From Propositional to Perceptual Asymmetry: Extending Frictive Policy Optimization to Asymmetric Partial Information Dialogue

Yifan Zhu and Kyeongmin Rim and James Pustejovsky
Brandeis University, Waltham, MA, USA
{zhuyifan, krim, jamesp}@brandeis.edu

Abstract

Frictive Policy Optimization (FPO; Pustejovsky and Krishnaswamy, 2025) treats friction in collaborative dialogue – misalignment, misunderstanding, repair – as an epistemic signal essential to common-ground construction, rather than noise to be minimized. However, FPO and its implementations assume shared perceptual contexts, where friction arises from differently interpreted propositions over the same scene, which we define as *propositional asymmetry*. We extend FPO to *perceptual asymmetry*, where participants hold asymmetric partial information and the same referring expression yields different denotations depending on whose information state grounds the reference. We evaluate this through cross-corpora analysis and LLM probing on referentially asymmetric dialogue tasks, primarily the HCRC MapTask (Anderson et al., 1991). We find that FPO's friction functional is empirically valid only when evaluated from within each participant's information horizon: different landmark configurations produce qualitatively distinct grounding failure modes, with a small class of ambiguous configurations driving a disproportionate share of misunderstandings through trajectories that appear successful but silently diverge. The LLM probe confirms that having the "right perspective" matters more than having all perspectives: the informed single viewpoint outperforms omniscient access to both participants' contexts. We propose two annotation refinements: subtype decomposition of pending grounding states and accommodation-aware alignment classification.

1 Introduction

Collaborative dialogue relies on common ground, but building that common ground is inherently not frictionless: participants misinterpret, fall out of alignment, and then repair. Rather than treating this friction as noise, Frictive Policy Optimization (FPO; Pustejovsky and Krishnaswamy, 2025) reframes it as an epistemic signal: moments of misunderstanding expose belief-state divergences that policies can exploit for more robust alignment. However, existing implementations of FPO (FAAF; Nath et al., 2025) assume a shared perceptual context where participants reason over the same scene but arrive at different beliefs, which we call *propositional asymmetry*.

Recent work on collaborative construction under epistemic asymmetry points to a qualitatively different setting. In the Distributed Partial Information Puzzle (DPIP; Zhu et al., 2026), each participant holds a distinct partial view of the target structure, and successful communication requires reconciling information that was never jointly observed; we call this *perceptual asymmetry*. As formalized by lexical event models (Pustejovsky and Zhu, 2024), even identical surface expressions (e.g., "the red block is on top") can evoke different event structures depending on which perceptual frame grounds the reference, yielding referential asymmetry as a downstream consequence.

This perceptual divide motivates an extension of FPO to settings where alignment is evaluated relative to each participant's information horizon, rather than a shared frame. Concretely, we generalize the friction functional to a perspectival formulation with a role-specific decomposition: the speaker incurs friction from within-frame ambiguity at production, while the addressee incurs friction from cross-frame conflicts at interpretation. Under this view, common ground is not presupposed but must be actively constructed by reconciling perceptually disjoint evidence.

Specifically, we make the following contributions. First, through corpus analysis and LLM probing on the HCRC MapTask, we show that even in this controlled cooperative-grounding task, collaborative dialogue exhibits substantial perceptual asymmetry: participants reason over private, non-overlapping perceptual states, making fric-

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 403–413
August 2–5, 2026.

tion perspective-dependent and limiting the reach of shared-scene assumptions. Second, we introduce a perspective-bound evaluation of the friction functional and demonstrate that, compared to an omniscient view, this formulation reveals substantially more frictive states, capturing divergences that would otherwise remain undetected. Finally, motivated by these findings, we propose two targeted refinements to existing grounding-annotation schemes: (i) decomposing unresolved grounding states into finer-grained subtypes aligned with frictive substates, and (ii) distinguishing alignment achieved through negotiation from alignment achieved through silent accommodation by linking grounding trajectories with dialogue-act structure. Throughout these three contributions, this paper develops the *diagnostic* and conceptual machinery for perspective-bound FPO and demonstrates its empirical purchase on existing dialogue data.

2 Background

2.1 LLM Alignment

Mainstream LLM alignment methods, such as RLHF (Christiano et al., 2017), DPO (Rafailov et al., 2024), and GRPO (Shao et al., 2024), optimize the policy toward a fixed preferred response. In collaborative dialogue this can be counterproductive: an agent that confirms apparent agreement without verifying actual common ground becomes a source of false common ground rather than a safeguard against it. Frictive Policy Optimization (FPO; Pustejovsky and Krishnaswamy, 2025, 2026) proposes that deliberate resistance to immediate task completion via clarification, verification, or challenge, is a rational control action. FPO factors the dialogue policy into two complementary components: an *intervention policy* $\pi_{\text{int}}(a \mid h)$ that decides *whether* (and what type of) friction action a to take given history h, and a *generation policy* $\pi_{\text{gen}}(y \mid h, a)$ that realizes the chosen action as an utterance y. FPO further defines a structured friction functional $F(h, y) = F^{+}(h, y) - F^{-}(h, y)$ that decomposes epistemic risk into independently measurable components. The Frictional Agent Alignment Framework (FAAF; Nath et al., 2025) provides the first empirical implementation, showing that friction-based alignment improves collaborative reasoning on tasks with shared perceptual context. We adopt FPO's friction functional as our theoretical foundation and extend it to settings where FAAF's shared-context assumption does not hold (§4).

