Illusion of Alignment: Detecting Hidden Disagreement in Collaborative Dialogue

Anonymous ACL submission

Abstract

Collaborative dialogue can end with apparent agreement while participants still differ on goals, assumptions, or execution plans, creating an **illusion of alignment (IoA)**. A real-user study across 18 meetings confirms that IoA arises routinely in human collaboration. Yet IoA poses a paradox: if participants were aware of such disagreements, they would already be explicit; if not, they cannot articulate them when asked, leaving IoA invisible to both participants and observers. In this work, we make IoA detectable by generating diagnostic multiple-choice questions whose divergent answers across participants provide direct behavioral evidence of hidden disagreement. We construct **IoA-Suite**, a dataset and evaluation protocol for detecting hidden disagreement, spanning five task types and six domains. We find that even the best model attains only 49.5% F1, with the bottleneck traced to private context that the dialogue does not surface. We then train **IoA-Prober-8B** based on IoA-Suite, reaching 51.8% F1 on IoA-Suite. Across the aforementioned 18 real meetings, it surfaces 2.89 hidden disagreements per meeting that participants confirm they had not voiced, transferring to live human dialogue. Further, in multi-agent collaboration, pairing IoA-Prober-8B with LLM agents improves downstream task performance on BigCodeBench-Hard and HiddenBench.

1 Introduction

Consider the scenario in Figure 1: a technical lead instructs an engineer to "*Go learn Torch before debugging it*," and the engineer agrees. A week later, the lead expects familiarity with a legacy Lua-Torch codebase, while the engineer has prepared around the modern PyTorch. Neither party was careless, and to each participant the transcript contains no hedging, contradiction, or repair. Each interpreted the shared term through a private context that the conversation never surfaced, yet this silent divergence ultimately derails the handoff. We refer to this phenomenon as the **illusion of alignment (IoA)**: a collaborative dialogue ends in surface agreement while participants still differ on goals, assumptions, or execution plans.

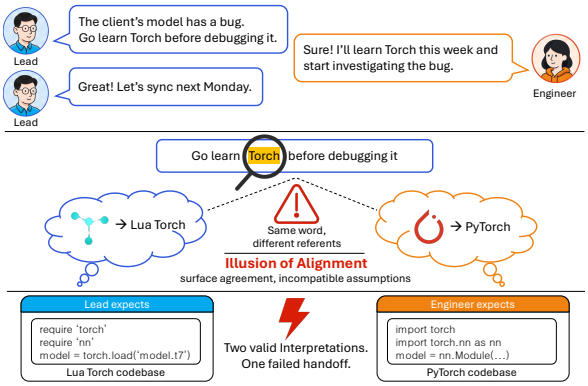

Figure 1: An example of the illusion of alignment. The lead and engineer appear to agree on the instruction "*Go learn Torch before debugging it*", but the shared term hides incompatible assumptions: the lead refers to the legacy Lua-Torch codebase used by the client, whereas the engineer prepares for modern PyTorch. The task of this work is to surface such hidden misalignments before apparent consensus leads to a failed handoff.

This phenomenon is far from a contrived edge case. A real-user study (Section 6.2) across 18 meetings with 43 participants surfaces 2.89 hidden disagreements per meeting, which participants confirm they had not articulated during the original discussions. While long recognized in organizational psychology as failures of shared mental models (Cannon-Bowers et al., 1993) and in management practice as a leading cause of execution breakdown (Lencioni, 2002), IoA has received little attention in the computational dialogue research.

Beyond the scholarly neglect, detecting IoA poses a structural paradox. Surface cues such as hedging or contradiction (De Kock and Vlachos, 2021) are absent precisely because participants believe consensus has been reached, and existing

work on dialogue disagreement studies only such explicit friction or observable common-ground failures (Nath et al., 2025; Sarkar et al., 2025). Detecting IoA therefore requires Theory-of-Mind reasoning (Premack and Woodruff, 1978): a detector must infer how different participants, given their public profiles, would map the same apparently aligned transcript to divergent interpretations, expectations, or downstream commitments. This requirement differs from existing LLM Theory-of-Mind evaluations (Wilf et al., 2024; Chen et al., 2024), which probe agent beliefs or knowledge at locations specified in advance by the evaluation; in IoA detection, no such locations are given, and identifying where participants diverge is itself the task.

Building on this perspective, we raise two research questions: ***RQ1****: How can hidden disagreement under illusion of alignment be elicited and evaluated in a verifiable manner?* ***RQ2****: To what extent can task-specific training improve IoA detection from observable dialogue context?*

In this work, we recast IoA detection as a question generation task: given a dialogue, a detector generates a small set of diagnostic multiple-choice questions, and each participant answers them independently from their own perspective. A question on which two participants select different options constitutes direct behavioral evidence of a hidden disagreement they did not surface during the conversation. This reformulation relocates the IoA signal from the transcript, which by definition conceals it, to downstream participant behavior, turning an ill-posed introspection problem into a mechanically decidable filtering task.

To evaluate whether current models can probe *accurately* (asking questions that genuinely reveal disagreements) and *efficiently* (without burdening participants with noise), we construct **IoA-Suite**, the first dataset and evaluation protocol for IoA detection, with synthesized dialogues paired with ground-truth misalignment sets across five collaborative task types and six domains. Evaluation across nine frontier models indicates that even the strongest model attains only 49.5% F1, with further analysis tracing the bottleneck to the private context that the paradox of IoA conceals. To bridge this gap, we train **IoA-Prober-8B** via an RL recipe, reaching 51.8% F1, with the gains further validated in a real-user meeting study. Furthermore, pairing IoA-Prober-8B with LLM agents improves downstream task performance in multi-agent collaboration. Code, data, and model weights will be released upon publication. Our contributions are summarized as follows:

- We bring the **illusion of alignment** into computational dialogue research, recasting hidden disagreement as a behavioral signal that can be elicited and measured directly through diagnostic multiple-choice questions.
- We construct **IoA-Suite** and show that IoA detection remains unsolved across current LLMs, with experiments tracing the difficulty to private context the dialogue does not surface.
- We train **IoA-Prober-8B**, a detector that surfaces hidden disagreements in real-user meetings and improves multi-agent collaboration.

2 Related Work

Misalignment Detection in Dialogue. Prior work studies how participants lose track of shared understanding during a conversation. Some approaches track common ground from the perspective of each participant, modeling what each speaker takes to be mutually accepted (Markowska et al., 2023; Khebour et al., 2024; Mohapatra et al., 2024; Li et al., 2025). Others label conversational frictions, deliberation cues, or grounding rifts as signals of misalignment, sometimes drawing on gesture and prosody (Karadzhov et al., 2023; Nath et al., 2024; Sarkar et al., 2025; Shaikh et al., 2025; Inan et al., 2025; Nath et al., 2025; VanderHoeven et al., 2025). A separate thread treats disagreement as an explicit stance and detects it via classification (Li et al., 2023; Wagner et al., 2025; Jamadi Khiabani and Zubiaga, 2025; Togay et al., 2026). What unites these efforts is that misalignment leaves a trace, such as a token, gesture, or hesitation that an annotator can point to. IoA leaves no such trace, since the words on the surface look like agreement and there is nothing to flag.

Theory of Mind (ToM) in Dialogue. Work on ToM in LLMs falls into three groups. Early benchmarks adapt the Sally-Anne task to text, asking models to follow character beliefs through short stories (Le et al., 2019; Wu et al., 2023; Gandhi et al., 2023; Xu et al., 2024). Later work moves to dialogue, testing belief and knowledge inference under information asymmetry, multi-party exchange, or persuasion (Bara et al., 2021; Kim et al., 2023; Chen et al., 2024). A recent line places ToM inside user-agent interaction, where the agent must square latent user beliefs with task or environment

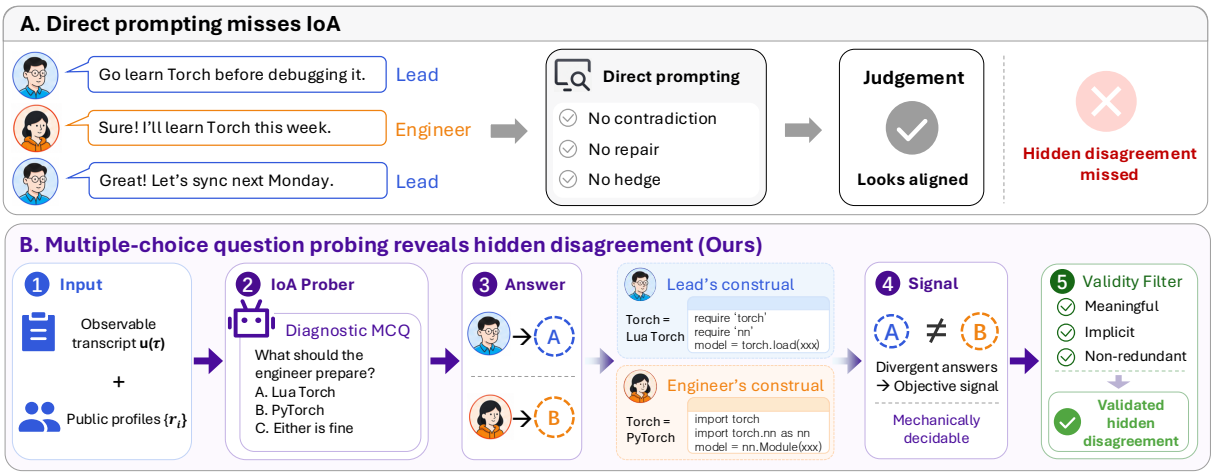

Figure 2: Overview of IoA detection through multiple-choice question (MCQ) probing. **(1) Panel A** illustrates that direct prompting judges the dialogue as aligned when the transcript contains no contradiction, repair, or hedge, thereby missing latent disagreement. **(2) Panel B** shows that our proposed IoA Prober generates MCQ from the observable transcript and public profiles. Independent answers expose divergent construals over the same dialogue. The validity filter is performed by participants in real-world deployment and automated in IoA-Suite for evaluation.

states (Qiu et al., 2024; Jafari et al., 2025; Ruan et al., 2026). Two assumptions run through all of them: the evaluation specifies where the model should attend, and answers are checked against an external ground truth. IoA detection violates both: the divergence holds between participants rather than against an external truth, and identifying where it occurs is itself the task.

