Mind the Gap: Theory-of-Mind-Grounded Friction for Epistemic Alignment

Anonymous ACL submission

Abstract

Productive dialogue alignment requires distinguishing *surface coordination* (acknowledgments, smooth task progression) from *epistemic alignment* (convergence of belief states); standard preference-based methods optimize the former without modeling the latter. We operationalize Theory-of-Mind (ToM) inference as a control signal within Frictive Policy Optimization (FPO; Pustejovsky and Krishnaswamy, 2026) by extracting, at each referring expression, a four-part belief structure: each participant's intended referent and each participant's model of the other's interpretation. This makes friction mechanically computable from epistemic state comparisons, in particular capturing *silent divergence*, where both participants proceed confidently while grounding to different referents. Two experiments evaluate the signal: (i) ablating the second-order channel cuts misunderstanding detection from 65% to 26%, and (ii) under FPO trainer adaptation, the reward-shaping and trust-region variants significantly outperform a friction-blind preference baseline on task-grounded metrics and pairwise LLM-judge preference. ToM-grounded friction is thus a trainable signal for managing belief dynamics across dialogue, reframing alignment as epistemic stewardship rather than task completion.

1 Introduction

Recent work on epistemically asymmetric interaction has begun to expose a structural gap in how collaborative dialogue is modeled. In settings such as the Distributed Partial Information Puzzle (DPIP; Zhu et al., 2026), each participant observes only a partial view of the environment, requiring inference not only about what others know, but also about what they can perceive and misinfer. Such interactions reveal a problem that standard turn-level evaluation obscures: interlocutors may appear aligned through acknowledgments, uninterrupted task progression, or lexical accommodation while still grounding expressions in different referents or maintaining incompatible interpretations. We refer to this gap as the distinction between *surface-level coordination* (observable dialogue behavior) and *epistemic alignment* (convergence of participants' belief states, including beliefs about others' beliefs); alignment is therefore better viewed as ongoing regulation of epistemic states rather than as successful surface coordination.

Frictive Policy Optimization (FPO; Pustejovsky and Krishnaswamy, 2026) addresses this distinction by formalizing alignment as a risk-sensitive epistemic control process governing belief, commitment, and uncertainty over time, rather than as turn-level optimization against a static preference signal. However, existing instantiations primarily operate on surface-level or indirectly represented frictive states, limiting their ability to model the latent belief divergence that arises in genuinely asymmetric interaction.

Closing this gap requires an explicit account of the higher-order beliefs that produce silent divergence. Theory of Mind (ToM) (Premack and Woodruff, 1978; Wimmer and Perner, 1983; Sperber and Wilson, 1986; Baron-Cohen et al., 1985) formalizes nested reasoning over others' mental states; in asymmetric settings, agents may maintain incompatible first-order beliefs while mutually presuming agreement at the second-order level, making misalignment invisible in surface dialogue and unrecoverable from transcripts alone. We therefore propose an operational revision of FPO: friction must be computed over epistemic states grounded in each participant's information horizon rather than over surface dialogue. Concretely, we extend FPO with a point-in-time Theory-of-Mind schema that tracks, for each referring expression, a four-part structure: speaker and addressee beliefs about the referent, and each participant's model of the other's belief. Friction is then defined over discrep-

ancies in this structured epistemic configuration, rather than over dialogue acts.

We make three contributions. First, we introduce a point-in-time ToM extraction schema for HCRC MapTask (Anderson et al., 1991) that reconstructs first- and second-order beliefs at each referring expression, enabling mechanically computable friction signals validated against gold annotations. Second, we show via channel attribution that many misunderstandings, including silent intent fixing where interaction proceeds despite divergent grounding, are recoverable only from second-order belief structure, demonstrating that surface and first-order signals are insufficient. Third, we find that reward-shaping (FAR) and trust-region control (FTR) yield context-sensitive intervention behavior that preference-based methods (e.g., DPO) cannot capture, suggesting per-row friction signals as a viable mechanism for implementing FPO in asymmetric dialogue settings.

2 Background

2.1 LLM Alignment

Mainstream alignment methods — RLHF (Christiano et al., 2017), DPO (Rafailov et al., 2023), GRPO (Shao et al., 2024) — treat alignment as a turn-level optimization toward responses preferred by human raters or proxy reward models. The contract is consistent across variants: the policy is rewarded for matching a static preference signal computed from the immediate input. This contract fits single-turn instruction following but breaks down in collaborative interaction, where appropriateness depends on participants' evolving epistemic states rather than static surface preferences. To bring this epistemic dimension into the training objective, Frictive Policy Optimization (FPO; Pustejovsky and Krishnaswamy, 2025, 2026) reframes alignment as risk-sensitive epistemic control using a friction functional $F(h, y) = F^{+}(h, y) - F^{-}(h, y)$; prior instantiations have operationalized F from natural-language descriptions of frictive states (Nath et al., 2025) or simulated partial-information environments (Nath et al., 2026), but the question of how F should be computed directly from participants' underlying belief structure in naturalistic asymmetric dialogue remains open.

2.2 Grounding, Perspective, and Epistemic Alignment

Common ground is not a static prerequisite for dialogue, but a dynamically negotiated achievement established through grounding (Clark and Brennan, 1991; Clark, 1996). Grounding fails silently when participants proceed as if mutual understanding has been achieved despite holding divergent interpretations, and such failures are difficult to detect from surface dialogue alone. Work on automated common ground tracking (Khebour et al., 2024) further shows that surface transcripts alone are insufficient even for tracking shared beliefs in collaborative tasks, underscoring the need for richer epistemic representations.