2.2 Grounding in Asymmetric Dialogue

Common ground, the mutual knowledge, beliefs, and assumptions that participants rely on when coordinating joint actions, is not a static precondition but a dynamic achievement, constructed incrementally through the process of *grounding* (Clark and Brennan, 1991; Clark, 1996). Grounding succeeds when both participants have sufficient evidence that their interpretations converge; it fails silently when both proceed as if convergence holds but their interpretations in fact diverge.

Several task-oriented dialogue corpora study collaborative grounding under information asymmetry: the Weights Task (Khebour et al., 2024), DeliData (Karadzhov et al., 2023), the HCRC MapTask (Anderson et al., 1991), OneCommon (Udagawa and Aizawa, 2019), and the Distributed Partial Information Puzzle (DPIP; Zhu et al., 2026). These corpora differ in *what* diverges across participants: in *propositional asymmetry*, participants share a scene S but assign different propositional content (Weights Task, DeliData); in *perceptual asymmetry*, each participant observes a different subset of evidence (MapTask, OneCommon, DPIP). We focus on the latter, and within it on *structural* asymmetry: stable, role-bound divergences originating in the organization of the task (e.g., the giver and follower roles in MapTask), which shape the distribution of accessible evidence throughout interaction. Such structurally differentiated access induces downstream perceptual and propositional asymmetries as dialogue unfolds. The relevant point of comparison across corpora is therefore not the surface task domain but the underlying role-conditioned asymmetry structure, which recurs across many collaborative dialogue settings; the perspective-bound formulation applies wherever the evidential history h becomes participant-specific.

Li et al. (2025) provide the first annotation scheme tracking grounding at the level of individual perspectives, separately recording the speaker's intended and addressee's interpreted landmark for each reference expression in MapTask (§ 3.1). Their key finding is that a small class of multiplicity discrepancies drives a disproportionate share of misunderstandings, which motivates the perspective-bound friction taxonomy in § 4. They also note an omniscience bias risk: their GPT-5 annotator, with access to both maps, may underestimate asymmetric knowledge by treating cross-

perspective contradictions as noise, a caveat we operationalize as a measurable quantity in § 6.

3 Dataset

3.1 HCRC MapTask

We draw on two annotation layers over the same 128 HCRC MapTask dialogues (Anderson et al., 1991). In each dialogue, a giver who holds a printed route guides a follower, who holds a different map of the same area, to reproduce the route. The two maps share most landmarks but differ in controlled ways: some landmarks appear on one map but not the other (*existence* discrepancies), some share a name but differ in label (*lexical* discrepancies), and a small number appear twice on one map but once on the other (*multiplicity*).

The first layer is the HCRC MapTask NXT annotation release (v2.1, 2011),[1] which packages the original corpus annotations in NXT XML format with dialogue act move labels (`instruct`, `check`, `clarify`, `acknowledge`, etc.) aligned to word-level timed units. We use the move labels to identify clarification and verification acts in our accommodation analysis (§5.2).

The second layer is the per-perspective annotation by Li et al. (2025), which we refer to as the Grounded Misunderstandings in MapTask dataset (GMMT, our abbreviation for convenience). GMMT annotates 13,077 reference expressions (REs) across the same 128 dialogues, separately recording the speaker's intended and the addressee's interpreted landmark for each RE, yielding three understanding states: *aligned*, *pending*, and *misunderstood*. REs sharing a concept ID within a dialogue form a *reference chain*, tracking how grounding for a single landmark evolves across the interaction. Crucially, GMMT introduces a unified landmark ID scheme that distinguishes multiplicity instances and lexical variants that the original NXT scheme collapses, exposing cases of false common ground invisible under the earlier annotation (Figure 1). This distinction matters: as Table 1 shows, multiplicity discrepancies, though only 6% of landmarks, account for a 12% misunderstanding rate—six times the corpus average.

We join the two layers at the (dialogue_id, utterance_id) level, giving each GMMT reference expression a list of co-occurring NXT move tags.

[1] https://groups.inf.ed.ac.uk/maptask/maptasknxt.html

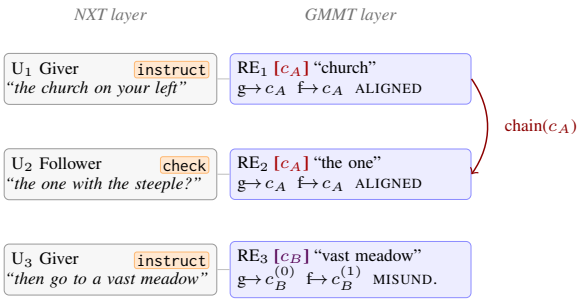

Figure 1: Joined annotation schema for a single dialogue d. **NXT layer**: each utterance carries a dialogue-act tag (orange). **GMMT layer**: each RE carries a unified concept ID (c_A, c_B; a superscript marks the multiplicity instance, e.g. $c_B^{(0)}$ vs $c_B^{(1)}$), the giver's intended landmark ($g \rightarrow$), the follower's interpreted landmark ($f \rightarrow$), and an understanding state in {ALIGNED, PENDING, MISUNDERSTOOD}. REs sharing a concept ID within d form a *reference chain* (red): chain(c_A) is stable in ALIGNED; chain(c_B) is a singleton r_3 where giver-side multiplicity yields silent misalignment despite a shared name. Joined at `(dialogue_id, utterance_id)`.

