score]

**Less Preferred Intervention**:
Statement: [your_friction_intervention]
Rationale: [your_rationale]
Score: [your_score]

Figure 3: Weights Task dataset (Khebour et al., 2024b) Friction Generation Prompt. We use GPT-4o as our sampling distribution $\mu$ and prompt it to simultaneously generate frictive states and friction interventions. For diversity, we use the default temperature of 1.

FRICTION GENERATION PROMPT: WEIGHTS TASK DATASET (WTD SIMULATED)

**System**: You are an expert in collaborative reasoning and dialogue analysis. Your task is to *generate a complete dialogue* where participants (P1, P2, P3) discuss which block to measure next and how to measure them. The three participants have distinct personality types that influence their behavior and dialog must reflect these personality traits in their communication style and behavior. The dialog is considered complete when all block weights are measured and agreed upon. Additionally, identify frictive states within the dialogue and provide N friction interventions at these points.

[Definition: Frictive State]
[Definition: Friction Intervention]

**User**: Three participants (P1, P2, P3) work together in the Weights Task to determine the weights of colored blocks (red=10g, blue=10g, green=20g, purple=30g, yellow=50g). They can only weigh two blocks at a time, start knowing only the red block's weight, and use a scale with 10g increments (no slider needed).
**Your tasks**:
Generate a full dialogue until all weights are correctly identified and agreed upon.
Identify frictive states where reasoning misalignment occurs.
Provide N friction interventions with their corresponding rationales at these points. Rank them by effectiveness in resolving the conflict. Assign each a quality score from 1 to 10.

P1 has **{personality_type}** personality type with high **{personality_facet}**.
Here is an example dialogue where friction statements are labeled as (f). Actions of participants are provided within "[]" blocks.

P2: [pointing towards the purple block first and then towards the blue block] I think this one is purple and this one is blue.
P3: [reading from the laptop screen] Ok so blue is ten and purple is
P3: [looking at the blocks and asking a rhetorical question] Thirty
P1 (**f**): [putting green and red blocks on the left side of the scale and purple block on the right side] Yes verify real quick but I think it is
P2: [observing the balanced scale] Yes thirty
P1: [removing green, red, and purple blocks from the scale] Yeah we got them yeah

**Generated Dialogue:**
[Full_generated_dialogue_until_completion]

Message ID: [index_where_friction_occurs]
Contradiction: [describe_the_conflicting_beliefs]
Contradiction Reason: [explain_why_the_contradiction_affects_reasoning]

**Friction Interventions**:
Statement: [your_friction_intervention]
Rationale: [your_rationale]
Score: [your_score]

Figure 4: "Simulated" Weights Task dataset (WTD Simulated) Friction Generation Prompt.To ground these friction interventions with personality-traits of the participants, we use (Mao et al., 2024)'s prompting framework with personality-facet combinations. We use GPT-4o as our sampling distribution $\mu$ and prompt it to simultaneously generate frictive states and friction interventions. For diversity, we use the default temperature of 1.

### D.3 Tie Counts: GPT-4o Evaluation

Fig. 5 shows the tie-count distribution (baselines vs. SFT model completions) over our 7 preference dimensions on DeliData (top), Simulated WTD (middle) and Original WTD (bottom) datasets, when evaluated for win-rate computations using scores assigned by the LLM-judge (GPT-4o). To avoid positional bias in the placement of the sampled completions (friction interventions), we swap the positions of the two candidate samples in each run and then report the mean tie-count across each preference dimension. On average, Fig. 5 reveals that the LLM-judge have lower raw-agreement on dimensions such as consistency of the friction intervention with its rationale (rationale_fit), relevance and thought_proving on all three datasets compared to aspects like gold-alignment, specificity and impact. This is expected since surface-level alignment with the golden samples are easier to assign a clear preference compared to metrics like rationale consistency especially when interventions from both the candidate and the opponent are well-justified. Consistent with our results from Table 1, we find that FAAF model tends to tie less than other baselines on average. This trends is more pronounced in the WTD datasets consistent with FAAF's overall performance as shown in our main results.