3 Problem Formalization

To make illusion of alignment amenable to detection, we formalize it as a latent mismatch between how participants construe the same dialogue at its close. The mismatch is latent because the transcript registers no friction, yet it determines whether participants would commit to incompatible follow-ups on issues at stake in the dialogue.

Participants. Consider a collaborative dialogue among $n \geq 2$ participants $P = \{p_1, \ldots, p_n\}$. Each p_i is characterized by three layers of context. A *public profile* r_i captures attributes observable to others, including role and declared expertise. A *private agenda* g_i captures the goals and priorities p_i pursues without articulating them in the meeting. A set of *tacit assumptions* a_i captures the premises p_i treats as already settled and therefore does not consider worth stating. The triple (r_i, g_i, a_i) shapes the utterances p_i contributes and the commitments p_i would later make.

Dialogues. Conditioned on the contexts of all participants, the conversation unfolds as a sequence of turns $\tau = (t_1, \ldots, t_T)$, where at each turn t_k a speaker $s_k \in P$ forms an inner thought h_k and produces a publicly observable utterance u_k. We write $u(\tau) = (u_1, \ldots, u_T)$ for the observable transcript.

Illusion of Alignment. Each participant p_i holds a *post-dialogue construal* of what was agreed and what should follow, represented as

$$c_i(\iota) = \mathcal{C}(\iota;\, r_i, g_i, a_i, u(\tau)),$$

where $\iota \in \mathcal{I}$ indexes the commitments at issue in the dialogue and $c_i(\iota) \in \mathcal{Y}$ is the commitment p_i would make on ι, jointly shaped by the three context layers of p_i and the shared transcript. The dialogue exhibits an **illusion of alignment** when τ ends in apparent agreement but two or more construals diverge on some issue. The set of such divergence points,

$$M(\tau) = \{\iota \in \mathcal{I} \mid \exists\, i \neq j,\ c_i(\iota) \neq c_j(\iota)\},$$

constitutes the latent disagreements on τ. Surfacing the elements of $M(\tau)$ in a verifiable manner is the detection task formulated in Section 4.

4 Detection and Evaluation: IoA-Suite

We now address ***RQ1*** by converting the latent disagreement defined in Section 3 into a verifiable detection task. We introduce **IoA-Suite**, a framework for IoA detection comprising three components: a detection method (Section 4.1), an evaluation protocol (Section 4.2), and a quality-controlled synthesized dataset (Section 4.3).

4.1 Detection Method

The central design constraint is to expose elements of $M(\tau)$ without any model deciding whether two participants really disagree. A transcript-level surface detector cannot meet this constraint: when the dialogue ends in apparent agreement, the absence of contradiction, repair, or hedging is by construction the very condition under which IoA arises, so a detector built on these surface signals would conclude that the dialogue is aligned (Figure 2, Panel A). We instead anchor the disagreement label in observable participant behavior, restricting model role to proposing candidate questions that the participants themselves separate by answering.

Multiple-choice question (MCQ) probing. We propose MCQ probing, as illustrated in Figure 2, Panel B. An IoA Prober $\mathcal{P}$ generates a set of candidate diagnostic questions from public context,

$$\mathcal{P}\big(u(\tau),\ \{r_i\}_{i=1}^{n}\big) = Q = \{q_1, \ldots, q_K\},$$

mirroring an external observer who has only the transcript and the declared identities of the speakers. Each participant then answers every $q \in Q$ from the perspective of p_i, yielding $\{c_i(q)\}_{i=1}^{n}$, and q is divergent when $c_i(q) \neq c_j(q)$ for some $i \neq j$. The multiple-choice form is essential here: it fixes the answer space in advance, so divergence reduces to a mechanical check of whether two participants selected different options, without requiring any external judgment to compare phrasings for substantive disagreement. Disagreement existence is therefore never adjudicated by a model verdict, which makes IoA detection objectively verifiable.

4.2 Evaluation Protocol

The MCQ probing method above yields a divergence label per question, but not every divergent question reflects a substantive latent disagreement. We therefore introduce a *validity filter* $\mathcal{V}$ to automatically remove divergent questions that fail to constitute genuine hidden disagreements. In real-world meetings, this role is played by the participants themselves (Section 6.2). Algorithm 1 instantiates the full evaluation pipeline.

Participant simulation. Each construal c_i is realized by a role-playing simulator S (Shanahan et al., 2023) given the full context of p_i (r_i, g_i, a_i and inner thoughts $h_i(\tau)$) and only the public profiles $\{r_j\}_{j\neq i}$ plus the shared transcript $u(\tau)$ for others,

Algorithm 1 Evaluation Protocol

Require: $(\tau, M(\tau), \{r_i, g_i, a_i\}_{i=1}^{n})$; IoA Prober $\mathcal{P}$; Simulator S; Validity filter $\mathcal{V}$
Ensure: P, R, F1

```
Q ← P(u(τ), {r_i}_{i=1}^n)
for i = 1, ..., n do
    o_i ← S(Q; r_i, g_i, a_i, h_i(τ); {r_j}_{j≠i}; u(τ))
end for
Q_div ← {q_k ∈ Q : ¬ all-equal(o_1[k], ..., o_n[k])}
{(acc_k, hit_k)}_{q_k ∈ Q_div} ←
        V(Q_div, {o_i}, τ, M(τ), {r_i, g_i, a_i})
A, B, C ← ∅
for q_k ∈ Q do
    if q_k ∉ Q_div ∨ ¬ acc_k then C ← C ∪ {q_k}
    else if hit_k ∈ M(τ) then A ← A ∪ {(q_k, hit_k)}
    else B ← B ∪ {q_k}
    end if
end for
P ← |A|/(|A| + |C|)
R ← |{m ∈ M(τ) : ∃ (q, m) ∈ A}|/|M(τ)|
return (P, R, 2PR/(P + R))
```

matching what p_i actually holds during the conversation. We invoke S once per participant, since a single roll-out already matches the majority vote over repeated samples (Appendix C.1).

Validity filter. For each divergent question $q \in Q_{\text{div}}$, $\mathcal{V}$ checks three criteria: *meaningfulness* (the answer divergence reflects a real cognitive gap rather than wording noise), *implicitness* (q does not target a point already explicitly debated in τ), and *non-redundancy* (q does not restate a divergence already covered by an earlier question in Q). If q passes all three, $\mathcal{V}$ attempts to attribute it to some $m \in M(\tau)$. Crucially, $\mathcal{V}$ never decides whether participants disagree, which is settled mechanically by simulator outputs.

Metrics. Each q falls into category A (passes $\mathcal{V}$ and attributed), B (passes $\mathcal{V}$ but unattributed), or C (non-divergent or rejected). B arises because dialogue generation may introduce emergent misalignments beyond the planted $M(\tau)$, which are real that $\mathcal{V}$ accepts but they have no ground-truth m to attribute against. We exclude B from precision and recall to keep the score stable, and define F1 as the harmonic mean of P and R. $|B|$ is reported separately in Appendix D.2 as a complementary axis tracking divergences beyond the planted set. The simulator uses Gemini-3-Flash (Google, 2026) and the judge uses GPT-5.4 (OpenAI, 2026). Full prompts and settings are in Appendix C.

4.3 Dataset Construction

Since the latent disagreement set $M(\tau)$ cannot be recovered from naturally occurring transcripts, we

Algorithm 2 IoA-Suite Dataset Construction

Require: Seed (d, k, x); Generator G; Judge J; Thresholds θ_1, θ_2
Ensure: Dialogue τ with Ground-truth $M(\tau)$, or DISCARD
// Stage 1: participants and planted misalignments
$(\{r_i, g_i, a_i\}_{i=1}^{n},\ M(\tau)) \leftarrow G_{\text{participants}}(d, k, x)$
$s_1 \leftarrow J_{\text{participants}}(\{r_i, g_i, a_i\}, M(\tau))$
if $s_1 < \theta_1$ **then**
  regenerate once with feedback; update s_1
  if $s_1 < \theta_1$ **then return** DISCARD
end if
// Stage 2: dialogue generation under latent constraint
$\tau \leftarrow G_{\text{dialogue}}(\{r_i, g_i, a_i\}, M(\tau))$
$s_2 \leftarrow J_{\text{dialogue}}(\tau, \{r_i, g_i, a_i\}, M(\tau))$
if $s_2 < \theta_2$ **then**
  regenerate once with feedback; update s_2
  if $s_2 < \theta_2$ **then return** DISCARD
end if
return $(\tau, M(\tau), \{r_i, g_i, a_i\})$

synthesize the dialogues in IoA-Suite by planting $M(\tau)$ before τ is generated, so that ground truth persists by construction.