Discrepancy	REs	Misunderstood	Rate
Identical	8,330	18	0.2%
Lexical	914	13	1.4%
Multiplicity	956	115	12.0%
Existence	2,877	93	3.2%
Total	13,077	239	1.8%

Table 1: Misunderstanding counts and rates by landmark discrepancy type in the GMMT corpus (replication of Li et al., 2025, Table 4).

3.2 OneCommon

The OneCommon corpus (Udagawa and Aizawa, 2019) provides a second referentially asymmetric task: two participants each see a partially overlapping set of dots on a 2D canvas with transformed coordinate frames, and must agree on a shared dot through text dialogue. Among 5,191 annotated dialogues, 1.0% of markables are flagged *ambiguous* (the expression matches multiple dots in the speaker's view) and 0.3% *unidentifiable* (the addressee cannot resolve the reference). These tags align with the giver-side and follower-side friction triggers we define in §4. However, OneCommon lacks the perspectivist dual-interpretation structure of GMMT, so we use it as a cross-corpus comparison point in the Discussion (§8), not a primary evaluation target.

4 Perspectival Frictive States

Existing FPO implementations target tasks with a shared perceptual context (e.g., the Weights Task, where all participants observe the same blocks on the same scale). In that regime, extracting frictive states from the full interaction history h is well-posed, because there is no private perceptual state to miss. When the information structure shifts to an asymmetric partial information setting, such as different maps, different dot configurations, different views of a structure, h is no longer shared: each participant's $h^{(i)}$ includes asymmetric, perspective-specific evidence. As established in Pustejovsky and Zhu (2024), epistemic states are shaped by perceived events, and perceptual asymmetries result in epistemic asymmetries. Frictive states therefore should be defined relative to a participant's private perspective.

Friction functional. The FPO formulation (Pustejovsky and Krishnaswamy, 2026) defines a structured friction functional as $F(h, y) = F^{+}(h, y) - F^{-}(h, y)$, where h is the interaction history, y is a candidate response, F^{+} captures productive friction (information gain from clarification or verification), and F^{-} aggregates four epistemic failure modes: uncertainty/miscalibration, contradiction, hazard, and value conflict.

Frictive state. We define a frictive state as a configuration of the interaction history at turn t in which $F^{-}(h_t, y) > 0$ for at least one plausible response y, a configuration where any continuation of the dialogue carries detectable epistemic risk under the friction functional. This definition makes three commitments. (i) A frictive state is a property of h, not of y: it is what π_{int} reads to decide *whether* to intervene, while π_{gen} produces the friction move *given* that decision. (ii) A frictive state is perspective-relative: when h is participant-specific ($h^{(g)}$, $h^{(f)}$), F^{-} must be evaluated as $F^{-}(h^{(i)}, y)$ for each i. (iii) A frictive state is not necessarily the absence of common ground: both participants may believe they are talking about a parked van (propositionally aligned) while grounding the expression to different physical entities (referentially frictive); the frictive state captures the *risk* this divergence surfaces as task failure.

Failure mode taxonomy. We adopt the four F^{-} failure modes from the FPO formulation above and instantiate each for referentially asymmetric grounding, specifying what it detects under a perspective-bound information horizon.

Uncertainty/miscalibration (within-context referential ambiguity). FPO measures this substate as miscalibration between expressed confidence and actual reliability. In referential grounding, uncertainty arises when a participant takes up a referring expression but cannot uniquely map it to a landmark: multiple candidates remain plausible, and confidence is low relative to true identifiability. This holds for both giver (unaware of ambiguity) and follower (facing competing candidates).

Contradiction (cross-perspective grounding conflict). FPO flags contradiction as logical inconsistency between a response and the interaction history. In referential grounding, it arises when participants commit to incompatible landmark hypotheses: e.g., the giver says "south of the crane bay," while the follower's only viable candidate lies north of it.

Hazard (confident miscommitment risk). FPO uses hazard to capture confident-but-wrong contributions that commit the interaction to costly trajectories. In referential grounding, it arises when a participant confidently acts on a misinterpretation, e.g., the follower routes past the wrong landmark and subsequent turns build on false common ground. The risk is not intent but propagation: high confidence plus undetected misalignment leads error to compound rather than self-correct.

Value Conflict (production underspecification). In FPO, value conflict measures the shift in inferred user intent after committing to y; high values indicate "silent intent fixing." In referential grounding, it arises when an underspecified expression is produced or accepted despite available ambiguity, e.g., the giver says "the vast meadow" with multiple candidates, or the follower commits to one without challenge. In both cases, an interpretation is silently fixed where clarification is warranted.

Table 2 reformulates these same four failure modes as diagnostic questions a single participant can ask from their own information horizon $h^{(i)}$, without omniscient access to the partner's evidence.