Theory of Mind (ToM) (Premack and Woodruff, 1978; Wimmer and Perner, 1983; Sperber and Wilson, 1986) provides the formal distinction between first-order beliefs and beliefs about others' beliefs. This distinction makes silent divergence representable: interlocutors may maintain incompatible first-order beliefs while mutually presuming agreement at the second-order level, producing misalignment that remains invisible in the transcript.

Recent perspectivist annotation work has begun to address this limitation. Li et al. (2026) introduce a scheme for the HCRC MapTask corpus (Anderson et al., 1991) that separately tracks each participant's interpretation of referring expressions, exposing referential mismatches obscured by earlier grounding annotations. By shifting grounding from dialogue-level outcomes to participant-specific belief trajectories, this framework provides the basis for structured ToM extraction in naturalistic interaction.

3 Data

We construct FPO training instances from Li et al. (2026), which augments the HCRC MapTask corpus with perspectivist grounding annotations. In MapTask, a *giver* guides a *follower* through a navigation task using mismatched maps that neither participant can see. Giver/follower roles remain fixed across a dialogue, while *speaker* and *addressee* vary turn by turn depending on who produces a referring expression, which is important for assigning first- and second-order beliefs (§4.1). Their annotations categorize map discrepancies into three types: *existence* (landmark on only one map), *lexical* (same landmark, different labels), and *multiplicity* (duplicated on one map only), with multiplicity serving as the main source of silent misalignment.

For each dialogue, we align inline transcript annotations of referring expressions (REs) with per-role belief-state annotations. Preprocessing stitches multi-line REs across same-speaker continuation

turns while preserving intervening utterances, retaining full conversational context. Each RE yields a training instance containing the dialogue history up to the reference, gold giver and follower beliefs, a frictive status label (PENDING, MISUNDERSTOOD, or ALIGNED), and the following eight utterances for completion-side evaluation.

We collapse the three-way status into a binary distinction: instances labeled PENDING or MISUNDERSTOOD are treated as *frictive* and used for friction-cache construction and FPO training, while ALIGNED instances are reserved as evaluation distractors. Dialogues are split into train, development, and test sets by stratified sampling over misunderstood-RE density, with development dialogues drawn from the distribution median to reduce prompt overfitting. After existence-query and underspecified-row filtering, the final data contains 99 training dialogues with 2,705 REs (147 misunderstood) and 24 test dialogues with 761 REs (69 misunderstood). The split preserves the corpus's natural misunderstanding rate (5.5% train, 9.1% test) while maintaining a broad range of dialogue difficulty.

4 Theory-of-Mind-Grounded Friction Construction

4.1 Point-in-Time ToM Extraction

We extend the perspectivist annotation framework of Li et al. (2026) in two ways: (1) replacing retrospective grounding outcomes with point-in-time belief attribution, and (2) augmenting first-order beliefs with explicit second-order Theory-of-Mind (ToM) beliefs. The resulting four-identifier schema makes silent divergence mechanically detectable through cross-party belief comparison.

Two earlier approaches exposed the need for explicit cross-party ToM structure. Local-context belief prediction remained below a constant-True baseline, indicating that silent divergence is often only recoverable retrospectively through downstream repair or task failure rather than local dialogue evidence. Adapting the retrospective grounding cascade of Li et al. (2026) improved contextual access but still failed structurally: grounding was represented as an intra-party binary status, while silent divergence arises precisely when both participants confidently ground different referents. These failures imply that the required signal must be both retrospectively attributable and explicitly structured over cross-party first- and second-order beliefs.

We therefore introduce a four-identifier schema that attributes, for each referring expression at time t: ① the speaker's intended referent, ② the addressee's inferred referent, ③ the speaker's belief about the addressee's interpretation, and ④ the addressee's belief about the speaker's intention. The model conditions on the full dialogue, using downstream repair and contradiction evidence to reconstruct latent beliefs at t. Silent divergence then becomes a property of identifier comparison rather than a single grounding label.

The four landmark identifiers come from two different sources. The first-order identifiers correspond directly to the perspectivist annotations of Li et al. (2026), which provide speaker-intended and addressee-interpreted landmark IDs for each referring expression. The second-order identifiers do not exist in prior annotations and are introduced in our ToM-extended schema. They are generated entirely through GPT-5 inference using our prompting pipeline over the dialogue context, map information, and dialogue-act annotations.

Beyond the four identifiers, the schema extracts three auxiliary fields through GPT-5 inference: UPTAKE_QUALITY, SURFACE_SIGNALS, and IS_EXISTENCE_QUERY.

UPTAKE_QUALITY is a categorical field in the `addressee_belief` block with four values: `committed`, `hesitant`, `withheld`, and `absent`. It captures how confidently the addressee responds after an RE, ranging from confident acknowledgment to explicit non-commitment or no response. The friction module maps these categories to the uncertainty component of F^{-}, treating withholding as the strongest signal of unresolved uncertainty.

SURFACE_SIGNALS is a multi-label field recording repair-oriented discourse markers in the addressee response, including `acknowledge`, `pause`, `request_clarify`, `repair`, and `contradict`. The friction module uses these signals both to estimate the contradiction component of F^{-} and to infer whether the addressee is aware of a grounding mismatch.