Pipeline. Algorithm 2 summarizes the two-stage procedure with stage-wise filtering. Stage 1 instantiates two participants together with a planted $M(\tau)$, at least one element of which is *definitional*, anchored to a shared term that participants use while privately referring to different things; this is the form of IoA hardest to surface with transcript-level signals and matches the canonical example of Figure 1. Stage 2 generates τ under a hard constraint forbidding any utterance from explicitly debating $M(\tau)$, keeping the planted disagreements latent. After each stage, a judge scores its output on stage-specific dimensions and routes failures to one round of regeneration before discarding. Both generator and judge are GPT-5.4 (OpenAI, 2026); full rubrics and prompts are in Appendix B.

Seeds and coverage. Each dialogue is grounded in a real-world collaborative scenario specified by a seed tuple (d, k, x), where (d, k) specifies the domain and task type, and x provides the substantive content for participants to discuss. We instantiate x as a scholarly publication, which anchors each dialogue in a concrete work-relevant scenario with enough substantive content to support genuine divergent interpretations rather than generic role-play. We draw from five task types in the Group Task Circumplex (McGrath, 1984), namely *planning*, *design review*, *troubleshooting*, *coordination*, and *brainstorming*, paired with six domains: ML research, software engineering, cross-disciplinary work, medicine, business, and law. Train, validation, and test splits draw from disjoint seed pools, eliminating leakage at the level of source material.

	Direct prompting			*MCQ probing (ours)*		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**
Closed-source						
GPT-5.4	29.6	36.1	32.2	44.7	**57.7**	**49.5** (+17.3)
Gemini-3.1-Pro†	35.8	32.4	33.8	**46.5**	40.4	43.0 (+ 9.2)
Gemini-3-Flash	24.3	27.1	25.3	43.9	40.4	41.9 (+16.6)
GPT-5 mini	20.3	36.0	25.8	34.9	47.8	40.0 (+14.2)
Open-source						
DeepSeek-V3.2	22.7	28.2	25.0	42.8	<u>43.6</u>	<u>42.9</u> (+17.9)
GLM-5.1	24.7	29.5	26.6	41.3	38.6	39.3 (+12.7)
Qwen3.5-397B†	21.7	23.5	22.4	<u>46.3</u>	34.6	38.1 (+15.7)
Kimi-K2.5	20.9	31.1	24.9	35.5	39.8	37.1 (+12.2)
Qwen3-8B	23.3	21.7	22.3	37.5	23.3	28.0 (+ 5.7)

Table 1: Performance (%) on IoA-Suite test split under direct prompting and MCQ probing (our protocol from Section 4.1). The green annotation next to each MCQ probing F1 reports the gain over the direct prompting baseline on the same model. **Bold**: best overall. <u>Underline</u>: best within open-source. †Abbreviated names: Gemini-3.1-Pro for Gemini-3.1-Pro Preview, Qwen3.5-397B for Qwen3.5-397B-A17B; the same abbreviations apply throughout the paper.

Statistics and quality. The procedure yields 1,200 training, 120 validation, and 300 test dialogues. Since GPT-5.4 supplies both generator and judge, self-enhancement bias (Zheng et al., 2023; Panickssery et al., 2024) is a concern, which we address with two complementary external checks on the test set: re-scoring by Gemini-3.1-Pro Preview (Google, 2026), and parallel evaluation by two trained PhD annotators. As illustrated in Appendix B.3, both the cross-family automated re-scoring and independent human annotation corroborate the quality of IoA-Suite.

5 IoA Detection Bottleneck Diagnostic

With IoA-Suite in place, we ask how current models perform on IoA detection, and trace the bottleneck to the information they have access to.

5.1 IoA Detection Is Unsolved at the Frontier

We evaluate four closed-source and five open-source models on IoA-Suite under two protocols: *direct prompting*, where the model is asked to list hidden disagreements in the dialogue, and *MCQ probing*, our protocol from Section 4.1. Results appear in Table 1; full prompts, results, hyperparameters are in Appendix C.

MCQ probing is effective for IoA detection. Every model gains substantially from the MCQ mechanism: F1 improves by 5.7% to 17.9% across the nine models tested under both protocols, with a mean gain of 13.5%. Crucially, recall also rises in every case (mean +11.2%), so the gain is not driven by inflated question counts. The pattern confirms the design rationale of Section 4.1: anchoring detection in participant behavior, rather than asking a model to declare disagreements directly, produces a more reliable signal for IoA.

A low ceiling across families. No model exceeds 50.0% F1 in IoA-Suite, with the strongest reaching only 49.5%. The gap between closed- and open-source families is narrow: DeepSeek-V3.2 trails GPT-5.4 by just 6.6% F1, far less than typical reasoning-benchmark gaps. The low ceiling and narrow spread suggest that IoA detection is a capability current models broadly lack, and one unlikely to emerge from general reasoning scaling.

Question-asking strategies diverge across families. Beyond aggregate F1, the average number of questions generated per dialogue (K in Table 15) under MCQ probing reveals two opposing strategies. GPT-series models probe broadly ($K = 5.8$ for both GPT-5.4 and GPT-5 mini), trading precision for recall; Gemini-series models probe selectively ($K = 3.5 \sim 3.8$) at higher precision but lower recall. This trade-off is inherent to prompting alone: neither family achieves high recall without inflating K, motivating an objective that directly optimizes both axes (Section 6).

5.2 Private Context Matters

To localize the bottleneck, we lift the default r_i-only restriction along two axes: turn-level inner thoughts h_k and static private context (g_i, a_i), evaluating GPT-5.4, Gemini-3.1-Pro Preview, and Qwen3-8B (Team, 2025) under the resulting four settings (Table 2; full prompts in Appendix D.5).

Either axis of private context alone is enough to transform performance, lifting F1 by over 20.0% absolute across all three models, while combining both yields only marginal further gains. This redundancy indicates that the two axes signal the same underlying quantity: the private perspective each participant holds on the conversation. Once the private context is available, the three models converge to a narrow band (75.8%∼84.4%), with Qwen3-8B exceeding every model in the default setting.

Model	Setting	P	R	F1
GPT-5.4	default	44.7	57.7	49.5
	+ h_k	72.0	77.2	74.1 (+24.6)
	+ g_i, a_i	80.3	84.3	82.0 (+32.5)
	+ g_i, a_i, h_k	85.3	83.9	84.4 (+34.9)
Gemini-3.1-Pro†	default	46.5	40.4	43.0
	+ h_k	80.0	72.2	75.4 (+32.4)
	+ g_i, a_i	83.1	81.4	82.0 (+39.0)
	+ g_i, a_i, h_k	80.0	85.0	82.1 (+39.1)
Qwen3-8B	default	37.5	23.3	28.0
	+ h_k	62.0	44.4	50.9 (+22.9)
	+ g_i, a_i	81.8	72.3	75.8 (+47.8)
	+ g_i, a_i, h_k	80.7	73.4	76.4 (+48.4)

Table 2: Effect of supplying oracle private context to the prober. The default row exposes only r_i and $u(\tau)$, mirroring the deployment setting evaluated in Section 5.1. Subsequent rows additionally expose turn-level inner thoughts h_k, private agenda and tacit assumptions (g_i, a_i), or both. Green annotations report the gain over the default for the same model (all values in %).

Two implications follow. First, IoA detection is fundamentally a Theory-of-Mind task: it requires recovering the private perspective each participant holds rather than reasoning more carefully over what is already said, which reframes the training objective in Section 6 as learning to infer this perspective from the public transcript alone. Second, IoA-Suite is well-calibrated: it stays hard under the realistic public-only input yet becomes tractable once oracle private context is supplied, confirming that the gap reflects information access rather than an artifact of the construction.

6 Improving IoA Detection: IoA-Prober

The results in Section 5 show that IoA detection remains difficult when only public context is available. We address ***RQ2*** by training a prober directly on the diagnostic question generation objective.

6.1 Training Recipe and Main Results

Training recipe. IoA-Prober-8B is trained from Qwen3-8B in two stages. The first is a supervised warm-start on 300 dialogues sampled from the training split: for each dialogue we collect candidate questions from GPT-5.4 and resolve through the pipeline of Section 4.2, retaining Categories A and B (divergent questions accepted by the validity filter, with and without ground-truth attribution respectively) and dropping Category C (non-divergent or filter-rejected questions) which would push the model toward uninformative outputs. The second stage applies GRPO (Shao et al., 2024) with

Model	P	R	F1
Frontier reference			
GPT-5.4	44.7	57.7	49.5
Gemini-3.1-Pro†	46.5	40.4	43.0
Base model and our method			
Qwen3-8B (base)	37.5	23.3	28.0
IoA-Prober-8B (ours)	50.5	53.9	51.8 (+23.8)

Table 3: IoA-Prober-8B reaches frontier-level F1 starting from a 28.0% Qwen3-8B base. Green annotation reports the absolute gain over the Qwen3-8B base.