The four components split naturally across the two speaker roles: Value Conflict and Uncertainty fire for a speaker at production, Contradiction and Hazard for an addressee at interpretation. Table 2 gives the diagnostic question for each substate, answerable from participant i's information horizon $h^{(i)}$ alone, without omniscient access to the partner's private state. We test whether LLMs can

Substate	Diagnostic from $h^{(i)}$
Unc	Can I commit to a referent, or does my evidence admit multiple candidates?
Contr	Given the spatial claims I attribute to my partner, is there a candidate on my map that satisfies both?
Haz	Am I (or my partner) acting confidently on a grounding that I have reason to doubt?
ValConf	Is the expression I am producing (or accepting) uniquely identifying for me, and if not, have I flagged it?

Table 2: Perspective-bound diagnostic for each frictive substate.

replicate these perspectival distinctions in § 6.

We exclude F^+ (Information Gain) from this taxonomy as F^+ scores potential future *interventions*, not on past interaction history, treating it as a state type would conflate trigger with response. Its role in connecting state detection to the intervention policy π_{int} is discussed in §7.

If the multi-component structure is empirically warranted, the four components should be separately anchored in real dialogue evidence. We test this across §5.1.

5 Corpus Analysis

The taxonomy in § 4 makes two testable predictions: (1) the multi-component structure should be visible in the annotation patterns, different discrepancy types should produce different grounding failure profiles; and (2) when alignment occurs on difficult landmarks, it should be achieved through friction (negotiation) rather than silent capitulation (accommodation). We test both using the GMMT and NXT annotations. The analysis below cross-tabulates three classification axes: landmark *discrepancy type* (Table 1), GMMT *pending subtype* (Table 3), and the four F^- *failure modes* from §4 (Table 2). The first two are observational, the third theoretical; relating them is the core empirical move.

5.1 When is Friction Warranted?

The GMMT annotation pipeline asks GPT-5 (Singh et al., 2025) to answer five binary questions per RE in a fixed order. The cascade terminates as soon as a question produces a "pending" exit; if all five are answered, the system compares each participant's assigned landmark ID to determine *aligned* (same landmark) or *misunderstood* (different landmarks). GMMT reports results using the three-way status

Subtype	Id.	Lex.	Mul.	Ex.	Tot.
quantificational	70.1	44.2	41.1	37.0	1604
speaker_query	4.7	8.3	10.1	14.1	363
unspecified	11.0	19.2	8.0	7.5	308
unaccommodated	5.5	5.8	2.7	4.3	151
ungrounded	8.7	22.4	**38.2**	**37.2**	977
# pending	949	156	487	1811	3403

Table 3: Pending subtype profile by discrepancy type (% within column). Id. identical, Lex. lexical, Mul. multiplicity, Ex. existence. Tot. is the row-total count of pending REs per subtype; # pending is the column-total count per discrepancy type. Bold highlights the *ungrounded* enrichment under multiplicity and existence; identical landmarks are instead dominated by *quantificational*.

only, but each pending exit point corresponds to a distinct sub-mode already recorded in the released `extra.subtype` field: *quantificational* (existence queries), *speaker_query* (speaker-initiated checks), *unspecified* (no addressee uptake), *unaccommodated* (addressee signals comprehension failure), and *ungrounded* (addressee cannot link the expression to a specific landmark). We cross-tabulate these existing labels against the discrepancy type of the landmark each RE references; no re-annotation is required.

Each RE in the corpus references a specific landmark via its concept ID, and each landmark has a *discrepancy type* determined by the map design: *identical* (appears once on both maps), *lexical* (appears on both but with different names), *existence* (appears on one map only), or *multiplicity* (appears twice on one map, once on the other). Table 3 cross-tabulates the pending subtype of each RE against the discrepancy type of the landmark it references. The subtype profiles differ radically across discrepancy types ($\chi^2 = 430.2$, $p < 10^{-84}$, Cramér's $V = 0.205$). Identical-landmark REs are dominated by *quantificational* pending (70.1%): participants confirm whether shared landmarks exist on both maps, a process that resolves quickly, reflecting F^+ information-seeking rather than F^- risk. Multiplicity REs, by contrast, show a sharply elevated rate of *ungrounded* pending (38.2% vs. 8.7% for identical; $V = 0.388$): the addressee takes up the reference but cannot confidently link it to a specific instance, because their map contains only one entity where the speaker's contains two, the follower-side Uncertainty diagnostic of Table 2. Existence REs show a third pattern, enriched for *speaker_query* (14.1%) and *ungrounded* (37.2%):

the speaker proactively checks whether the landmark exists on the partner's map ($V = 0.374$), the giver-side ValConf diagnostic.

These results confirm the hypothesis from § 4: friction-warranting contexts have qualitatively different signatures across discrepancy types, not merely different rates. A friction agent that collapses all pending states into a single "not yet aligned" signal would miss the distinction between a routine existence check and a genuine referent-disambiguation failure.