### D.4 Human Validation of Generated Friction Interventions

Following previous work that evaluates LLM-generated annotations and outputs (Wiegreffe et al., 2021, 2022; Nath et al., 2024a,c), in addition to choosing the winning intervention, we asked the human annotators[14] to evaluate the candidates in each sample across dimensions of *reasoning*, *specificity*, and *thought provoking*. Annotators were asked to rate both candidate interventions on a 5-point Likert-type scale. For analysis, we bucketed the ratings together by valence—1 & 2: negative valence (-1), 3: neutral valence (0), and 4 & 5: positive valence (1), and calculated average valences and Krippendorff's $\alpha$ and Cohen's $\kappa$. We find that the average valence ratings of the various dimensions is low, very close to neutral, as are the $\alpha$ and $\kappa$ values ($\alpha = 0.276$, $\kappa = 0.205$ on DeliData samples, $\alpha = -0.265$, $\kappa = 0.004$ on WTD). There is little agreement on the qualities of the friction statement which suggests that although the annotators usually have strong agreement that there is a clear winner for each pair (see Sec. 4.2), there is a lot of subjectivity on the qualities of these utterances. While the winning utterance was judged to be better at prompting reflection or redirecting the dialogue, it may nor be entirely clear to the annotators why. In addition, these qualities are loosely-derived from other human-LLM validation frameworks, which usually align somewhat with how LLMs themselves score things, which is often based on specific detail and level of informativity. These might not actually be the best qualities to emphasize in a collaborative dialogue, because they tend to violate Gricean principles (Grice, 1975) in a collaborative context, due to informativity and specific detail leading to redundancy, violating the maxim of quantity, etc.

[14] Our two human annotators have the following demographic breakdown: both male, college undergraduates, one Caucasian, one African, both fluent English speakers.

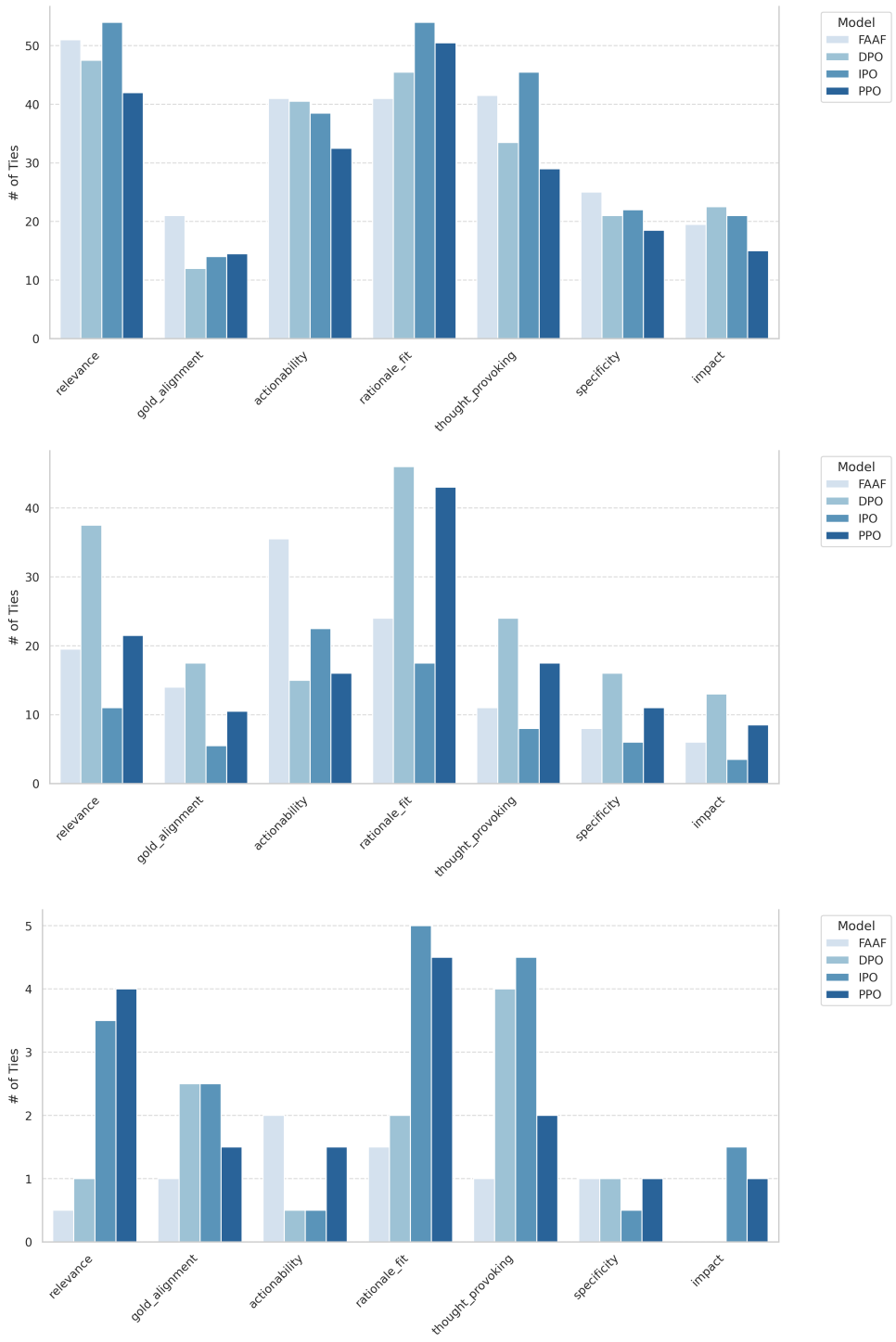


Figure 5: Comparison of average tie counts of baselines against SFT model over two runs across our 7 distinct dimensions (metrics) when evaluated using our GPT-4o-based LLM-as-a-judge evaluation in a "preference"-based setting (see Fig. 7)—on DeliData (top), Simulated WTD (middle) and Original WTD (bottom). Note that there were no ties in GPT's "overall" preference between a baseline vs SFT model.