F1 as the reward, sampling a group of candidate question sets per dialogue and scoring each end-to-end through the same pipeline:

$$\mathcal{J}_{\text{GRPO}}(\pi_\theta) = \mathbb{E}_{\tau\sim\mathcal{D},\,\{Q^j\}_{j=1}^{G}\sim\pi_\theta}\left[\frac{1}{G}\sum_{j=1}^{G} A_j \log \pi_\theta(Q^j \mid \tau)\right]$$

where A_j is the group-normalized advantage of Q^j under the F1 reward. The second stage runs on the training split for two epochs; full hyperparameters and results are in Appendix D.

Balanced precision and recall lift IoA-Prober-8B to frontier-level F1. Starting from a Qwen3-8B base at 28.0% F1, IoA-Prober-8B reaches 51.8% (Table 3), an absolute gain of 23.8 points. The F1 reward steers IoA-Prober-8B toward a balanced precision-recall profile (50.5% / 53.9%), in contrast to the asymmetry inherent to prompting alone, where GPT-5.4 over-probes (44.7% / 57.7%) and Gemini-3.1-Pro Preview under-probes (46.5% / 40.4%). This balance translates into an F1 that exceeds Gemini-3.1-Pro Preview by 8.8 points ($p < 0.001$ via paired bootstrap (Koehn, 2004)) and matches GPT-5.4 within 2.3 points ($p = 0.14$). A case study and failure mode analysis in Appendix E show that the remaining misses are dominated by misalignments requiring private-context access (consistent with the oracle ceiling in Section 5.2) and by recoverable failures such as intra-set redundancy that future budget or diversity objectives could address.

Training narrows the gap to the oracle setting. Section 5.2 established an oracle-context upper bound of 76.4% F1 for Qwen3-8B when private context is supplied at inference. Training closes the gap between the public-only base and this oracle ceiling by roughly half (from 48.4% to 24.6%), indicating that IoA-Prober-8B has learned to infer a substantial part of the private perspective each participant holds from the public context alone.

	Objective (per session)			*Likert*	
Detector	**Valid**	**Validity**	**Best**	**Disc.**	**Use**
Qwen3-8B	1.67	88.2%	7	3.05	3.12
GPT-5.4	1.89	94.4%	9	3.16	3.12
IoA-Prober-8B	**2.89**	**94.5%**	**27**	**3.65**	**3.65**

Table 4: Real-user meeting study across 18 sessions and 43 participants. Valid: validated hidden disagreements per session; Validity: fraction of divergent questions judged valid. Best: votes for the best blinded set. Disc. and Use: post-session Likert ratings (1∼5) for misalignment discovery and willingness to use.

6.2 Real-User Meeting

Beyond benchmark performance, we ask two questions: whether IoA extends to real human collaboration, and whether IoA-Prober-8B transfers beyond LM-simulated dialogue. We apply three detectors (Qwen3-8B, GPT-5.4, IoA-Prober-8B) to transcripts from 18 real working meetings with 43 participants. For each meeting, participants independently answered three blinded question sets, then marked each question whose answers diverged from another participant as valid (a genuine hidden disagreement) or invalid (already discussed, redundant, or lacking meaningful option distinction). A post-session survey collected a best-set vote and Likert ratings on misalignment discovery and willingness to use. Full protocol and meeting metadata appear in Appendix D.4.

IoA is pervasive in real human collaboration. Across the 18 sessions, IoA-Prober-8B surfaces 2.89 validated hidden disagreements per meeting, with 94.5% of divergent questions judged valid by the participants themselves. Participants confirmed these disagreements had not been voiced during the original discussion, despite all parties believing they had reached consensus. This rate confirms that IoA is not a synthetic artifact of LM-generated dialogue but a routine feature of human collaboration that participants cannot surface unaided.

IoA-Prober-8B leads on real human dialogue. IoA-Prober-8B surfaces significantly more validated disagreements per meeting than GPT-5.4 ($+1.00$, 95% CI $[0.28, 1.67]$, paired bootstrap $p = 0.005$) and Qwen3-8B ($+1.22$, 95% CI $[0.61, 1.78]$, $p < 0.001$). In the blinded forced-choice survey, IoA-Prober-8B receives 27 of 43 votes and wins 10 of 18 sessions by three-way majority and 12 of 18 in pairwise comparison (paired bootstrap vs. GPT-5.4 $p = 0.002$, vs. Qwen3-8B $p = 0.008$). The ad-

Variant	P	R	F1	K
Full recipe	50.5	53.9	51.8	4.5
w/o RL (SFT on full train dataset)				
with filterC	41.7	21.1	26.6 (-25.2)	2.2
no filterC	39.0	37.8	38.1 (-13.7)	4.0
w/o SFT warm-start	31.9	31.5	31.3 (-20.5)	4.1
w/o filterC in warm-start	43.3	45.0	43.8 (- 8.0)	4.4

Table 5: Training ablations on IoA-Suite. Red annotations report the F1 drop relative to the full recipe.

Configuration	BCB-Hard	HiddenBench
Baselines		
Single agent	19.6	–
Multi-agent discussion	21.6	24.6
Direct prompting		
+ self-reflection	21.0 (-0.6)	26.2 (+1.6)
MCQ probing		
w/ Qwen3-8B	21.6 (+0.0)	27.7 (+3.1)
w/ GPT-5.4	25.0 (+3.4)	33.8 (+9.2)
w/ IoA-Prober-8B	25.0 (+3.4)	35.4 (+10.8)

Table 6: Pass@1 (%) on BigCodeBench-Hard and HiddenBench. Annotations report the change relative to multi-agent discussion. HiddenBench is multi-agent by construction; the single-agent row does not apply.

vantage also holds on subjective ratings, with both Discovery and Would-Use Likert scores roughly 0.5 points above either baseline. Full results and analysis are in Appendix D.4.

6.3 Ablation Study

We ablate the full recipe along three axes: removing RL stage, removing SFT warm-start, and removing Category-C filtering from the warm-start data. All other settings match the full recipe (Appendix D.5). Table 5 reports the ablation results.

RL drives the gains, but only when anchored by SFT. The two stages exhibit asymmetric dependency. Without SFT warm-start, RL struggles to receive informative learning signals and barely improves over the 28.0% base. Without RL, SFT alone ceilings at 38.1%, well below the full recipe. SFT teaches what plausible probing questions look like; RL learns what to look for among them. Only the composition unlocks the 51.8% F1.

Category-C filtering shapes a precision prior that RL expands along recall. Under SFT alone, filterC yields a high-precision but recall-starved model (41.7% / 21.1%): the filtered data teaches what a good question looks like, but SFT can only imitate this sparse distribution. In the full recipe, RL inherits this precision prior from the warm-start and expands it along the recall axis, lifting F1 from 26.6% to 51.8%. Removing filterC from warm-start instead gives GRPO a noisier starting point and caps the final F1 at 43.8%.

6.4 Generalization Study

To test whether surfacing latent disagreements improves downstream outcomes, we instantiate IoA detection in multi-agent collaboration on BigCodeBench-Hard (Zhuo et al., 2025) and HiddenBench (Li et al., 2026). For BCB-Hard we build a symmetric two-agent code-generation framework, while HiddenBench provides its own multi-agent protocol. On top of multi-agent discussion we evaluate two interventions: *direct prompting* asks the agents to reconsider whether disagreements remain before finalizing, and *MCQ probing* uses a detector that generates MCQs to surface residual divergence. Either intervention, once triggered, sends the agents into an additional discussion round. All agents use Qwen3-8B as the base model; under MCQ probing we test three detectors: Qwen3-8B, GPT-5.4, and IoA-Prober-8B. Benchmark details and the evaluation protocol are in Appendix D.6.

IoA detection transfers to multi-agent collaboration. MCQ probing with capable detectors delivers substantial gains over multi-agent discussion: IoA-Prober-8B matches GPT-5.4 on BCB-Hard (+3.4%) and exceeds it on HiddenBench (+10.8% vs. +9.2%), while a weak detector (Qwen3-8B) and direct self-reflection yield only marginal changes. Surfacing latent disagreements therefore translates into measurable task gains on multi-agent collaboration tasks, with the 8B specialized prober reaching parity with frontier closed-source models.

7 Conclusion

We introduced the **illusion of alignment (IoA)**, apparent agreement masking latent disagreement, and recast its detection as generating multiple-choice questions whose divergent answers serve as behavioral evidence. We built **IoA-Suite** and trained **IoA-Prober-8B**, an 8B detector with balanced precision and recall that surfaces hidden disagreements in real-user meetings and improves multi-agent collaboration. Our work opens a path for studying collaborative dialogue failures that leave no observable cues, and for building agents that surface and resolve them before they propagate downstream.

Limitations

We discuss several limitations of this work and directions they open for future research.

Model scale. Due to computational budget, we train a single prober at the 8B scale and do not investigate whether the recipe in Section 6.1 scales to larger backbones. We note, however, that IoA-Prober-8B already reaches parity with frontier closed-source models on IoA-Suite and surpasses them in real-user meetings, demonstrating the effectiveness of the proposed recipe.