IS_EXISTENCE_QUERY is a boolean field in the `speaker_belief` block identifying references that explicitly ask whether a landmark exists (e.g., "do you have a forest?"). These cases are excluded from FPO training because discovering asymmetry through questioning is treated as successful information exchange rather than grounding failure.

Pairwise comparisons among the four identi-

Configuration	①-②	③-①	④-②	Description
Aligned	$=$	$=$	$=$	Both parties land on the same landmark and correctly believe they have.
Silent intent fixing	$\neq$	$=$	$=$	Divergence exists but neither party detects it; both proceed confidently.
Asymmetric (speaker-aware)	$\neq$	$\neq$	$=$	Divergence exists; speaker has detected it, addressee has not.
Asymmetric (addressee-aware)	$\neq$	$=$	$\neq$	Divergence exists; addressee has detected it, speaker has not.
Mutual awareness	$\neq$	$\neq$	$\neq$	Divergence exists and both parties have detected it; repair is imminent or underway.

Table 1: The five canonical epistemic configurations produced by the four-identifier schema. Columns report pairwise agreement among the four identifiers (§4.1): ①-② = referent agreement between speaker and addressee (first-order); ③-① = whether the speaker correctly anticipates the addressee's interpretation; ④-② = whether the addressee correctly recovers the speaker's intent. "$=$" is agreement, "$\neq$" is divergence.

Component	Condition	Score
UNC	committed	0.05
	hesitant	0.55
	withheld	0.70
	absent	0.40
CONTR	contradict	0.85
	repair	0.65
	request_clarify	0.55
	pause	0.20
	acknowledge	0.05
	none	0.00
HAZ	existence asymmetry	0.65
	multiplicity divergence	0.55
	lexical variant	0.30
	otherwise	0.00
VALCONF	no divergence	0.00
	diverged, neither aware	0.95
	diverged, one aware	0.55
	diverged, both aware	0.25

Table 2: Friction component surrogates.

fiers induce five canonical epistemic configurations shown in Table 1. *Aligned* is the only non-frictive configuration, while *silent intent fixing*, as the central friction target, captures cases where interlocutors anchor to different referents while mutually presuming agreement.

4.2 Friction Computation

The friction module Φ maps the four-identifier schema to FPO's friction signal:

$$F^-(h) = w_{\text{UNC}}\,\text{UNC}(h) + w_{\text{CONTR}}\,\text{CONTR}(h) + w_{\text{HAZ}}\,\text{HAZ}(h) + w_{\text{VALCONF}}\,\text{VALCONF}(h) \quad (1)$$

$$F^+(h, a) = \text{INFOGAIN}(h, a) \quad (2)$$

All F^- components are deterministic scalar surrogates computed from the four-identifier schema and auxiliary dialogue-observable fields, normalized to $[0, 1]$, and summarized in Table 2. UNC operationalizes confidence–reliability divergence from addressee uptake behavior. CONTR operationalizes contradiction with interaction history via repair-oriented discourse markers, taking the maximum active surface signal to prevent strong markers from being diluted by weaker hesitation. HAZ operationalizes downstream propagation risk from corpus-level landmark metadata, taking the maximum applicable condition and scaling by remaining task length (*len_mult* $\in \{0.7, 1.0, 1.3\}$). VALCONF operationalizes second-order belief conflict directly from the four-identifier structure: divergence is ① $\neq$ ②; speaker awareness is ③ $\neq$ ①; addressee awareness is ④ $\neq$ ② or the presence of repair-oriented surface signals.

Composition. The composite assigns weights $w = (0.15, 0.20, 0.30, 0.35)$ for (UNC, CONTR, HAZ, VALCONF) respectively, with $\sum_i w_i = 1$, so $F^-(h) \in [0, 1]$.

The weighting follows FPO's distinction between recoverable precursor signals and structurally persistent alignment failures. Hesitation and repair behaviors (UNC, CONTR) indicate surfaced divergence that remains partially recoverable, whereas propagation-prone misalignment and silent referential conflict (HAZ, VALCONF) may remain undetected while distorting downstream coordination. Accordingly, most weight mass is assigned to HAZ and VALCONF, with the latter receiving the highest weight because mutually incompatible yet confidently maintained interpretations constitute the strongest form of latent epistemic divergence.

Because the components are computed independently, multiple risks may co-occur within a single referring expression. We additionally introduce a structural fallback that increases HAZ and VALCONF for concepts belonging to the dataset's multiplicity-divergent set despite identical pre-

Configuration	clar.	verif.	redir.	refuse
Divergent:				
`silent_intent_fixing`	0.85	**0.95**	0.55	0.20
`asym_aware_speaker_only`	0.85	0.80	0.45	0.15
`asym_aware_addressee_only`	0.85	0.70	0.45	0.15
`open_dispute`	0.70	0.80	0.50	0.10
`multiplicity_suspicious`	0.75	**0.85**	0.40	0.10
Aligned:				
`aligned_deep`	0.05	0.10	0.05	0.02
`aligned_with_repair`	0.40	0.40	0.20	0.05
Underdetermined:				
`underspecified`	**0.90**	0.50	0.25	0.05
`existence_query`	0.20	0.30	0.10	0.05

Table 3: $F^{+}(h, a)$ values by referential configuration and intervention type.

dicted identifiers.