At the chain level, among the 108 multiplicity–discrepancy chains, 37.0% contain at least one misunderstood referring expression (RE), compared to 0.9% for identical chains. More strikingly, 38.9% of multiplicity chains exhibit non-monotonic state trajectories, reaching an aligned state before reverting to pending or misunderstood, whereas this occurs in only 9.9% of identical chains. This pattern reflects a referent-renegotiation dynamic: alignment on a difficult landmark may be locally achieved yet remains globally unstable, deteriorating as subsequent references reintroduce ambiguity. Such behavior exemplifies the propagation dynamic captured by the Hazard component of F^-, namely trajectories that appear successful but silently diverge.

This pattern is robust across speaker roles. For giver-produced REs, the misunderstanding rate is 13.9% on multiplicity landmarks, compared to 0.2% on identical landmarks; for follower-produced REs, the corresponding rates are 8.4% and 0.4%. The higher rate for givers is consistent with the role-based decomposition in §4: because the giver must choose between multiple candidate instances of a landmark, they are more likely to produce underspecified expressions that trigger misalignment downstream.

5.2 Accommodation versus Negotiation

The GMMT "aligned" state covers two functionally different cases: *negotiated* alignment, where both participants converge through explicit friction moves (checks, clarifications), and *accommodated* alignment, where one participant silently flips their interpretation without any intervening friction move. We distinguish these by joining the GMMT per-RE interpretation trajectories with NXT dialogue act move tags (§ 3.1). Operationally, an aligned RE is classified as *accommodated* if the participant's grounded landmark changed from the previous RE in the same chain and no friction-positive move (`check`, `query_yn`, `query_w`, `clarify`, `align`) occurs within a 3-utterance window.

	Aligned REs		Chains	
Disc. type	**Neg. %**	**Acc. %**	**Non-mon. %**	n
Identical	1.2	0.1	9.9	939
Lexical	4.6	0.7	29.1	79
Multiplicity	11.9	1.1	38.9	108
Existence	0.5	0.0	11.9	539

Table 4: Alignment mode and chain-level non-monotonicity by discrepancy type. *Neg.* = negotiated (interpretation flip with a friction move in the window); *Acc.* = accommodated (flip without friction move); *Non-mon.* = chain contains aligned→pending or aligned→misunderstood.

Results in Table 4 indicate that non-monotonic trajectories are strongly concentrated on multiplicity chains (38.9% vs. 9.9% for identical; $\chi^2 = 88.6$, $p < 10^{-19}$, $V = 0.231$), confirming that alignment decay is a structural property of referential ambiguity, not a corpus-wide noise floor.

The accommodation/negotiation split yields a largely negative result: silent accommodation is rare across the entire corpus (15 cases, 0.2% of aligned REs; only 4 in multiplicity chains). The dominant pattern is *negotiated* alignment: 11.9% of multiplicity aligned REs involve a referent flip accompanied by an explicit friction move, vs. 1.2% for identical landmarks ($\chi^2 = 37.7$, $p < 10^{-8}$). Early accommodation does not predict later chain decay (Spearman $\rho = -0.05$, $p = 0.58$, $n = 108$).

This negative result is itself informative. MapTask participants *are* effective at deploying friction when interpretations shift: they check, clarify, and renegotiate rather than silently capitulating, and the friction *works* at the RE level. The 39% chain-level decay rate therefore reflects *new* REs in the same chain resurfacing the underlying ambiguity, not failure of the original repair — the bottleneck for friction-aware agents is therefore not learning *to* repair, but learning to track that a resolved ambiguity remains structurally available and may resurface across the chain.

6 LLM Probing: Does Perspective Help Detection?

§5 showed that referential misalignment in GMMT is structured by perspective. In this section, we test whether this perspectival structure is recoverable by an LLM: does restricting the model to a single

participant's view improve detection of referential misalignment relative to omniscient access over both maps?

Data. From GMMT we construct a balanced evaluation set of 200 referring expressions, sampled from a pool of 6,098 candidate REs spanning 12,157 utterances across 51 dialogues:

- *Friction-warranted* ($n = 100$): REs from the *multiplicity* class with *misunderstood* status, or labeled as *ungrounded* under *pending*. We prioritize human-verified dialogues and complete the set via uniform sampling with a fixed seed.
- *Control* ($n = 100$): REs with *identical* discrepancy and *aligned* status, from the same dialogues, controlling for lexical and topical variation.

The 200-item set is sized for high-precision per-condition comparison rather than corpus-wide generalization; the effect sizes we report (F_1 differences > 0.45 in every model) are well outside the 95% confidence intervals at this sample size.

Models. We probe three instruction-tuned LLMs spanning closed and open families, dense and mixture-of-experts architectures, and full- and quantized-precision deployments: **GPT-5** (closed, frontier-scale; OpenAI API), **Qwen3.6-35B-A3B** (open, MoE; vLLM, bf16), and **Llama-4-Scout-17B-16E** (open, MoE; vLLM, w4a16 quantized). Qwen reasoning is disabled to keep outputs short and parseable.

Conditions. Each model is queried under three binary yes/no conditions, each tailored to the epistemic role it instantiates:

C1 – OMNISCIENT. Both maps + dialogue prefix. Question: do giver and follower refer to the same physical landmark? Gold positive iff `misunderstood`.