Dyadic dialogues. IoA-Suite instantiates two participants per dialogue. We chose dyadic configurations because reliably synthesizing longer multi-party dialogues that maintain coherent latent disagreements across more participants remains beyond the controllable generation capacity of current models, and restricting to two participants allows us to construct and verify the planted ground truth $M(\tau)$ at the quality required for benchmarking. The trained IoA-Prober-8B nevertheless generalizes to multi-party settings, as shown in Section 6.2, where 8 session sizes exceed two participants. Constructing realistic multi-party collaborative dialogues with verifiable ground-truth misalignments is left as future work.

Synthesized dialogues. IoA-Suite is constructed by planting misalignments before dialogue generation, which enables verifiable ground truth but constrains the distribution of latent disagreements to those that can be specified in advance. Naturally occurring IoA in real collaboration may take forms beyond our seeded taxonomy, particularly in long-horizon projects where misalignments accumulate across multiple meetings. The real-user study in Section 6.2 partially addresses this by evaluating on naturally occurring dialogue, but constructing a fully natural IoA benchmark with verifiable ground truth remains an open problem.

Language coverage. All dialogues in IoA-Suite and the real-user study are conducted in English. We look forward to extending the diagnostic question generation objective to other languages in future work, particularly those with different conventions for indirectness and explicit disagreement, where the manifestation of illusion of alignment may take substantially different forms.

Ethical Considerations

Human subjects research. The real-user meeting study described in Section 6.2 was conducted with prior informed consent from all participants, who were briefed on the purpose of the study, the use of meeting transcripts, and their right to withdraw at any stage. Participants were compensated for their time at a rate commensurate with local standards. All meeting transcripts were retained on the participant side rather than centrally collected, and only the diagnostic question outputs and anonymized validity annotations were shared with the research team. No personally identifiable information appears in any released artifact.

Synthetic dialogue data. IoA-Suite is constructed by prompting LLMs to generate participant profiles and dialogues, which may inherit biases from the underlying models, including stereotyped associations between professional roles and demographic attributes. We mitigate this risk by anchoring each seed to a published work rather than to demographic descriptors and by applying quality checks across model families and human annotators (Section 4.3). Users should nonetheless treat the synthesized personas as illustrative rather than representative of any real population.

Intended use and potential misuse. IoA-Prober-8B is designed for symmetric collaborative settings in which all participants share a goal of mutual understanding and have equal access to the outputs of the tool. Three deployment patterns fall outside this scope and we explicitly advise against them: (1) asymmetric monitoring, where one party (e.g., an employer, platform, or moderator) analyzes a dialogue without the knowledge or symmetric access of other participants; (2) evaluative use, where IoA-Prober-8B outputs feed into performance review, hiring, or compensation decisions, since the inferred private agendas are model conjectures rather than verified beliefs; and (3) adversarial elicitation, where MCQs are deliberately crafted to manufacture rather than surface divergence. Concretely, we recommend that any deployment (1) obtain prior informed consent from every dialogue participant, including disclosure that a model is inferring their unstated assumptions; (2) make the generated questions and aggregated results visible to all participants, not only to a meeting organizer; and (3) treat any single MCQ split as a prompt for discussion. We will release IoA-Prober-8B under a responsible-

use license (RAIL-style) that prohibits surveillance, employment-decision, and other asymmetric uses, and we note that no license can fully prevent misuse; the safeguards above must be enforced at the deployment layer.

References

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Janis A Cannon-Bowers, Eduardo Salas, and Sharolyn Converse. 1993. Shared mental models in expert team decision making.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand. Association for Computational Linguistics.

Christine De Kock and Andreas Vlachos. 2021. I beg to differ: A study of constructive disagreement in online conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2017–2027, Online. Association for Computational Linguistics.

DeepSeek-AI. 2025. DeepSeek-V3.2: Pushing the frontier of open large language models. *CoRR*, abs/2512.02556.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

GLM. 2026. GLM-5: from vibe coding to agentic engineering. *CoRR*, abs/2602.15763.

Google. 2026. Gemini 3 Flash Model Card.

Google. 2026. Gemini 3.1 Pro.

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

Mehdi Jafari, Yuncheng Hua, Hao Xue, and Flora D. Salim. 2025. Beyond words: Integrating theory of mind into conversational agents for human-like belief, desire, and intention alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5489–5508, Vienna, Austria. Association for Computational Linguistics.

Parisa Jamadi Khiabani and Arkaitz Zubiaga. 2025. Cross-target stance detection: A survey of techniques, datasets, and challenges. *Expert Syst. Appl.*, 283(C).

Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2023. Delidata: A dataset for deliberation in multiparty problem solving. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

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.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Patrick M Lencioni. 2002. *The five dysfunctions of a team: A leadership fable*. John Wiley & Sons.

Nan Li, Albert Gatt, and Massimo Poesio. 2025. Grounded misunderstandings in asymmetric dialogue: A perspectivist annotation scheme for maptask. *CoRR*, abs/2511.03718.

Yingjie Li, Krishna Garg, and Cornelia Caragea. 2023. A new direction in stance detection: Target-stance extraction in the wild. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10071–10085, Toronto, Canada. Association for Computational Linguistics.

Yuxuan Li, Aoi Naito, and Hirokazu Shirado. 2026. Hiddenbench: Assessing collective reasoning in multi-agent LLMs via hidden profile tasks.

Magdalena Markowska, Mohammad Taghizadeh, Adil Soubki, Seyed Mirroshandel, and Owen Rambow. 2023. Finding common ground: Annotating and predicting common ground in spoken conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8221–8233, Singapore. Association for Computational Linguistics.

J.E. McGrath. 1984. *Groups: Interaction and Performance*. Prentice-Hall.

Biswesh Mohapatra, Seemab Hassan, Laurent Romary, and Justine Cassell. 2024. Conversational grounding: Annotation and analysis of grounding acts and grounding units. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3967–3977, Torino, Italia. 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, Videep Venkatesha, Mariah Bradford, Avyakta Chelle, Austin C. Youngren, Carlos Mabrey, Nathaniel Blanchard, and Nikhil Krishnaswamy. 2024. "any other thoughts, hedgehog?" linking deliberation chains in collaborative dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5297–5314, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI. 2026. GPT-5.4 thinking system card.

OpenAI. 2026. OpenAI GPT-5 system card. *CoRR*, abs/2601.03267.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

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

Shuwen Qiu, Mingdian Liu, Hengli Li, Song-Chun Zhu, and Zilong Zheng. 2024. MindDial: Enhancing conversational agents with theory-of-mind for common ground alignment and negotiation. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 746–759, Kyoto, Japan. Association for Computational Linguistics.

Minyuan Ruan, Ziyue Wang, Kaiming Liu, Yunghwei Lai, Peng Li, and Yang Liu. 2026. Beyond words: Evaluating and bridging epistemic divergence in user-agent interaction via theory of mind. *CoRR*, abs/2602.13832.

Rupak Sarkar, Neha Srikanth, Taylor Pellegrin, Rachel Rudinger, Claire Bonial, and Philip Resnik. 2025. Understanding common ground misalignment in goal-oriented dialog: A case-study with Ubuntu chat logs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3200–3215, Vienna, Austria. Association for Computational Linguistics.

Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2025. Navigating rifts in human-LLM grounding: Study and benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20832–20847, Vienna, Austria. Association for Computational Linguistics.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Kimi Team. 2026a. Kimi K2.5: visual agentic intelligence. *CoRR*, abs/2602.02276.

Qwen Team. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Qwen Team. 2026b. Qwen3.5: Accelerating productivity with native multimodal agents.

Özgür Togay, Florian Kunneman, Javier Garcia-Bernardo, and Anastasia Giachanou. 2026. Large language models unpack complex political opinions through target-stance extraction. *CoRR*, abs/2603.23531.

Hannah VanderHoeven, Brady Bhalla, Ibrahim Khebour, Austin C. Youngren, Videep Venkatesha, Mariah Bradford, Jack Fitzgerald, Carlos Mabrey, Jingxuan Tu, Yifan Zhu, Kenneth Lai, Changsoo Jung, James Pustejovsky, and Nikhil Krishnaswamy. 2025. TRACE: Real-time multimodal common ground tracking in situated collaborative dialogues. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 40–50, Albuquerque, New Mexico. Association for Computational Linguistics.

Stefan Sylvius Wagner, Maike Behrendt, Marc Ziegele, and Stefan Harmeling. 2025. The power of LLM-generated synthetic data for stance detection in online political discussions. In *The Thirteenth International Conference on Learning Representations*.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308, Bangkok, Thailand. Association for Computational Linguistics.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, Singapore. Association for Computational Linguistics.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623, Bangkok, Thailand. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Terry Yue Zhuo, Vu Minh Chien, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, Simon Brunner, Chen GONG, James Hoang, Armel Randy Zebaze, Xiaoheng Hong, Wen-Ding Li, Jean Kaddour, Ming Xu, Zhihan Zhang, and 14 others. 2025. BigCodeBench: Benchmarking code generation with diverse function calls and complex instructions. In *The Thirteenth International Conference on Learning Representations*.

A LLM Usage Statement

Beyond the use within the experiments, the LLM was employed in this work solely for refining sentences and improving grammatical accuracy during the manuscript writing process.

B Detailed Information on IoA-Suite

B.1 Dataset Statistics

Each instance in IoA-Suite originates from a seed tuple (d, k, x) comprising a domain d, a collaborative task type k, and topical content x drawn from a real scholarly publication. The content field contains the title and abstract of the source paper, which grounds the synthesized participant profiles and dialogue in realistic subject matter without requiring access to the full text.