Productive friction. $F^{+}(h, a)$ scores intervention utility over the four FPO actions (`clarify`, `verify`, `redirect`, `refuse`), with the no-intervention baseline fixed at $F^{+}(\texttt{answer}) = 0$ following Pustejovsky and Krishnaswamy (2026). In the offline setting, $\textsc{InfoGain}(h, a)$ is implemented as the lookup table in Table 3. Diagnostic actions (`clarify`, `verify`) receive the highest values on divergent states, especially `silent_intent_fixing`, where neither participant locally detects the mismatch. Corrective actions (`redirect`, `refuse`) are scored lower because they assume the source of divergence is already identifiable and impose greater conversational cost. Aligned states receive near-zero values, encoding that unnecessary intervention yields little epistemic gain.

5 Friction Signal Validation

Before using the friction cache for trainer adaptation (§6), we validate the signal along three axes: (i) whether the schema recovers usable belief structure and misunderstanding labels, (ii) whether the composite F^{-} monotonically separates frictive from non-frictive contexts, and (iii) whether the signal materially depends on the second-order ToM structure introduced in §4.1. Results are reported on both train and test partitions throughout.

5.1 Schema Validation and Belief Accuracy

The four-identifier schema achieves 98.5% parse success on train (2665/2705) and 100% on test (761/761); the remaining train failures are primarily long-dialogue API truncations and are filtered before downstream training. Belief accuracy against Li et al. (2026) is shown in Table 4.

Quantity (loose match)	Train	Test
Speaker belief	59.8%	64.1%
Addressee belief	81.3%	80.6%
Speaker ToM consistency	35.9%	43.5%
Addressee ToM consistency	41.4%	42.0%

Table 4: Four-identifier accuracy on train and test partitions.

Addressee-side belief recovery substantially exceeds speaker-side recovery (80.6–81.3% vs. 59.8–64.1% loose match), which is expected because addressee interpretations are more directly observable through uptake and repair behavior, whereas speaker intent must be inferred from production-side evidence alone. Since both friction computation and trainer adaptation primarily condition on addressee-side state, the stronger follower-side accuracy is the operationally relevant quantity.

Second-order ToM consistency is lower (35.9–43.5% speaker ToM; 41.4–42.0% addressee ToM), but these values are reported as soft signals rather than gold-aligned accuracy because no direct annotation for second-order beliefs exists in the corpus. To directly assess the reliability of GPT-5 second-order predictions, we conducted a stratified blind annotation pilot on 60 samples. A human annotator independently assigned second-order belief labels without access to the model outputs, after which the annotations were compared against GPT-5 predictions. Agreement reached 82% for speaker-ToM and 80% for addressee-ToM, corresponding to Cohen's κ values of 0.80 and 0.77, respectively. These results indicate substantial agreement with human judgments, supporting the use of the inferred second-order beliefs as supervisory signals during training.

At the task level, the schema recovers misunderstanding labels with 67.6% recall on train (98/145) and 65.2% on test (45/69), improving more than 60 points over the cascade-based alternatives discussed in §4.1.

5.2 F^{-} Distribution and Monotonic Ordering

The composite F^{-} is intended to function as a monotonic risk signal: genuinely frictive contexts should receive systematically higher values than non-frictive ones. This ordering holds consistently across both partitions (Table 5). Mean F^{-} increases from 0.185 to 0.375 on train ($\Delta = 0.190$) and from 0.193 to 0.390 on test ($\Delta = 0.197$) when moving from pending to misunderstood contexts. The near-identical cross-partition gap suggests that

Partition	Status	n	Mean F^-	Median
Train	Pending	2521	0.185	0.182
	Misund	145	0.375	0.340
Test	Pending	692	0.193	0.203
	Misund	69	0.390	0.340

Table 5: F^- distribution by gold status on both partitions

Detection channel	Train (%)	Test (%)
ToM-required configurations	44.2	39.1
First-order divergence	1.4	–
Multiplicity fallback	22.1	26.1
Total recall	67.6	65.2

Table 6: Channel decomposition of detected misunderstandings.

the composite generalizes rather than fitting a single split. This monotonicity is required for downstream trainer adaptation, where $F^-(h)$ directly controls trust-region width: a non-monotonic signal would incorrectly relax constraints on high-risk contexts or tighten them on benign ones.

5.3 Detection Channels and ToM Contribution

We further decompose detected misunderstandings by the channel responsible for recovery: (i) ToM-dependent configurations involving silent intent fixing or asymmetric awareness, (ii) direct first-order divergence, and (iii) metadata-based multiplicity fallback. Results are summarized in Table 6.

Second-order ToM structure accounts for a substantial portion of recovered misunderstandings. ToM-dependent configurations contribute 44.2% of gold misund train cases and 39.1% on test, including all silent-intent-fixing instances — the canonical FPO target. These cases are not recoverable from surface dialogue alone because interlocutors appear aligned while grounding different referents. Multiplicity fallback contributes an additional $\sim$ 22–26%, whereas direct first-order divergence accounts for only a negligible fraction.

The channels are non-redundant. Removing the ToM channel reduces recall to 23.4% on train and 26.1% on test; removing multiplicity fallback further reduces recall to 1.4% on train and 0% on test. Overall, more than half of recovered misunderstandings are ToM-attributable (65.3% on train, 60.0% on test), indicating that the friction signal depends materially on the second-order belief structure introduced by the four-identifier schema.

Most remaining misses involve underspecified or visually ambiguous references that the upstream extractor cannot disambiguate from dialogue and map evidence alone, reflecting extraction limitations rather than failures of the friction computation itself.