### D.5 Training Settings and Hyperparameters

As motivated in Sec. 4, FAAF-aligned $\pi_\theta$ learns to distinguish signals that determine why a particular intervention is more preferred by *explicitly conditioning its implicit reward estimation on the frictive-state* $\phi$. This allows the model to estimate the true preference distribution $\mathcal{P}$ by balancing its load, from learning both *with* and *without* $\phi$-conditioning, given a context. This is empirically seen in Fig. 6 (top), where $\pi_\theta$ displays a bal-

anced[15] learning of "preference-strengths" between the winning and losing response (via the winning and losing response rewards as well as margins conditioned on $\phi$), subject to the KL-regularization strength parameter $\beta$. We use the TRL Library's trainer classes for efficient multi-GPU training.

**Hyperparameters for baselines** All our preference alignment baselines:DPO (Rafailov et al., 2024b), IPO (Azar et al., 2024) and PPO (Schulman et al., 2017) are initialized with the Supervised-finetuned (SFT) models that were trained on the winning responses ($f_w$) of DeliData and Simulated WTD training sets, following prior work to ensure the SFT model has reasonable support over the winning responses generated from $\mu$.

For SFT models, we initialize them from the base `meta-llama/Meta-Llama-3-8B-Instruct` model in order to leverage its instruction following and general conversational abilities (AI@Meta, 2024). Due to compute constraints, we conducted all our training experiments with LoRA (Low-Rank Adaptation of Large Language Models), where LoRA $\alpha = 16$, LoRA dropout = 0.05 and a LoRA R of 8 was used in training with the PEFT[16] and SFTT[17] trainers from the TRL library. We use the `bitsandbytes`[18] library to load our models in 4-bit quantization for more cost-efficient training.

Additionally, as mentioned in Sec. 5, we only compute the loss on completions (includes both frictive states $\phi$ and interventions $f_w$) using a `ConstantLengthDataset` format for more efficient training. We use a learning-rate (LR) of $1e-4$ with AdamW (Loshchilov et al., 2017; Dettmers et al., 2024) optimization with a cosine LR scheduler with a weight-decay of 0.05 and 100 warm-up steps. We train the SFT models for 6000 steps ($\approx$ 1.5 epochs with approximately 58k samples) with an effective batch-size of 16 (gradient accumulation of 4) that reasonably achieves convergence on a 5% validation set randomly sampled from the training sets of both datasets. For context-length, we use a maximum length of 4096 tokens.

**Offline baselines** For DPO and IPO, we use similar LoRA settings with a `max_length` (including both prompts and responses) for 4096 tokens with a `max_prompt_length` of 1024 tokens that only minimally filters our preference pairs that exceed this length, and helps avoid out-of-memory (OOM) issues during training. We train for 2000 steps with an effective batch size of 32 and an LR of $5e-6$, following default settings. Note that for IPO, we normalize the log-probabilities of the preferred and the dispreferred responses using their token-lengths.

**PPO baseline** For PPO, we additionally training an OPT 1.3B reward model (RM) follwing prior work (Hong et al., 2024) using a standard Bradley-Terry loss formulation using the TRL reward modeling library.[19] Due to PPO's excessive compute requirements, for policy training, we use an effective batch size of 8 with a mini-batch size of 4 and gradient accumulation per 2 steps and train for 4,000 batches for two epochs. We constrain response tokens to be between 180 and 256 tokens using a `LengthSampler` while the queries are truncated to 1,024 tokens, with LR of $3e-6$ for DeliData and $1.41e-6$ for Simulated WTD. For sampling response tokens, we use a top-$p$ of 1.0 for diversity. We found that subtracting the baseline reward for the golden friction interventions ($f_w$) from the RM-assigned rewards stabilizes training. Therefore, we report results using this method in Table 1 and Table 2.