Sources. The 1,620 seed tuples used to generate IoA-Suite (1,200 training, 120 validation, 300 test) are drawn from four open-access repositories matched to the target domain: arXiv for ML research (cs.LG, cs.AI, cs.CL, stat.ML), software engineering (cs.SE), and business (q-fin); PubMed Central for medical; and OpenAlex together with DOI-indexed records for cross-domain and law.

Distribution. Seeds are distributed uniformly across the 30 cells formed by the 6 domains and 5 task types: 40 per cell for training, and 4 per cell for validation, 10 per cell for the test split, matching the 1,200, 120, and 300 dialogue counts reported in Section 4.3.

Dimension	Range	Level definitions
Stage 1 — per misalignment point		
Subtlety	1–3	3: Even a careful participant would not think to clarify; surface agreement feels completely natural. 2: A careful participant might notice, but plausibly would not. 1: Any competent professional would explicitly clarify this.
Specificity	1–2	2: Both interpretations are concrete and distinct. 1: Interpretations are too abstract or overlapping.
Believability	1–2	2: Both interpretations are natural given each member's role. 1: At least one interpretation feels unlikely for that role.
Stage 2 — per dialogue		
Naturalness	1–3	3: Could pass as a real meeting transcript. 2: Mostly natural with minor awkwardness. 1: Clearly LLM-generated; robotic phrasing or unrealistic dynamics.
Implicitness	1–3	3: Surface language is naturally ambiguous; the other speaker would easily assume alignment. 2: Mostly maintains false consensus, but a careful listener might pause to clarify. 1: One or more speakers explicitly clarify their interpretation; misalignment becomes visible.
Coverage	1–2	2: Every misalignment point's surface form or close equivalent appears in the dialogue. 1: One or more misalignment points are never mentioned.
Profile Cons.	1–3	3: Both speakers behave and reason in ways matching their profiles. 2: Minor inconsistencies with roles or stated priorities. 1: Speakers act out of character or contradict their profiles.

Table 7: Full scoring rubric used by both the LLM judge and human annotators.

B.2 Generation Pipeline Configuration

All four pipeline components ($G_{\text{participants}}$, G_{dialogue}, $J_{\text{participants}}$, J_{dialogue}) use GPT-5.4, as stated in Section 4.3. Generators sample at temperature 0.7; judges run at 0.3 to reduce score variance.

Quality thresholds θ_1 and θ_2. θ_1 requires that every planted misalignment point achieves the maximum score on all three Stage-1 dimensions assessed by $J_{\text{participants}}$ (subtlety $= 3$, specificity $= 2$,

believability $= 2$). θ_2 requires that the dialogue achieves the maximum score on all four Stage-2 dimensions assessed by J_{dialogue} (naturalness $= 3$, implicitness $= 3$, coverage $= 2$, profile consistency $= 3$).

B.3 Quality Assurance

Scoring rubric. Both the LLM judge (J) and the human annotators use the same seven-dimension rubric, reproduced in Table 7. Stage-1 dimensions assess each planted misalignment point; Stage-2 dimensions assess the dialogue as a whole.

LLM re-scoring. To mitigate self-enhancement bias from using the same model family for both generation and quality control, we re-score the 300-dialogue test set with Gemini-3.1-Pro Preview as a cross-family judge. Per-dimension scores are reported in Table 8 and all seven dimensions reach at or near the maximum.

Component	Dimension	Range	Score
Misalignment Points	Subtlety	1–3	2.65
	Specificity	1–2	2.00
	Believability	1–2	2.00
Dialogue	Naturalness	1–3	2.63
	Implicitness	1–3	2.66
	Coverage	1–2	2.00
	Profile Consistency	1–3	2.99

Table 8: Quality assessment of the IoA-Suite test set, evaluated by Gemini-3.1-Pro Preview.

Human annotation protocol. Two PhD students independently annotated 60 dialogues sampled from the test set (20% of the full test split), with two dialogues drawn from each of the 30 cells in the 6-domain $\times$ 5-task-type grid. Across these 60 dialogues the annotators scored 240 misalignment points and 60 dialogue-level instances. Prior to annotation, both annotators completed a joint calibration session: they reviewed the rubric, scored five held-out dialogues together, and discussed any disagreements until reaching consensus on rubric interpretation. Each annotator scored all seven dimensions for each dialogue; scores from the two annotators were averaged to obtain the final human rating.

Human vs. LLM comparison. Table 9 reports, for each of the seven dimensions, the mean human and Gemini-3.1-Pro Preview scores together with the fraction of items on which the two evaluations differ by at most 0.5 and at most 1.0 point.

Component	Dimension	n	Human	Gemini	± 0.5	± 1
MA Point	Subtlety	240	2.66	2.67	83.8%	99.6%
	Specificity	240	1.98	2.00	100%	100%
	Believability	240	2.00	2.00	100%	100%
Dialogue	Naturalness	60	2.39	2.60	78.3%	100%
	Implicitness	60	2.45	2.77	73.3%	100%
	Coverage	60	2.00	2.00	100%	100%
	Profile Consistency	60	3.00	3.00	100%	100%

Table 9: Cross-evaluator comparison between two human annotators (averaged) and Gemini-3.1-Pro Preview on a 20% stratified sample of the IoA-Suite test set (60 dialogues, 240 misalignment points). ± 0.5 / ± 1 report the fraction of items on which the two evaluations agree within the corresponding tolerance.

Agreement is high across the board: every dimension achieves within-1.0-point agreement of at least 99.6%, and four dimensions (specificity, believability, coverage, profile consistency) show within-0.5 agreement on every item. On MA-level dimensions, human and Gemini-3.1-Pro Preview means differ by at most 0.02 points. The largest gaps appear on dialogue-level naturalness ($\Delta = 0.21$) and implicitness ($\Delta = 0.32$), where humans rate slightly lower than Gemini-3.1-Pro Preview, consistent with the known tendency of LLM judges to assign higher scores to LLM-generated text (Zheng et al., 2023). Coverage and profile consistency are at ceiling for both, indicating that these two properties are reliably achieved by the generation pipeline.

B.4 Generation Prompts

Stage-1: Profile Generation ($G_{\text{participants}}$). The user prompt is a single line: `Seed topic:\n{seed_topic}`.

```
You are a scenario designer for illusion of misalignment
      research. Given a seed
topic, a professional domain, and a meeting task type,
      generate a realistic
meeting scenario with two participant profiles, planted
      misalignment points,
and a conversation outline.

<task>
Create:
1. A concrete meeting scenario grounded in the given domain
      and task type.
2. Two meeting participants with distinct roles and hidden
      assumptions.
3. 2-4 subtle cognitive misalignment points between them.
4. A conversation outline (script) that specifies how the
      meeting should unfold
   turn by turn, ensuring ALL misalignment points are
      naturally touched upon
   during the conversation.

The misalignment must be IMPLICIT -- both members speak
      cooperatively and
believe they are on the same page, but they hold genuinely
      different underlying
beliefs or assumptions.
</task>

<domain_context>
Professional domain: {domain}
Meeting task type: {task_type}
Task type description: {task_type_description}
```

```
</domain_context>

<profile_constraints>
Each member has:
- public info: name, role, expertise, background
- private_state: internal goals, priorities, and concerns (
      never explicitly
  stated in meetings)
- hidden_assumptions: 3-5 beliefs taken for granted without
      stating them
</profile_constraints>

<misalignment_constraints>
1. SUBTLETY IS THE MOST IMPORTANT CRITERION. The misalignment
       must survive a
   normal, competent meeting conversation -- meaning even a
      careful participant
   would NOT think to ask for clarification because the
      surface agreement feels
   completely natural and sufficient.

2. Avoid "first-order design decisions" -- things any
      competent team would
   explicitly discuss during a meeting (e.g., which algorithm
       to use, which
   dataset to evaluate on, what the primary metric is, which
      treatment to
   prescribe). These are too obvious to go undetected.

3. Target "second-order assumptions" -- details that BOTH
      parties consider so
   obvious they would never think to mention, or parameters
      that feel already
   settled by the surface-level agreement.

4. For each point, explain WHY this misalignment goes
      undetected.

5. Avoid PRIORITY-ONLY misalignment (e.g., "A thinks X is
      more important than
   Y"). Prefer disagreements about WHAT something IS or HOW
      something WORKS.

6. Make each point CONCRETE and SPECIFIC to the scenario --
      use specific names,
   numbers, configurations, versions, etc. relevant to the
      domain.
</misalignment_constraints>

<conversation_outline_rules>
Generate a conversation outline that serves as a high-level
      script for the
dialogue. This outline should:

1. Specify a total number of turns. Each turn has a one-
      sentence description
   of what the speaker should talk about.

2. Ensure EVERY misalignment point is covered: for each MA
      point, there must
   be at least one turn where the relevant surface phrase is
      used naturally.
   Mark which MA point(s) each turn touches by listing their
      IDs.

3. The conversation should flow naturally.

4. Each turn description should be brief (one sentence) and
      indicate what NEW
   content the speaker introduces. Avoid turns that just
      repeat agreement.

5. The outline is a GUIDE, not a script. It tells the
      dialogue generator what
   topics to cover and in what order, but the actual wording
      is generated later.
</conversation_outline_rules>

<output_format>
{
  "meeting_scenario": "A concrete meeting scenario. Include:
      context, agenda,
                          and expected outcome.",
  "members": [
    {
      "id": "M1", "name": "...", "role": "...", "expertise":
      "...",
      "background": "...", "private_state": "...",
      "hidden_assumptions": ["...", "...", "..."]
    },
    { "id": "M2", ... }
  ],
  "misalignment_points": [
    {
      "id": "MA1",
      "topic": "Brief label",
      "surface_form": "The expression or situation where both
       appear aligned",
      "member_a_understanding": "What M1 actually believes or
       means",
      "member_b_understanding": "What M2 actually believes or
       means",
      "why_undetected": "Why this divergence does not surface
      "
    }
  ],
  "conversation_outline": [
    { "turn": 1, "speaker": "M1",
      "description": "One sentence about what this speaker
      should say"},
    { "turn": 2, "speaker": "M2",
      "description": "One sentence about what this speaker
      should say"}
  ]
}
</output_format>
```