6 FPO Trainer Adaptation

We adapt a friction-blind DPO baseline (Rafailov et al., 2023), and three FPO variants, FAR, FPP, and FTR (Pustejovsky and Krishnaswamy, 2026) to our offline MapTask corpus. Although Pustejovsky and Krishnaswamy (2026) admits both participant-agent and mediator-agent interpretations, we instantiate the participant-agent setting (not as a mediator): it observes only its own map and dialogue history, and decides per turn whether to `answer` or intervene. Friction scores are computed offline from our four-identifier ToM annotations (§4.1) and injected as reward, filter, or trust-region signals. GRFR is excluded as it requires multi-turn rollouts unavailable in a single-turn RE corpus.

Training instances. Each instance pairs a referring expression with its immediate follower response. The input combines dialogue history with an addressee-side belief preamble; unobservable identifiers are masked so the agent's training-time input mirrors its inference-time information state. After filtering underspecified and existence-query rows, the cache holds 663 training rows (216 divergent, 447 aligned) and 278 test rows. Pair-format methods (DPO, FPP) use corpus-retrieved rather than synthetic completions: for divergent rows the gold acknowledgment is y^- and a retrieved repair is y^+; reversed for aligned rows.

DPO (friction-blind baseline). Standard DPO (Rafailov et al., 2023) with $\beta = 0.1$ and a frozen reference policy; the loss reads only the chosen/rejected token-level tensors and is by definition friction-blind.

FAR (Friction-Aware Reward). FAR shapes the policy-gradient reward so that friction is encouraged selectively — only when epistemic risk is high — and penalized when it imposes unnecessary cost: $R' = R_{\text{task}} + \alpha\, g(\text{Risk})\, F(h, y) - \beta\, C_{\text{fric}}(a)$, with Risk grounded in the HAZ component of F^-. Applying this offline required two fixes. On divergent rows the gold is a silent acknowledgment, so a negative R_{task} collapses training without an alternative target; we retarget these rows to the corpus-retrieved repair template, making $R_{\text{task}}{=}{+1}$ everywhere. The on-policy gradient update also requires

rollouts the corpus cannot supply, so we substitute REINFORCE with running-mean baseline subtraction (Williams, 1992) and reward clipping.

FPP (Friction Preference Pairing). FPP fits a DPO-style logistic loss to pairs ordered by productive friction: y^+ must have strictly higher F^+ than y^-. Three offline fixes were adopted. First, $F^+(\texttt{answer})$ is undefined in the friction module, which collapses the ordering on every aligned pair; we restore it with $F^+(\texttt{answer} \mid \text{aligned}) = 1 - \max_a F^+(h, a)$, leaving the module itself untouched. Second, F^- varies by an order of magnitude across configurations, so we reweight the loss by $w(h) \propto F^-(h) + \varepsilon$ to concentrate gradient on rows where the preference signal is most informative. Third, MapTask interventions are encoded prosodically rather than lexically — only 9.3% of divergent gold templates trigger the rule-based classifier — so we prepend a minimal class-marker token to divergent chosen completions that do not already surface the correct pattern (*sorry*→clarify, *so*→verify, *no*→redirect, *i can't*→refuse).

FTR (Friction-Conditioned Trust Region). FTR uses epistemic risk to govern how much the policy is allowed to change during training. The key insight is that the base model's default behavior is most dangerous precisely in high-friction contexts, so those are the contexts where the policy should be freest to depart from it. FTR encodes this by making the KL trust-region radius a function of $F^-(h)$, and enforcing it via Lagrangian relaxation, which integrates naturally with the supervised CE objective on the gold completion, unlike PPO-style clipping:

$$L_{\text{FTR}}(h) = L_{\text{CE}}(y_{\text{gold}}|h) + \beta(h) \cdot \widehat{\text{KL}}(\pi_\theta \| \pi_0 | h),$$
$$\beta(h) = \frac{1}{\epsilon_0 + \kappa F^-(h)}.$$

Three choices fill gaps the FPO specification leaves open. Exact KL is intractable for a single gold completion, so we use the k_3 estimator $\widehat{\text{KL}} = r - 1 - \log r$ (Schulman, 2020), which asymmetrically penalizes decreases in gold probability — an asymmetry that signed- and squared-log-ratio estimators failed to preserve in pilots. We source $\text{Risk}(h) = F^-(h)$ directly rather than training a separate risk head, since F^- already aggregates all four epistemic risk components (§4.2). Finally, $\epsilon_0 = 1.0$ and $\kappa = 3.0$ place $\beta(h) \in [0.36, 1.0]$; tighter regimes over-anchored the policy and inflated ECE.

Backbone and training. All methods share Qwen3.5-27B fine-tuned with 4-bit QLoRA (Qwen Team, 2025; Dettmers et al., 2023), differing only in loss and data format. At inference, cross-party ToM annotations are unavailable; the model receives only its own dialogue history and the observable projection of the belief preamble — a privileged-information distillation setup (Vapnik and Izmailov, 2015) that keeps the approach deployable without cross-party access at test time.