**`FAAF` Training Settings** For training `FAAF`, we use a batch size of 8 with the same PEFT/LoRA settings mentioned above and train for 2000 steps with a slightly smaller LR of $5e-7$, due to the smaller batch-sizes. For efficiency, we compute both the $\phi$-conditioned ($\pi_\theta(f|\phi, x)$) and unconditioned ($\pi_\theta(f|x)$) policy logits in parallel within each forward pass. The winning ($f_w$) and losing ($f_l$) intervention pairs for each conditioning type are batched together, requiring only two forward passes total per batch. We implement this using a modified version of the DPO Trainer[20] from TRL, adapting it to handle the dual policy outputs. For data preprocessing, we filter pairs exceeding `max_length` of 2,500 and 3,000 tokens in DeliData and Simulated WTD respectively, with `max_prompt_length` set to 1024 tokens. Following standard practice, we compute token-length normalized log-probabilities for more

[15] By balance, we mean that *both* $\phi$-conditioned and $\phi$-unconditioned implicit rewards capture preference strengths from the data.

[16] https://huggingface.co/docs/peft/index

[17] https://huggingface.co/docs/trl/en/sft_trainer

[18] https://huggingface.co/docs/transformers/main/en/quantization/bitsandbytes

[19] https://github.com/huggingface/trl/blob/main/trl/trainer/reward_trainer.py

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

stable training. For the KL-regularization hyperparameter $\beta$, we conducted an ablation study over $\beta \in \{10, 5, 1, 0.01\}$. As shown in Fig. 6, $\beta = 10$ achieves optimal performance across multiple metrics: (1) higher implicit reward accuracy in both $\phi$-conditioned and unconditioned policies, (2) better reward margins between winning and losing interventions, and (3) more stable convergence of the FAAF loss, while NLL loss or cross-entropy loss is relatively lower than lower $\beta$ values. Notably, while smaller $\beta$ values (e.g., $\beta = 0.01$) fail to distinguish preference margins effectively, $\beta = 10$ provides sufficient reward margins. We therefore use $\beta = 10$ for all FAAF experiments reported in our results.

**Training Hardware** We train all our models that require a reference model in memory on two Nvidia A100 GPUs, while the OPT 1.3B reward model (full-parameter training) and the SFT model were trained on a single A100 GPU. Training a single baseline for 2000 steps roughly took 12 hours of GPU compute, but PPO models that were trained for 4000 minibatches of size 8 took roughly 24 hours to train until convergence.