C2 – GIVER. Only the giver's map + dialogue prefix. Question: is the target RE underspecified on the giver's own map? (Production-time ValConf diagnostic.) Gold positive iff the base landmark has multiple instances on the giver's map.

C3 – FOLLOWER. Only the follower's map + dialogue prefix. Question: does the giver's spatial description conflict with the follower's map? (Interpretation-time Contr diagnostic.) Gold positive tracks `misunderstood`.

Evaluated against the human-verified gold from Li et al. (2025), Table 5 shows a consistent pattern across all three models: the omniscient condition performs substantially worse than the giver-perspective condition, *despite having strictly greater access to information*. The giver-versus-omniscient F_1 gap is $+0.69$ for GPT-5, $+0.59$ for Qwen3.6, and $+0.47$ for Llama-4-Scout. In every case the gap is driven by the friction subset: control accuracy under the giver condition matches or exceeds that under the omniscient condition, while friction accuracy collapses from ≥ 0.926 (giver) to ≤ 0.430 (omniscient).

Model	Cond.	Acc.	P	R	F_1	Fr.	Ct.
GPT-5	Omni.	.684	.191	.346	.247	.289	.990
	Giver	**.938**	**.946**	**.926**	**.936**	**.926**	.949
	Foll.	.600	.184	.220	.200	.239	.977
Qwen3.6-35B	Omni.	.625	.289	.512	.370	.290	.960
	Giver	**.955**	**.960**	**.950**	**.955**	**.950**	.960
	Foll.	.725	.167	.070	.098	.490	.960
Llama-4-Scout	Omni.	.437	.277	1.000	.434	.430	.444
	Giver	**.895**	**.838**	**.980**	**.903**	**.980**	.810
	Foll.	.455	.183	.442	.259	.280	.630

Table 5: Per-condition probe performance for three LLMs on 200 REs. Conditions (Cond.): OMNI. shows both maps; GIVER/FOLL. show only that participant's map. P, R, F_1 are computed against the per-condition gold label; Fr. and Ct. report accuracy restricted to the 100-item friction and 100-item control subsets. **Bold** marks the best value per metric within each model.

This is therefore not a uniform degradation but a *specific insensitivity to referential misalignment* that holds across closed and open, dense and MoE, full-precision and quantized models. When the relevant signal — a single landmark with two instances — is embedded within the full set of cross-map relations, the model loses the foregrounding that the giver-only condition provides for free. Restricting the model to one participant's view does not strip information; it surfaces the structurally relevant subset.

Error analysis. Three patterns clarify *how* the omniscient failure manifests at the item level. (i) Llama-4-Scout's omniscient F_1 is propped up by a near-constant "yes" predictor: under C1 it answers yes on 100% of friction items and 55.6% of controls (recall $= 1.000$, false-alarm rate $= 0.541$), yet under C2 the same model recovers an F_1 of 0.903. GPT-5 and Qwen3.6 are more balanced but still positive-biased on friction (60.5%, 72.0%). (ii) The friction pool splits into 43 multiplicity-misunderstood items (gold-positive under C1) and 57 pending-ungrounded items (gold-negative under C1). All three models *conflate* the two under C1 —

Qwen3.6 catches 51% of misunderstood but rules out only 12% of pending-ungrounded; GPT-5 35% and 26% respectively — yet classify *both* subtypes at $\geq 88\%$ under C2. (iii) 35 friction items are misclassified by all three models under C1 (pairwise Jaccard 0.50–0.64); only 1 is misclassified by all three under C2. The omniscient failures are therefore partly item-driven, while the giver-perspective successes are robust across models. Qualitatively, GPT-5 and Qwen3.6 free-text reasons for shared C1 false negatives *enumerate* the giver's two candidate instances (e.g. "*the shared instance (instance 1)*") and then conclude alignment. The omniscient prompt thus elicits the structural information but does not license acting on it; the giver prompt foregrounds it as the question.

7 Perspective-Bound F^- and Repair-Productive Friction

§ 6 established that a perspective-restricted view improves an LLM's ability to *detect* referential misalignment. We now test the policy-relevant question: when F^- is computed per-perspective and a frictive state is declared by *disagreement between perspectival readings*, do the detected states correspond to friction the dialogue subsequently *repairs*—i.e., is the perspectival signal not only detectable but productive?

A note on scope: our experiments probe only F^- on history up to the target turn. The F^+ component is defined as the expected belief-entropy reduction from a hypothetical intervention, which we do not simulate here. RepairScore (below) serves as a retrospective surrogate: it measures whether flagged frictive states were in fact resolved downstream, operationalizing the F^+/F^- coupling that FPO leaves abstract. A prospective computation of F^+, requiring next-turn belief simulation, is deferred to future work.

Perspective-bound contradiction. Following § 4, we instantiate the Contradiction component of F^- for referential grounding as

$$\mathrm{Contr}(h^{(i)}, y) = \mathbf{1}\Big[\mathrm{ref}(y \mid h^{(g)}) \neq \mathrm{ref}(y \mid h^{(f)})\Big] \quad (1)$$

where $\mathrm{ref}(y \mid h^{(i)})$ is the landmark to which expression y resolves under participant i's private information horizon $h^{(i)}$. Unlike the omniscient formulation, which evaluates $\mathrm{Contr}(h, y)$ over a global h and is forced to adjudicate cross-map relations, Eq. 1 reads each perspective independently and declares friction precisely when the two readings diverge. Rather than asking the model a perspective-aligned question, we *compose* two single-perspective resolutions into a frictive-state detector.