7 Evaluations

We evaluate on the held-out MapTask test set (n=278) using three metrics from the FPO evaluation framework (Pustejovsky and Krishnaswamy, 2026). **ClarifyScore/F1** measures the net rate of *appropriate* intervention: it rewards intervening when divergence is present and penalizes intervening when it is not, so a model that intervenes on every row gets a high recall but is punished for false alarms. **ECE** is the expected absolute gap between the model's expressed confidence (first-token probability) and empirical correctness — lower is better, and we report it separately on warranted rows (divergent configurations, n=109) and aligned rows (n=169) to distinguish calibration on risky vs. safe contexts. **InfoEff** is the expected epistemic gain per unit of intervention cost, using raw F^+ as the Info-Gain proxy — it rewards interventions that reduce uncertainty efficiently rather than indiscriminately. Pairwise win rates use Gemma-4-31B-IT (Team and DeepMind, 2026) with position bias debiased by querying each pair in both orderings.

7.1 Headline Results

FAR and FTR both improve significantly over DPO across all three metric families: F1 (+21% relative for both), ECE (FAR −13%), and the Gemma judge (FAR +15.8pp, $p = 4.5 \times 10^{-4}$; FTR +13.3pp, $p = 2.9 \times 10^{-3}$). The two methods land in the same statistical neighborhood, indicating that the friction signal transfers across two structurally distinct insertion points — reward shaping and trust-region control.

FPP provides an instructive failure case. Although it achieves the highest F1 (0.603), this result is degenerate: FPP intervenes on every row (ask rate 100%), yielding perfect recall by construction and a strongly negative judge preference (−91.4pp). The collapse stems from two interacting factors. First, pairwise DPO-style objectives constrain only relative ranking and provide no absolute behavioral anchor, allowing optimization

Policy	F1	ECE (global)	InfoEff	ECE (warr.)	ECE (align.)	Net win vs DPO
DPO	0.344 [0.253, 0.431]	0.174 [0.135, 0.238]	3.99 [3.21, 4.82]	0.432 [0.349, 0.513]	**0.148** [0.115, 0.227]	—
FAR	**0.416*** [0.332, 0.498]	**0.152*** [0.109, 0.216]	3.92 [3.31, 4.57]	0.227 [0.159, 0.316]	0.208 [0.156, 0.289]	**+15.8**** [+7.2, +24.1]
FPP†	0.603 [0.545, 0.657]	0.568 [0.511, 0.626]	3.78 [3.39, 4.21]	0.000 [0.000, 0.000]	1.000 [1.000, 1.000]	−91.4 [−95.0, −87.4]
FTR	**0.417*** [0.332, 0.502]	0.189 [0.138, 0.249]	3.94 [3.23, 4.71]	**0.178** [0.133, 0.290]	0.220 [0.175, 0.307]	**+13.3**** [+5.0, +21.6]

Table 7: Held-out evaluation (n=278), 95% bootstrap CIs (B=2,000). Net win is paired win-minus-loss % vs DPO under Gemma-4-31B-IT. †FPP intervenes on every row (ask rate 100%); its F1 and ECE reflect this degenerate configuration).

to collapse toward the dominant acknowledgment distribution. Second, MapTask encodes many interventions prosodically rather than lexically, leaving weak observable supervision for preference pairing: only $20/216$ (9.3%) of divergent chosen completions contain explicit lexical intervention markers. FAR and FTR avoid this failure because they condition directly on per-row friction through $R'(h, a, y)$ and $\beta(h) = 1/\epsilon(h)$ rather than completion contrasts.

The warranted-subset ECE is the load-bearing calibration signal. FAR and FTR are the only methods whose warranted ECE drops below DPO's (0.227 and 0.178 vs 0.432), while DPO is well calibrated on aligned rows but badly miscalibrated on warranted ones — its calibration tracks label frequency rather than epistemic state.

7.2 Per-Configuration and Length Robustness

Configuration (n)	FAR	FTR
silent_intent_fixing (35)	**42.9**	**42.9**
multiplicity_suspicious (43)	**39.5**	30.2
asym_aware_addressee (28)	**42.9**	35.7
aligned_with_repair (24)	**41.7**	**41.7**
aligned_deep (145)	29.7	29.7

Table 8: Gemma win rate vs DPO by configuration (%). FPP omitted ($\leq$3% on all rows).

Stratifying by referential configuration shows FAR and FTR win most on the configurations with the largest belief asymmetries (`silent_intent_fixing`, `multiplicity_suspicious`, `asym_aware_addressee`), and smaller but positive estimates on aligned classes. Subgroup CIs are wide given small n; we treat these as directional evidence and rely on the full-sample headline for statistical claims. FAR and FTR also produce shorter responses than DPO (median $16/15$ vs 29 tokens), consistent with friction-conditioned policies preferring concise interventions. On the length-balanced subset (n=78/77), neither FAR (15.4% win, 16.7% loss, p=1.0) nor FTR (11.7% win, 9.1% loss, p=0.80) reaches significance — confirming that the LLM judge has a known length-preference bias (Singhal et al., 2023). We therefore treat task-grounded F1/ECE/InfoEff as primary evidence, since they score functional appropriateness independently of length.

8 Conclusion

Interaction systems are typically evaluated on response quality rather than whether agents converge on shared belief states, allowing epistemic divergence to accumulate silently until downstream tasks fail. We show that this failure stems from missing representational machinery rather than incidental error. A four-identifier Theory-of-Mind schema makes second-order belief divergence measurable within individual references, revealing that many misunderstandings arise from untracked belief asymmetries rather than surface ambiguity; the ablation confirms these signals are not recoverable from first-order information alone. Once explicit, this structure becomes actionable. FAR and FTR both operationalize ToM-grounded friction as a training signal and outperform friction-blind DPO across F1, calibration, and judge preference. That these gains transfer across distinct optimization methods suggests that the improvement derives from the friction representation itself rather than any particular objective. More broadly, these results recast alignment as a state-tracking problem: the goal of interaction is not merely to generate plausible responses, but to maintain and repair shared belief states over time. Successful communication is measured not by what is said, but by whether interaction preserves or restores alignment in what is believed.