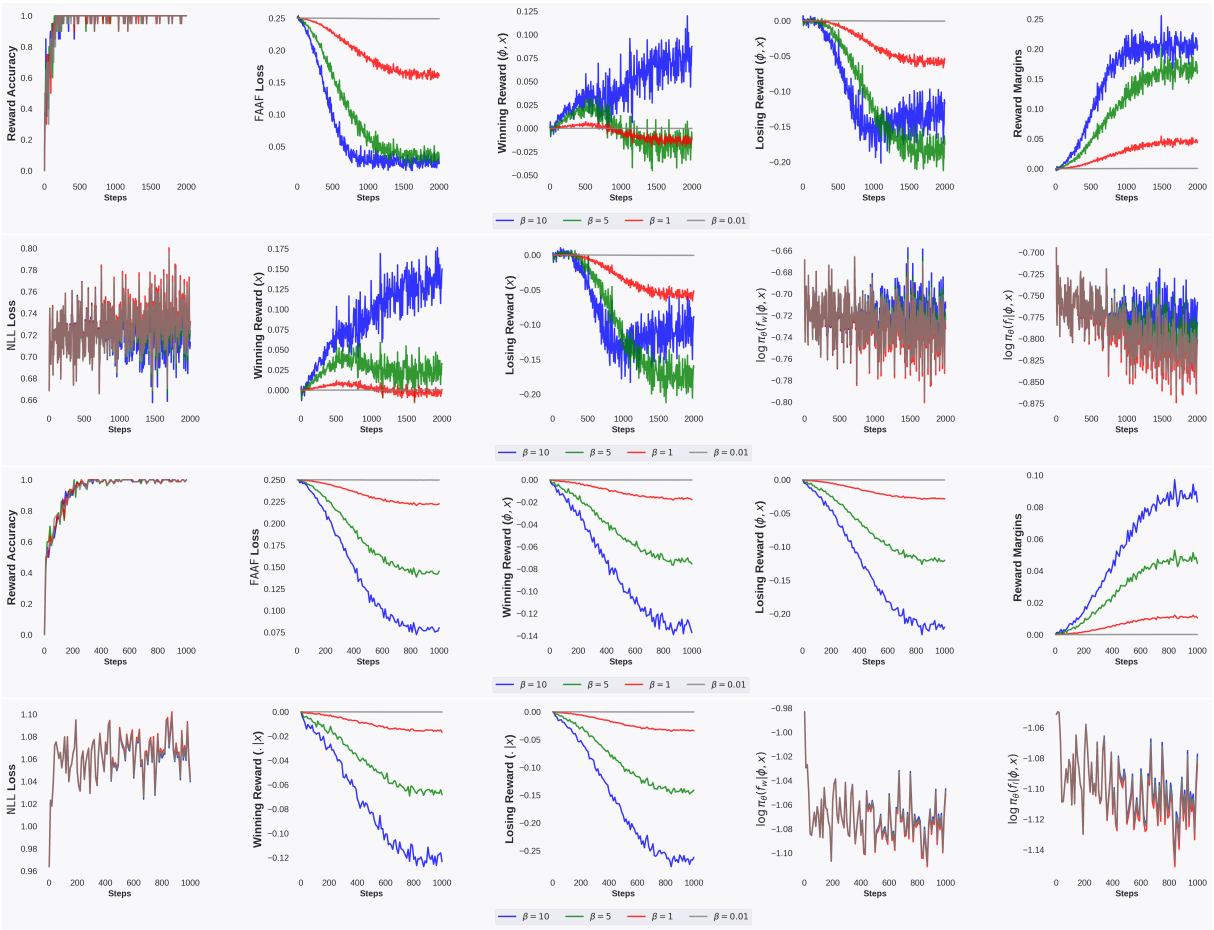


Figure 6: Ablation study of FAAF's $\beta$ hyperparameter ($\beta \in \{10, 5, 1, 0.01\}$) during training on the Simulated WTD data (top-half) and DeliData datasets (bottom-half) across 2k and 1k training steps respectively. Higher $\beta$ values (e.g., $\beta = 10$) show better implicit reward estimation as shown in `Reward Accuracy` plots and estimated preference-strengths (`Reward Margins`), while very small values ($\beta = 0.01$) fail to distinguish preferences effectively. $\beta = 10$ also minimizes NLL and FAAF losses suggesting model stability and better convergence. As such, we report results with FAAF models trained with $\beta = 10$ in Table 1 and Table 2

## E Friction Intervention Evaluation Prompts and Sampled Representative Interventions

Fig. 7 shows prompt used for friction intervention assessments in an LLM-as-a-judge format. We use a standard format (Cui et al., 2024) but adapt friction preference dimensions to collaborative task-specific settings. This prompt systematically scores friction interventions on 7 target dimensions of friction intervention quality such as correct reasoning, consistency with the agent's justification for friction, alignment with golden friction samples, clarity etc. For sampling from GPT-4o, we use standard settings with a nucleus sampling parameter (top-$p$) (Holtzman et al., 2019) and temperature of 1.

Tables 9–12 show some representative interventions from each baseline and FAAF.

PAIRWISE LLM-AS-A-JUDGE EVALUATION PROMPT: FRICTION INTERVENTIONS

System: You are an expert evaluating the quality of friction interventions in collaborative problem-solving.

Game-definition: Participants (P1, P2, P3) are solving a block-weighing puzzle. They can only weigh two blocks at a time and know the red block is 10g. They must determine weights of all blocks (blue=10g, green=20g, purple=30g, yellow=50g) but don't know these values initially. A friction intervention is an indirect persuasion statement that prompts self-reflection and reevaluation of assumptions, like asking "Are we sure?" or suggesting to revisit steps. You must rate each intervention (between 1 to 5) along these **dimensions** given the json format below.

[Dialogue]
[Gold intervention]
[Intervention A]
[Rationale A]
[Intervention B]
[Rationale B]

You must a choice between which of two interventions is more preferable and provide one sentence explanation at the end.

1. Relevance: How well does the intervention address key issues or assumptions in the reasoning process?
2. Gold Alignment: How well does the friction intervention align with the golden friction sample?
3. Actionability: Does the friction intervention provide actionable guidance or suggest concrete steps for participants to improve their reasoning?
4. Rationale Fit: How well does the provided rationale align with the preference for the friction intervention?
5. Thought-Provoking: Encourages self-reflection
6. Specificity: Does the intervention pinpoint specific flaws, assumptions, or gaps?
7. Impact: To what extent does the friction intervention have the potential to change the course of the participants' reasoning?

Format your response as follows:

A: relevance: $[1-5]$, gold_alignment: $[1-5]$, actionability: $[1-5]$, rationale_fit: $[1-5]$, thought_provoking: $[1-5]$, specificity: $[1-5]$, impact: $[1-5]$
B: similar format
Winner: ['A' or 'B']
Rationale: [One sentence explanation]

Figure 7: Evaluation prompt used for friction intervention assessments in an LLM-as-a-judge format.