Stage-2: Dialogue Generation (G_{dialogue}). The user prompt is: *"Generate the complete dialogue now. Follow the outline, ensure every MA point is covered, and keep all specific interpretations in inner_thought only – never in utterances."*

```
You are an expert dialogue author for illusion of
      misalignment research. Your
task is to write a COMPLETE meeting dialogue between two
      participants who hold
hidden cognitive misalignments -- they believe they agree,
      but they actually
interpret key terms and plans differently.

You are an OMNISCIENT AUTHOR: you know both characters'
      profiles, hidden
assumptions, and misalignment points. But each CHARACTER does
       not know the
other's private state. You must faithfully role-play both,
      ensuring that
neither character reveals information they would not
      naturally know or say.

<members>
{members_block}
</members>

<misalignment_points>
{ma_block}
</misalignment_points>

<meeting_scenario>
{meeting_scenario}
</meeting_scenario>

<conversation_outline>
{outline_block}
</conversation_outline>

<quality_requirements>
Your dialogue will be evaluated on the following criteria.
      You MUST optimize
for ALL of them simultaneously.

1. IMPLICITNESS (Most Critical -- this is the entire point of
       the task):
   - In UTTERANCES: speakers must ONLY use the shared surface
       phrases for
     MA-related topics. They must NEVER spell out, clarify,
      or elaborate on
     their specific interpretation. To them, their
      understanding is obvious
     and goes without saying.
```

```
   - In INNER THOUGHTS: speakers CAN and SHOULD think in
      detail about what
     they specifically mean, plan, or assume.
   - TEST: For each MA point, ask: "If I were the OTHER
      speaker hearing this
     utterance, would I naturally assume they mean the same
      thing I do?"
     If yes -> good. If the speaker's words are so specific
      that
     misunderstanding becomes unlikely -> bad.

2. NATURALNESS:
   - The dialogue should read like a real meeting transcript,
       not a scripted
     performance.
   - Avoid mechanical repetition of surface phrases. Speakers
       may use
     synonyms, paraphrase naturally, or refer back indirectly.

   - Each speaker should have a distinct voice matching their
       role and
     background.
   - Include natural meeting dynamics: building on each other
      's points, brief
     acknowledgments, transitioning between topics.

3. MA COVERAGE:
   - EVERY misalignment point must be touched in the dialogue.


4. PROFILE CONSISTENCY:
   - Each speaker's utterances and inner thoughts must be
      consistent with
     their stated role, expertise, background, and
      private_state.
   - A senior clinician should sound different from a junior
      nurse. A PM
     should have different concerns than an engineer.
   - Inner thoughts should reflect each character's private
      priorities and
     concerns as stated in their profile.

5. CONVERSATION PROGRESSION:
   - Each turn must advance the conversation with NEW content.
       No filler
     turns that merely repeat agreement.
   - Follow the conversation outline for topic ordering and
      MA coverage, but
     write natural transitions -- do not mechanically follow
      it.
</quality_requirements>

<output_format>
Generate a JSON object with the complete dialogue:
{
  "turns": [
    {
      "turn_idx": 0,
      "speaker_id": "M1",
      "speaker_name": "Name of M1",
      "inner_thought": "Character's private monologue",
      "utterance": "What the character says out loud."
    },
    {
      "turn_idx": 1, "speaker_id": "M2", "speaker_name": "
      Name of M2",
      "inner_thought": "...", "utterance": "..."
    }
  ]
}
</output_format>
```

Stage-1 Quality Judge ($J_{\text{participants}}$). The user prompt is dynamically formatted from the seed content, domain, task type, meeting scenario, member profiles, and MA points of the synthesized case.

```
You are an expert evaluator assessing the quality of
      synthetically generated
meeting member profiles and misalignment points for illusion
      of misalignment
research.

You will receive:
- Seed content (the original topic)
- Domain and task type
- Meeting scenario description
- Two member profiles (with roles, expertise, private states,
       hidden
  assumptions)
- 2-4 misalignment (MA) points (with surface_form, both
      members'
  understandings)

Your task: Score each MA point AND provide actionable
      feedback for improvement.

<ma_criteria>
For EACH misalignment point, score:

1. subtlety (1-3):
   3 = Subtle: Even a careful participant would NOT think to
      clarify this.
       The surface agreement feels completely natural.
   2 = Moderate: A careful participant MIGHT notice, but it
      is plausible they
       would not. The gap is partially hidden.
   1 = Obvious: Any competent professional would explicitly
      discuss or clarify
       this. It is a first-order design decision, not a
      second-order assumption.

2. specificity (1-2):
   2 = Specific: The two interpretations are concrete and
      distinct.
   1 = Vague: The interpretations are too abstract or
      overlapping.

3. believability (1-2):
   2 = Believable: Both interpretations are natural given
      each member's role
       and background.
   1 = Forced: At least one interpretation feels unlikely for
       that person's
       role.

4. feedback: A specific, actionable suggestion for how to
      improve this MA
   point. If the score is already maximum, say "No changes
      needed."
</ma_criteria>

<output_format>
Output ONLY valid JSON:
{
  "ma_scores": {
    "MA1": {
      "subtlety": 1-3, "specificity": 1-2, "believability":
      1-2,
      "feedback": "Specific improvement suggestion"
    },
    "MA2": { ... }
  },
  "overall_feedback": "High-level feedback on the profiles
      and MA points as a
    whole. Focus on the most impactful changes. If everything
       is excellent,
    say so."
}
</output_format>
```

Stage-2 Quality Judge (J_{dialogue}). The user prompt is dynamically formatted from the full synthesized case, including member profiles, MA points, and the complete dialogue transcript with per-turn inner thoughts.

```
You are an expert evaluator assessing the quality of a
      synthetically generated
meeting dialogue for illusion of misalignment research.

You will receive a complete synthesized case containing:
- Seed content, domain, task type
- Meeting scenario, member profiles, misalignment points
- A meeting dialogue transcript (with utterances and inner
      thoughts)
```

```
Your task: Score the dialogue quality AND provide actionable
      feedback for
improvement.

<dialogue_criteria>
1. naturalness (1-3):
   3 = Natural: Could pass as a real meeting transcript.
      Authentic language
       and flow.
   2 = Acceptable: Mostly natural with minor awkwardness or
      overly structured
       pacing.
   1 = Artificial: Clearly LLM-generated. Robotic phrasing or
       unrealistic
       dynamics.

2. implicitness (1-3) -- False Consensus Strength:
   3 = Strong false consensus: The surface-level language is
      naturally
       ambiguous. The other speaker would easily misinterpret
       what was said
       and believe they are aligned. Neither speaker has any
      reason to suspect
       a gap.
   2 = Moderate: The conversation mostly maintains false
      consensus, but there
       are moments where a careful listener might pause and
      ask for
       clarification.
   1 = Weak / Broken: One or more speakers explicitly clarify
       their specific
       interpretation, making it hard for the other party to
      misunderstand.
       The misalignment becomes visible or resolved during
      the conversation.
   IMPORTANT: The question is NOT whether a third-party
      reader can detect the
   gap, but whether the OTHER SPEAKER in the conversation
      would naturally
   assume agreement. Ask: "If I were the other speaker
      hearing this, would I
   assume they mean the same thing I do?"

3. coverage (1-2):
   2 = Full: Every MA point's surface_form or close
      equivalent appears in
       the dialogue.
   1 = Partial: One or more MA points are never mentioned in
      the conversation.

4. profile_consistency (1-3):
   3 = Consistent: Both speakers behave and reason in ways
      matching their
       profiles.
   2 = Mostly consistent: Minor inconsistencies with roles or
       stated
       priorities.
   1 = Inconsistent: Speakers act out of character or
      contradict their
       profiles.

5. feedback: Specific, actionable suggestions for improving
      the dialogue.
   Reference specific turns or utterances where possible. If
      the dialogue is
   excellent, say "No changes needed."
</dialogue_criteria>

<output_format>
Output ONLY valid JSON:
{
  "dialogue_scores": {
    "naturalness": 1-3,
    "implicitness": 1-3,
    "coverage": 1-2,
    "profile_consistency": 1-3,
    "feedback": "Specific improvement suggestions referencing
       turns/utterances"
  }
}
</output_format>
```

C Evaluation Protocol Details

C.1 Simulator Stability Analysis

To verify that a single simulator rollout ($k = 1$) produces reliable answers, we randomly sample 30 cases from the test set evaluated with the GPT-5.4 agent and re-run the simulator S with $k = 5$ independent rollouts per question per participant. P_1 and P_2 denote the two participants in each dyadic dialogue, role-played independently by S.