Limitations

Our experiments are conducted on the HCRC MapTask corpus, a controlled navigation task chosen because its dense referential grounding and mismatched participant maps provide an ideal testbed for evaluating latent epistemic alignment. While this structured environment allows for precise evaluation of the four-identifier ToM schema, exploring its generalization to less constrained open-domain or multi-party dialogue remains an objective for future work. Additionally, the misunderstanding rate in MapTask (5.5–9.1%) reflects its specific task asymmetric design and may vary in other collaborative corpora.

Because direct human annotations for second-order mental states do not exist in the original corpus, the second-order belief identifiers are derived through GPT-5 inference. To mitigate potential extraction bias, we validated these labels against independent human annotations in a 60-sample pilot study, yielding high agreement ($\kappa = 0.80/0.77$) that supports the reliability of the inferred labels as supervisory signals.

Finally, this instantiation computes friction scores offline to leverage gold-standard grounded trajectories during training, rather than estimating them incrementally in a fully online FPO loop. While this design choice properly controls for training stability and clarifies the evaluation of individual extraction channels, it leaves multi-turn rollout variants like GRFR out of scope. At inference time, the system deployably relies on the belief-asymmetry sensitivity acquired during privileged-information distillation training; evaluating the bounds of this transfer under severe out-of-distribution dialogue conditions remains an active area of ongoing exploration.

Potential Risks

The primary artifact introduced in this work is a friction signal derived from Theory-of-Mind inference over dialogue. While the immediate application is alignment training for collaborative dialogue agents, the same underlying machinery could, in principle, be deployed to detect and strategically exploit belief asymmetries between interlocutors rather than to repair them. We explicitly flag this dual-use potential for adversarial manipulation; however, we note that our framework is structurally engineered for epistemic stewardship, optimizing exclusively for transparency, verification, and mutual repair signals.

Additionally, our second-order belief inference pipeline relies on GPT-5, a proprietary model whose latent representations may reflect demographic or stylistic biases present in its training data. Systematic errors in ToM attribution—particularly speaker-side belief inference, where baseline accuracy is naturally bounded —could cascade into miscalibrated friction signals that penalize appropriate conversational progress or incentivize redundant, costly interventions. Consequently, we strongly advise that any downstream adaptation of friction-conditioned policies in high-stakes domains maintain strict human-in-the-loop oversight to audit model calibration during continuous interaction.

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC Map Task corpus. *Language and Speech*, 34(4):351–366.

Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Preprint*, arXiv:2305.14314.

Ibrahim Khalil Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard A. Brutti, Christopher Tam, Jingxuan Tu, Benjamin A. Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, and James Pustejovsky. 2024. Common ground tracking in multimodal dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3587–3602, Torino, Italia. ELRA and ICCL.

Nan Li, Albert Gatt, and Massimo Poesio. 2026. Grounded misunderstandings in asymmetric dialogue: A perspectivist annotation scheme for MapTask. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4988–5001. ELRA and ICCL.

Abhijnan Nath, Carine Graff, Andrei Bachinin, and Nikhil Krishnaswamy. 2025. Frictional agent alignment framework: Slow down and don't break things. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11042–11089, Vienna, Austria. Association for Computational Linguistics.

Abhijnan Nath, Hannah VanderHoeven, and Nikhil Krishnaswamy. 2026. CRAFT: Grounded multi-agent coordination under partial information. *Preprint*, arXiv:2603.25268.

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

James Pustejovsky and Nikhil Krishnaswamy. 2025. Frictive policy optimization for LLM agent interactions. 2025 International Conference on Autonomous Agents and Multiagent Systems.

James Pustejovsky and Nikhil Krishnaswamy. 2026. Frictive policy optimization for LLMs: Epistemic intervention, risk-sensitive control, and reflective alignment. *Preprint*, arXiv:2604.25136.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388. We use the `Qwen3.5-27B` variant from the same model family.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

John Schulman. 2020. Approximating KL divergence. `http://joschu.net/blog/kl-approx.html`. Blog post.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and 1 others. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023. A long way to go: Investigating length correlations in RLHF. *Preprint*, arXiv:2310.03716.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA.

Gemma Team and Google DeepMind. 2026. Gemma 4: Open multimodal models based on Gemini 3 technology.

Vladimir Vapnik and Rauf Izmailov. 2015. Learning using privileged information: similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16(61):2023–2049.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256.

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Yifan Zhu, Mariah Bradford, Kenneth Lai, Timothy Obiso, Videep Venkatesha, James Pustejovsky, and Nikhil Krishnaswamy. 2026. Distributed partial information puzzles: Examining common ground construction under epistemic asymmetry. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4974–4987. ELRA and ICCL.