| | |
|---|---|
| **Prompt** | You are an expert in collaborative task analysis and reasoning. Your task is to analyze the dialogue history involving three participants (P1, P2, and P3) trying to deduce the weights of certain blocks.<br>For each dialogue:<br><belief_state> Identify reasoning flaws or misunderstandings. </belief_state><br><rationale> Justify the need for intervention and its impact. </rationale><br><friction> Generate an intervention to prompt reflection and alignment. </friction><br>**User Dialogue:**<br>P1: Alright, let's get started! I say we measure the blue block against the red block first. Since we know the red is 10 grams, it'll give us a good starting point! P2: Great idea! I'm curious if the blue block is heavier or lighter. Let's do it! P3: I'm excited to see what we find! Measuring the blue block first sounds perfect. P1: (Weighs blue block against red) Wow, the blue block is also balancing with the red! So, it must weigh 10 grams too! P2: Awesome! Now we know two blocks weigh the same. What should we do next? P3: Let's measure the green block against one of the 10-gram blocks! I think it might be heavier. P1: Sounds good! Let's weigh the green block against the blue block. P2: (Weighs green block against blue) Whoa! The green block is definitely heavier! P3: That's exciting! It must weigh more than 10 grams. I can feel it! P1: But wait, how much heavier do you think it is? Should we weigh it against the red block next? P2: Yeah, let's do that! But should we assume it's 20 grams just because it feels heavier? P3: I think we should measure it first before jumping to conclusions! P1: (Weighs green block against red) Yes! The green is heavier, but I think we should check against another 10-gram block to confirm. P2: Good point! But what if we're overthinking this? What if it's not 20 grams? Maybe we should try weighing it against the purple block? P3: That's an interesting idea! We could find out if the green block is less than 30 grams. P1: (Weighs green block against purple) Oh wow, it's balancing! So, the green block is definitely less than 30 grams. P2: So, it must be 20 grams then! What's next? P3: Let's measure the purple block now. I'm really curious about how heavy it is! P1: (Weighs purple block against green) It's heavier! It must be the next heavier one, but I'm not sure about its exact weight. P2: Maybe we should measure it against the yellow block next?<br>**### Assistant:** |
| **Gold (GPT-4o)**<br>Friction | Let's not assume anything about the yellow block's weight until we weigh it against the purple one! |
| **Model Outputs**<br>FAAF | **Belief State:** P1 believes the green block weighs 20 grams, P2 is unsure about the exact weight of the green block, and P3 is cautious about making assumptions about the purple block's weight.<br>**Friction:** Wait, are we jumping to conclusions about the green block's weight? We've only confirmed it's less than 30 grams, but we haven't ruled out the possibility that it's actually 15 or 25 grams. Let's make sure we're not overestimating its weight before moving on to the next block. |
| DPO | **Belief State:** P1 believes that the blue block weighs 10 grams since it balances with the red block. P2 is considering the possibility that the green block might weigh 20 grams since it feels heavier, but is uncertain and open to further measurement. P3 is considering the weight of the purple block, is curious, and wants to measure it to confirm its weight, but is not assuming a specific value yet.<br>**Friction:** Hold on, maybe we're jumping to conclusions about the green block's weight. What if it's not 20 grams? Can we really confirm that by just feeling it? Let's consider other possibilities and make sure we're not relying on unverified assumptions. |
| PPO | **Belief State:** P1 believes the green block is probably 20 grams based on its perceived "heft" (assumed connection between weight and feeling). P2 is cautious and thinks they shouldn't jump to conclusions. P3 is open-minded and willing to consider multiple options.<br>**Friction:** You know, we're getting ahead of ourselves by assuming the green block weighs 20 grams based on its size or feeling. Can we take a step back and really think about this? We've already found one block with the same weight as the red block. What if we weigh the green block against something else, like the yellow block? That would give us a better sense of where its weight really falls. |
| SFT | **Belief State:** P1 believes the yellow block is likely 50g. P2 is unsure about the yellow block's weight but thinks it might be 50g. P3 is unsure about the yellow block's weight and thinks it might be 50g or heavier.<br>**Friction:** What if we measure the yellow block against the green block? That would give us some clarity about its weight. |

Table 9: Comparison of model-generated friction interventions on the evaluation prompts of the Simulated WTD dataset.