Setting	Gemma-4 Omni.	Gemma-4 Persp.	GPT-5 Omni.	GPT-5 Persp.
Repair Count (of 13)	5	**13**	12	**13**
Det.→Rep. rate	.116	**.302**	.279	**.302**

Table 6: Repair-productive friction detection on 200 REs from 51 dialogues. Repair Count counts repair events preceded by a flag (max 13); Det.→Rep. rate is RepairScore) normalized by N_{contr}.

Linking F^- to repair via RepairScore. Detection alone is insufficient: the FPO framework requires that flagged states correspond to recoverable misalignment, not noise. We use the definition Eq. 2 of *RepairScore* in Pustejovsky and Krishnaswamy (2026) to evaluate the performance:

$$\mathrm{RepairScore} = \mathbb{E}\left[\sum_t \mathbf{1}[\mathrm{Contr}(h_t, y_t) > \tau]\, \mathbf{1}_{\mathrm{repair}}(t)\right] \quad (2)$$

where $\mathbf{1}_{\mathrm{repair}}(t)$ marks a gold status transition to `aligned` within the next five REs of the same dialogue. RepairScore operationalizes the F^+/F^- coupling left to future work in the FPO formulation: a frictive-state detector is policy-useful only insofar as the states it flags are ones the dialogue can actually resolve.

Setup. We evaluate on 200 REs from 51 dialogues (43 gold contradictions, 13 gold repairs), using both `Gemma-4-31B` (Team and DeepMind, 2026) and `GPT-5`. We compare an *omniscient* condition (joint map access) against a *perspectival* condition (independent single-map resolutions with disagreement as the frictive signal).

Results. Table 6 shows that the perspectival setting captures *all 13* gold repair events (under our 5-RE window) for both Gemma-4-31B and GPT-5, whereas the omniscient setting captures 5 and 12 respectively; the detection-to-repair rate converges to 0.302 under perspective-bound F^- for both models. The omniscient baseline improves with scale, but the gap does not close: even GPT-5 under omniscient access misses one recoverable misalignment, while perspective-bound F^- recovers all under a substantially smaller open-weights model. The perspectival formulation thus does not simply increase detection; it selectively surfaces *the frictions that matter*—those the dialogue is positioned to repair.

F^- is empirically valid when evaluated within each participant's information horizon, and the resulting signal is repair-productive in a way that omniscient access, even at frontier scale, does not match.

8 Discussion

Annotation refinement 1: subtype decomposition of *pending*. §5.1 shows that subtypes of *pending* distribute differently across discrepancy conditions: multiplicity contexts are dominated by *ungrounded* pending (referent-disambiguation failures), whereas identical-landmark contexts are dominated by *quantificational* pending (routine existence checks). We recommend that future perspectivist grounding annotations expose pending subtypes alongside the three-way status, since this information is already in the annotation cascade.

Annotation refinement 2: accommodation-aware alignment. The *aligned* label similarly conflates *negotiated* alignment (through explicit clarification) and *accommodated* alignment (silent reinterpretation). Although accommodation is rare in MapTask (§5.2), the conflation hides chain-level non-monotonicity such as the 38.9% decay rate in multiplicity chains. We recommend distinguishing these modes by linking per-RE trajectories with dialogue-act annotations, as our (`dialogue_id`, `utterance_id`) join does without re-annotation.

Evidence from a second corpus. The OneCommon dataset provides complementary evidence that referential asymmetry drives failures of common ground. Its continuously varying asymmetry yields a 56% selection failure rate, compared to the 1.8% RE-level misunderstanding rate observed in MapTask. Dialogues containing at least one ambiguous markable also require more turns to complete (5.64 vs. 4.53), reinforcing our finding that referential difficulty increases interactional effort. However, surface-level proxies of friction, such as question frequency or hedging, do not predict task success beyond base rates: failed dialogues are in fact longer and contain more hedging than successful ones ($p < 10^{-19}$). This indicates that the quantity of friction is not itself predictive of outcome. Rather, the key determinant is whether friction is *productive* (resolves ambiguity through targeted clarification) or *unproductive* (hedging that fails to repair misalignment), consistent with the F^+/F^- decomposition. A fully perspectival annotation of OneCommon, applying the grounding cascade separately to each participant's dot configuration, would operationalize both refinements proposed above and remains ongoing work.

9 Conclusion

Collaborative dialogue is not a shared-context problem: when participants hold private perceptual state, friction lives in the gap between their information horizons. We develop the diagnostic and conceptual machinery for perspective-bound FPO, defining frictive states per-perspective with each F^- component computable from one participant's evidence alone. Three findings follow: (i) multiplicity configurations drive most misunderstandings through trajectories that locally succeed but globally decay; (ii) perspective-bound probing outperforms omniscient access at every scale, a structural rather than capacity-bound limitation; (iii) the same signal tracks the misalignments dialogue actually repairs. This cuts against a default in agent design—that misalignment is a context-coverage problem solvable by larger windows or stronger models: more context obscured the signal, and a smaller perspectival model beat an omniscient frontier one. Wherever agents coordinate under private information—multi-agent LLM systems, human-AI teaming, distributed task handoff—friction-aware systems must reason from a perspective, not over all of them.