A Sensitivity to Design-Choice Constants

The friction module relies on two sets of hand-set constants: the component weights $W = (w_{\text{UNC}}, w_{\text{CONTR}}, w_{\text{HAZ}}, w_{\text{VALCONF}}) = (0.15, 0.20, 0.30, 0.35)$ used to combine the four components into F^-, and the categorical-to-scalar lookup tables for surface signals (e.g., `withheld`$\rightarrow$ 0.70, `request_clarify`$\rightarrow$ 0.55). We perturb each weight w_k multiplicatively by $\delta \in \{-0.20, -0.10, +0.10, +0.20\}$ (independently and in concert), and separately perturb all scalar mappings uniformly by the same δ range. For every perturbation we recompute F^- over the 278 test rows, re-evaluate the configuration ordering, and re-derive the FTR trust-region coefficients $\beta(h) = 1/(\epsilon_0 + \kappa F^-(h))$.

The perturbation sweep yields three results. First, the mean F^- ordering of the five canonical configurations is preserved on all five structurally informative pairs under every perturbation, with the sole exception of the near-tie pair separated by 0.005 in the reference. Second, the cross-class gap between the most- and least-frictive configurations stays in $[0.37, 0.56]$, well above any threshold that would distinguish meaningful from incidental separation. Third, FTR's per-row $\beta(h)$ keeps a roughly $2\times$ ratio between aligned and divergent contexts; perturbing the F^+ table changes INFOEFF proportionally but does not flip the inter-method ranking. The design constants therefore function as robust structural surrogates: their numerical values can be moved by $\pm 20\%$ without disrupting any conclusion in §7.

Perturbation	Cfg Kendall τ	Cfg gap	β-ratio	InfoEff rank
Reference	$0.867^{\star}$	$+0.46$	2.12	DPO $<$ FAR/FTR
$\delta = -20\%$ (uniform W)	$0.867^{\star}$	$+0.37$	1.93	preserved
$\delta = -10\%$ (uniform W)	$0.867^{\star}$	$+0.42$	2.03	preserved
$\delta = +10\%$ (uniform W)	$0.867^{\star}$	$+0.51$	2.17	preserved
$\delta = +20\%$ (uniform W)	$0.867^{\star}$	$+0.56$	2.18	preserved
$\delta = \pm 20\%$ (single w_k, worst case)	≥ 0.733	$+0.40$	1.93	preserved
$\delta = \pm 20\%$ (scalar map, uniform)	≥ 0.733	$+0.37$	1.93	preserved

Table 9: Sensitivity to friction-module constants (n=278). **Cfg** τ: Kendall rank-correlation of per-configuration mean F^- vs. reference. **Cfg gap**: F^-(`silent_intent_fixing`) $-$ F^-(`aligned_deep`). β**-ratio**: β(`aligned_deep`)/β(`silent_intent_fixing`) under FTR. **InfoEff rank**: whether DPO $<$ {FAR,FTR} survives. *Single near-zero tie between `multiplicity_suspicious` and `aligned_with_repair` (Δ=0.005); all five structurally informative pairs are concordant in every perturbation.

B Human Annotation Protocol for the ToM Validation Pilot

To evaluate the reliability of GPT-5 second-order ToM predictions (§5.1), we conducted a blind human-annotation pilot on a stratified sample of 60 referring expressions drawn from the held-out test partition ($n_{\text{test}} = 761$). Sampling enforced a minimum of eight items per configuration class whenever class frequency permitted, ensuring coverage of all eight epistemic configurations, including divergent cases such as `silent_intent_fixing` and `multiplicity_suspicious`.

The annotator was a co-author of this paper, fluent in English and familiar with both the HCRC MapTask corpus and the four-identifier ToM schema introduced in §4.1. The pilot was designed to evaluate whether GPT-5 second-order predictions aligned with human-interpretable referential judgments under the proposed schema, rather than to establish inter-annotator agreement for the schema itself.

For each item, the annotator had access to the dialogue context preceding the target referring expression, the target expression itself, speaker role (giver or follower), dialogue and map identifiers, and the corresponding MapTask map images for both the giver and the follower. This matches the GPT-5 prompting condition (§4.1) on the dialogue-text and map-vision channels; the only information unavailable to the annotator was the inline dialogue-act annotation stream additionally consumed by GPT-5 for surface-signal grounding. GPT-5 outputs were withheld throughout annotation to prevent anchoring effects.

The worksheet header provided the following written instructions:

> *For each row, check the corresponding map and read the* `dialogue_context` *and* `target_re_text`*. Second-order ToM refers to what each participant believes the other participant meant or understood. Fill in predicted landmark identifiers for* `user_speaker_tom_landmark` *and* `user_addressee_tom_landmark`*. Use corpus landmark identifiers where possible (e.g.,* `m3_sandstone_cliffs#0@g`*); free-text descriptions are also acceptable. Enter* `unknown` *if the referent cannot be determined from dialogue evidence alone.*

No additional verbal instructions were provided. Annotation was completed in a single session of approximately two hours, producing 120 second-order labels (60×2 identifier slots). The worksheet was edited in spreadsheet format and exported to CSV.

Agreement was computed using an automated scoring script that canonicalized map prefixes (`m3_`, `m10_`), instance indices (`#0`), map-role suffixes (`@f`, `@g`), common spelling variants (`wheat_fields`/`wheatfields`), and leading articles before exact-match comparison. Cohen's κ is reported over the canonicalized labels.

The HCRC MapTask corpus (Anderson et al., 1991) is publicly released for academic research and contains no personally identifying or sensitive material. The annotation task involved only referential interpretation over existing corpus data. The annotator received no separate compensation and performed the work as part of routine research activity. Because the pilot involved only author annotation over publicly available data, IRB review was not sought.