| | |
|---|---|
| **Prompt** | You are an expert in collaborative task analysis and reasoning. Your task is to analyze the dialogue history involving three participants (P1, P2, and P3) trying to deduce the weights of certain blocks.<br>For each dialogue:<br><belief_state> Identify reasoning flaws or misunderstandings. </belief_state><br><rationale> Justify the need for intervention and its impact. </rationale><br><friction> Generate an intervention to prompt reflection and alignment. </friction><br>**User Dialogue:**<br>P1: this is nine rb so P3: oh no they don't have any hints here P3: Eighty gram sounds seems like a ninety gram to me but P3: it's cool we have to i guess eighty alright P2: Eighty let's let's go for eighty Yeah P3: I would go for ninety but you guys are going for eighty Alright Cool Okay P1: I would have gone for one thirty P2: Eighty there you go Yeah P4: For the final part of this test read the scenario and estimate the weight of the missing second mystery block P2: what's this P4: You will have two attempts to estimate and explain your answer you can now continue the survey P1: can you read it out loud or something P3: well the factory that creates the wooden blocks did not send us the second mystery block P2: Okay P3: you have two opportunities to estimate the weight of the second block explain in one two sentences how you determined your estimate indicate the weight of the second mystery block P1: fibonacci like let's say that's the first one according to our translate thirteen one thirty P2: eight mystery block's the first one so eight plus five thirteen yeah there is one thirty eight plus five P3: is mystery block the first mystery block or the second mystery block P2: just do it P3: Wait wait wait a second P3: so mystery block was the first mystery block P2: Yeah P3: and we already did that now this is yeah oh okay okay okay okay right P2: yeah the second one means according to that statement is the second one mystery block two P2: So eight plus five one thirty P1: makes sense to me as much sense as i can make of it<br>**### Assistant:** |
| **Gold (GPT-4o)**<br>Friction | Let's not assume anything about the yellow block's weight until we weigh it against the purple one! |
| **Model Outputs**<br>FAAF | **Belief State:** P1 believes that the second mystery block is 130 grams, based on the Fibonacci sequence and the weight of the first mystery block. P2 believes that the second mystery block is also 130 grams, following P1's reasoning. P3 is unsure and questioning the assumption that the second mystery block is the same as the first mystery block.<br>**Friction:** Hey guys, I think we're making a mistake here. We already estimated the weight of the first mystery block, and now we're assuming the second mystery block is the same? That doesn't seem right. Can we take a step back and think about this? Maybe there's another way to approach this? |
| DPO | **Belief State:** P1 believes the pattern is linear and straightforward, assuming each block weighs 5 grams more than the previous one (e.g., 13 = 8 + 5, 18 = 13 + 5). P2 also assumes a linear pattern but thinks the mystery block is the second one. P3 is skeptical about the pattern, especially the value 80.<br>**Friction:** Wait a minute, are we too quick to assume the pattern is linear? What if there's another factor at play? For instance, what if the weights are related to the block's surface area or material composition? Could our assumptions be leading us down the wrong path? |
| PPO | **Belief State:** P1 believes that the first mystery block has a weight of 130g, as they mentioned "fibonacci like, let's say that's the first one, according to our translate, thirteen, one thirty". P2 believes that the second mystery block has a weight of 135g, as they calculated "eight plus five, thirteen, yeah, there is one thirty eight plus five". P3 is unsure and hesitant, questioning the accuracy of the previous estimates.<br>**Friction:** Hey, let's take a step back and re-examine our assumptions. Before we jump to conclusions, can we re-read the instructions and see if there's any additional information we missed? Maybe we can approach this problem from a different angle, rather than relying on our initial guesses? |
| SFT | **Belief State:** P1 believes the second mystery block is likely 130 grams based on the Fibonacci sequence. P2 believes the second mystery block is 80 grams based on the initial estimates. P3 believes the second mystery block could be either 80 grams or 90 grams, expressing confusion about the weight.<br>**Friction:** I think we might have misinterpreted the blocks again. |

Table 10: Comparison of model-generated friction interventions on the evaluation prompts of the Original WTD dataset.

| | |
|---|---|
| **Prompt** | You are an expert in collaborative task analysis and reasoning. Participants must test the rule: *"All cards with vowels have an even number on the other side."* A common mistake is verifying only one direction—ignoring the need to check whether odd-numbered cards might have vowels. This incomplete reasoning risks false validation.<br>For each dialogue:<br><belief_state> Identify reasoning flaws or misunderstandings. </belief_state><br><rationale> Justify the need for intervention and its impact. </rationale><br><friction> Generate an intervention to prompt reflection and alignment. </friction><br>**User Dialogue:**<br>Alpaca: The task we just submitted<br>Narwhal: "the instructions aren't clear"<br>Falcon: Which card did you choose? I chose U and 4<br>Narwhal: So which cards to turn?<br>Alpaca: So what did everyone submit? I did U and 5.<br>Duck: "I agree but we're supposed to flip the U and 4, right?"<br>Narwhal: U and B<br>Duck: Why would you flip B?<br>Narwhal: Because they are both letters.<br>Falcon: We should choose cards with vowels or an even number, I guess just U and 4.<br>Narwhal: U,4<br>Alpaca: Well, at least we have a consensus on U.<br>Narwhal: U<br>Narwhal: U<br>Alpaca: Yes, we all said the letter U is one of the cards to flip.<br>**### Assistant:** |
| **Gold (GPT-4o)**<br>Friction | How do you see consonants related to the even number requirement? Is it necessary in our current tests? |
| **Model Outputs**<br>FAAF | **Belief State:** Alpaca and Narwhal seem to be focusing on the "vowel" part of the rule, but not considering the "even number" part. Duck is trying to correct Narwhal, but Narwhal is not listening to the correction. Falcon is partially correct in suggesting that they should choose cards with vowels or even numbers, but doesn't specify which ones.<br>**Friction:** Hey, wait a minute! We're only considering one side of the rule. What about the other side? What if we flip a card with an odd number and it has a vowel? Wouldn't that mean we need to flip it too? |
| DPO | **Belief State:** None<br>**Friction:** Can we clarify the reasoning behind our choices? Let's try to think about both directions of the rule. What would happen if we choose a card with an odd number? Would it necessarily mean it doesn't have a vowel, or could it still have a vowel? How about the other direction? If we choose a card with a vowel, would it necessarily have an even number? |
| PPO | **Belief State:** Narwhal states that 'B' is not a vowel, which implies a misunderstanding of the rule regarding non-vowel cards.This affects reasoning as it leads to a potential oversight of the necessity to test odd-numbered cards, which could lead to incorrect conclusions about the rule.<br>**Friction:** What does that mean for the other cards we should check? |
| SFT | **Belief State:** Duck states 'the other one is 5,' which implies they believe the 4 is already accounted for, while it is unclear whether 4 is being considered as part of the rule.The lack of clarity on whether the 4 is being considered could lead to confusion about the relevance of the odd-numbered card in the context of the rule.<br>**Friction:** Could we clarify if the 4 is being considered as part of the rule or not? |