Future work. Three directions worth pursuing. (i) A fully perspectival re-annotation of OneCommon (Udagawa and Aizawa, 2019), applying the grounding cascade to each participant's dot configuration separately, would test whether the same F^- diagnostics transfer to continuously varying asymmetry. (ii) Transfer across other structurally asymmetric tasks (DPIP, OneCommon) would evaluate the perspective-bound formulation beyond MapTask. (iii) Existing alignment labels often conflate surface coordination with epistemic alignment, obscuring silent accommodation in which participants appear aligned despite persistent interpretive divergence. A perspectival analysis that tracks Theory-of-Mind belief states could expose these latent divergences and provide a more precise account of epistemic alignment.

Acknowledgments

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grants DRL 2019805 and DRL 2454151. The opinions expressed are those of the authors and do not represent views of the NSF.

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.

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.

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

Ibrahim Khalil Khebour, Mariah Bradford, Kenneth Lai, 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.

Nan Li, Albert Gatt, and Massimo Poesio. 2025. Grounded misunderstandings in asymmetric dialogue: A perspectivist annotation scheme for MapTask. *arXiv preprint arXiv:2511.03718*. To appear in LREC 2026.

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.

James Pustejovsky and Nikhil Krishnaswamy. 2025. Frictive policy optimization for LLM agent interactions. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

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

James Pustejovsky and Yifan Zhu. 2024. Lexical event models for multimodal dialogues. In *International Conference on Human-Computer Interaction*, pages 174–192. Springer.

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

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*.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

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

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7120–7127.

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. *arXiv preprint arXiv:2603.05450*.

A Perspectival Probe: Implementation Details

We document the construction, model, prompt, parsing, and metric details of the probe reported in §6.

A.1 Item selection

The 200-item evaluation set is sampled from the GMMT release (Li et al., 2025) under fixed seed 42, in two strata:

- **Friction-warranted ($n = 100$).** REs whose annotated `discrepancy_type` is `multiplicity` *and* (`status=misunderstood` *or* `subtype=ungrounded`). The three human-verified dialogues (`q1ec2`, `q1nc3`, `q1nc7`) are included first; the remainder is filled by uniform sampling without replacement. The resulting subtype composition is 43 multiplicity-misunderstood and 57 pending-ungrounded items.

- **Control ($n = 100$).** REs with `discrepancy_type=identical` and `status=aligned`, drawn uniformly from the dialogues used by the friction stratum.

Gold labels are derived per condition: $g_{C1} = \mathbb{1}[\text{status} = \text{misund.}]$; $g_{C2} = \mathbb{1}[\text{giver's map has} \geq 2 \text{ instances of the base concept}]$; $g_{C3} = g_{C1}$.

A.2 Models and decoding

- **GPT-5**: OpenAI `gpt-5` via the Chat Completions API; `max_completion_tokens=2000`, default temperature. One call per (item, condition), with 60 s back-off on rate-limit responses. The exact model fingerprint is recorded in every prediction row.
- **Qwen3.6-35B-A3B**: `Qwen/Qwen3.6-35B-A3B` served with vLLM $\geq$0.6 in bfloat16, `tensor_parallel_size=2`, `gpu_memory_utilization=0.85`, `max_model_len=32768`. Reasoning is disabled by passing `enable_thinking=False` to the chat template and appending a `/no_think` suffix to the user message; `max_tokens=1024`.
- **Llama-4-Scout-17B-16E**: served with vLLM using `quantization="compressed-tensors"` and otherwise identical settings to Qwen; `max_tokens=200`.

All open-weight decoding is greedy (T=0, `top_p`=1.0).

A.3 Prompts

Each condition wraps the dialogue prefix (transcript up to and including the target RE's utterance) and the relevant landmark map(s) into a single user-turn prompt that closes with the role-specific question and a strict response template:

- **C1 (Omniscient).** "*At this point in the dialogue, are the giver and follower referring to the SAME physical landmark?*" Both maps rendered as JSON.
- **C2 (Giver).** "*Is the referring expression* `[expr]` *underspecified on YOUR map?*" Giver's map only.
- **C3 (Follower).** "*Does the giver's spatial description conflict with YOUR map?*" Follower's map only.

Every prompt closes with the response schema `{"answer":"yes"|"no","reason":"..."}`.

A.4 Output parsing

The first balanced JSON object in the raw output is extracted by a brace-counting scan after stripping fenced code-block markers. The `answer` field is normalized (`yes`/`true` $\mapsto$ True, `no`/`false` $\mapsto$ False). If JSON extraction fails, a lenient fallback searches the raw text for quoted `"yes"` / `"no"` tokens; items still unresolved are dropped from metric computation. The parse-failure rate is below 1% across all nine model$\times$condition cells.

A.5 Metrics

For each (model, condition) we report accuracy, precision, recall, and F_1 against the gold label fixed by the condition above, together with per-subset accuracy on the friction and control strata.