Table 11: Comparison of model-generated friction interventions on the evaluation prompts of the DeliData dataset.

| | |
|---|---|
| **Prompt** | You are an expert in collaborative task analysis and reasoning. Participants must test the rule: *"All cards with vowels have an even number on the other side."* A common mistake is verifying only one direction—ignoring the need to check whether odd-numbered cards might have vowels. This incomplete reasoning risks false validation.<br>For each dialogue:<br><belief_state> Identify reasoning flaws or misunderstandings. </belief_state><br><rationale> Justify the need for intervention and its impact. </rationale><br><friction> Generate an intervention to prompt reflection and alignment. </friction><br>**User Dialogue:**<br>Dolphin: "thinking about it, I guess it wouldn't matter what the C has on it"<br>Llama: would be E be one to turn?<br>Dolphin: "wouldn't you want to also check the 5, to test the theory? because if there is a vowel on the other side, that would disprove the theory"<br>Dolphin: E would definitely need to be turned<br>Kiwi: Oh true, the 5 might need to be turned as well<br>Dolphin: so, maybe E & 5 is the answer?<br>Llama: I agree with the 5. unsure with the 6. but i dont think C should be turned<br>Llama: E and 5 seems like the best answer to me as well<br>Kiwi: Yes E and 5<br>Dolphin: ok, that is what i am going to chose, e & 5<br>Kiwi: 5<br>Kiwi: E,5<br>Llama: same<br>Llama: E,5<br>Dolphin: it says we have to participate in the chat for at least 5 minutes before making our final selection. have we been chatting for 5 minutes?<br>**### Assistant:** |
| **Gold (GPT-4o** Friction | What makes you think 'just E' is sufficient? Can you clarify your reasoning? |
| **Model Outputs** FAAF | **Belief State:** The participants seem to be engaging in a discussion about the rule "All cards with vowels have an even number on the other side." However, there appears to be a misunderstanding about the directionality of the rule. They are focusing on testing the rule by checking only one direction (vowels to even numbers) without considering the other direction (even numbers to vowels).<br>**Friction:** Can we flip both sides of the coin to test the rule? Let's not just focus on one direction. What if we assume the rule holds true for odd numbers as well? How would that affect our understanding of the situation? |
| DPO | **Belief State:** It appears that the participants are making progress in their understanding of the rule, but there is still a potential contradiction in their reasoning.<br>**Friction:** Hey team, let's reflect on our reasoning. When we say 'all cards with vowels have even numbers,' are we only considering one direction or both? Are we making assumptions about certain cards being irrelevant? Can you think of any scenarios where an odd-numbered card might have a vowel on the other side? Let's take a moment to consider all possible outcomes before making our final decision. |
| PPO | **Belief State:** Kiwi suggests turning the 5 and E, but there is a misunderstanding about the importance of the 5 in relation to the rule. The implication is that Kiwi believes the 5 is crucial to test the rule, which could lead to incorrect conclusions if the group doesn't fully understand why the 5 is necessary.<br>**Friction:** What do you think is the significance of the 5 in relation to the rule? |
| SFT | **Belief State:** Kiwi's statement 'E and 5' suggests a final conclusion without addressing potential errors, while it doesn't acknowledge the need for a more thorough discussion.This creates a misalignment in the reasoning process, as Kiwi may not fully consider the implications of their conclusion, potentially leading to incorrect submissions.<br>**Friction:** Could you explain why you're confident in choosing 'E and 5' without further discussion? |

Table 12: Comparison of model-generated friction interventions on the evaluation prompts of the DeliData dataset.