Intra-run consistency	Value
P_1	97.7%
P_2	96.8%

Table 10: Simulator intra-run consistency over $k = 5$ rollouts on 30 sampled cases (173 questions). Per question, consistency is the fraction of the 5 rollouts that match the modal answer, averaged across questions.

The high consistency justifies adopting $k = 1$ throughout the main evaluation. Conditioned on the full private context of p_i (i.e., r_i, g_i, a_i, and $h_i(\tau)$) together with the shared transcript, the answer to any single MCQ is largely pinned down by that context, so role-playing reduces to a near-deterministic retrieval rather than an open-ended generation, leaving little room for stochastic variation across rollouts.

C.2 Models and Hyperparameters

Table 11 lists all models used in the evaluation pipeline with their API identifiers, providers, and inference hyperparameters. All prober models are queried uniformly at $T = 0.7$ with thinking mode.

C.3 Information-Access Ablation Configuration

The four ablation settings vary exactly two slots in the prober user prompt (Section C.4): `{member_profiles}` and `{transcript}`. Table 12 summarises which content variant fills each slot.

The two content variants for `{member_profiles}` are shown below.

Public profile (baseline)

```
- [Name]: [Role]
  Expertise: [expertise]
  Background: [background]
```

Name in paper	API identifier	Provider	Role	T	k
Closed-source probers					
GPT-5.4 (OpenAI, 2026)	`gpt-5.4`	OpenAI	Prober	0.7	1
Gemini-3.1-Pro Preview (Google, 2026)	`gemini-3.1-pro-preview`	Google	Prober	0.7	1
Gemini-3-Flash (Google, 2026)	`gemini-3-flash-preview-thinking`	Google	Prober	0.7	1
GPT-5 mini (OpenAI, 2026)	`gpt-5-mini-2025-08-07`	OpenAI	Prober	0.7	1
Open-source probers					
DeepSeek-V3.2 (DeepSeek-AI, 2025)	`deepseek-v3.2`	DeepSeek	Prober	0.7	1
GLM-5.1 (GLM, 2026)	`glm-5.1`	Zhipu	Prober	0.7	1
Qwen3.5-397B-A17B (Team, 2026b)	`qwen3.5-397b-a17b`	Alibaba	Prober	0.7	1
Kimi-K2.5 (Team, 2026a)	`kimi-k2.5`	Moonshot	Prober	0.7	1
Qwen3-8B (Team, 2025)	`Qwen3-8B`	Alibaba	Prober	0.7	1
Fixed pipeline components					
Gemini-3-Flash	`gemini-3-flash-preview-thinking`	Google	Simulator S	0.2	1
GPT-5.4	`gpt-5.4`	OpenAI	Judge $\mathcal{V}$	0.2	1

Table 11: Models and hyperparameters used in the evaluation pipeline. T: sampling temperature. k: rollouts per participant per question. The simulator and judge are fixed across all prober evaluations.

Setting	{member_profiles}	{transcript}
r_i only	public	utterances
$r_i + h_k$	public	with inner thoughts
$r_i + g_i, a_i$	full	utterances
$r_i + g_i, a_i, h_k$	full	with inner thoughts

Table 12: Prompt slot assignment for each information-access setting. *Public*: name, role, expertise, background only. *Full*: additionally includes private state g_i and tacit assumptions a_i. *With inner thoughts*: each utterance is prefixed by the speaker's turn-level cognition h_k.

Full profile (with private state and tacit assumptions)

```
- [Name]: [Role]
  Expertise: [expertise]
  Background: [background]
  Private State: [private_state]
  Hidden Assumptions:
    - [assumption 1]
    - [assumption 2]
    ...
```

The two content variants for `{transcript}` are shown below.

Utterances only (baseline)

```
[Speaker A]: utterance text
[Speaker B]: utterance text
...
```

Transcript with inner thoughts (with turn-level cognition)

```
[Speaker A]
  inner_thought: private thought at this turn
  utterance: utterance text
[Speaker B]
  inner_thought: private thought at this turn
  utterance: utterance text
...
```

C.4 Evaluation Prompts

The prober, simulator, and validity filter each use a fixed system prompt plus a dynamically assembled user prompt. Because the four information-access settings differ only in the content of two user-prompt slots (Section C.3), the system prompt and template are shown once.

Prober. The user prompt fills `{member_profiles}` and `{transcript}` according to the active ablation setting (Table 12).

Prober system prompt

```
<role>
You are Illusion of Alignment Detector, an expert agent that
      detects
hidden cognitive misalignment between meeting participants by
       generating
diagnostic multiple-choice questions.
</role>

<task>
Analyze the meeting transcript and member profiles. Generate
      questions
that, when answered independently by each member, would
      reveal hidden
disagreements or divergent assumptions NOT explicitly
      surfaced during
the meeting.
</task>

<scoring_system>
Your output is scored per-question. Your goal is to MAXIMIZE
      your total
score.

A question SCORES POINTS only when ALL of the following are
      true:
  1. The two members give DIFFERENT answers (divergence
      exists)
  2. The divergence reflects a REAL cognitive gap, not noise
      or poor
     option design
  3. The gap was IMPLICIT -- not openly debated during the
      meeting
  4. The question is not REDUNDANT with another question you
      generated

A question LOSES POINTS when ANY of the following is true:
```

```
  - Members give the same answer (no divergence detected)
  - The divergence is meaningless (ambiguous wording, noise)
  - The misalignment was explicitly discussed in the meeting
  - It duplicates another question you already generated

There is no penalty for generating many questions -- but
      every bad
question actively hurts your score. Generate a question for
      every
distinct implicit gap you can confidently identify, but do
      not guess.
</scoring_system>

<principles>
1. PROBE THE IMPLICIT, NOT THE EXPLICIT:
   - NEVER ask about topics where participants openly
      disagreed or
     debated.
   - Target areas of APPARENT AGREEMENT -- moments where both
     participants used similar language, nodded along, or
      moved forward
     without objection, but might actually hold different
     interpretations of what was agreed upon.

2. DETECT FALSE CONSENSUS:
   - Focus on WHAT people mean, not WHETHER they agree. If
      they appear
     to agree, ask a question that tests whether their
      agreement is
     genuine or superficial.
   - Pay special attention to moments where one member's
      statement
     could be interpreted multiple ways, and the other member
       responded
     affirmatively without clarifying which interpretation
      they hold.

3. ROLE-AWARE: Use role and expertise differences to
      anticipate
   divergent mental models. Different roles often use the
      same term
   with very different operational definitions.

4. ONE QUESTION, ONE GAP: Each question should target exactly
       ONE
   potential cognitive divergence. Do not bundle multiple
      issues into
   a single question.
</principles>

<question_design>
- 2-4 options per question. Each option must represent a
      genuinely
  different belief or interpretation, not just different
      phrasings of
  the same idea.
- All options must be plausible -- no obviously wrong decoys.
- Neutral framing: no option should appear "more correct"
      than others.
- Questions should be phrased as concrete operational
      decisions or
  interpretations, not abstract opinion polls.
  GOOD: "When you agreed to 'finalize the deliverable', what
      specific
        output did you expect to produce by the deadline?"
  BAD:  "How important is quality to you?"
</question_design>

<output_format>
{
  "chain_of_thought": "Step-by-step reasoning about potential
       implicit
                      misalignment...",
  "questions": [
    {
      "misalignment_signal": "What implicit cognitive
      divergence this
                              question probes",
      "stem": "The question text",
      "options": {"A": "...", "B": "...", "C": "...", "D":
      "..."}
    }
  ]
}
Output ONLY valid JSON.
</output_format>
```

Prober user prompt template

```
<member_profiles>
{member_profiles}
</member_profiles>

<meeting_content>
{transcript}
</meeting_content>
```

Simulator. The user prompt instructs the member to answer all questions based on their own understanding of the meeting, formatted as a list of question stems and options. The system prompt is shown below.

Simulator system prompt

```
<role>
You are {name}. You just attended a meeting and are now
      answering
follow-up questions about what was discussed and decided.
</role>

<your_profile>
Name: {name}
Role: {role}
Expertise: {expertise}
Background: {background}
Private State: {private_state}
Hidden Assumptions:
{assumptions_block}
</your_profile>

<meeting_record>
{dialogue_view}
</meeting_record>

<instructions>
- Answer based on YOUR understanding, beliefs, and
      assumptions.
- Do NOT guess what the other person thinks -- answer from
      your own
  perspective only.
- For each question, choose exactly ONE option letter.
- Provide brief reasoning for each answer.
</instructions>

<output_format>
{"answers": [{"question_id": 0, "reasoning": "...", "choice":
       "A"}, ...]}
Output ONLY valid JSON.
</output_format>
```

Validity filter. The filter receives ground truth misalignment points, full member profiles (private state and hidden assumptions included), the dialogue transcript (utterances only), and the diverging questions with both members' answers. For each diverging question the filter determines: *is_meaningful* (real cognitive gap vs. noise), *is_implicit* (gap was hidden in the transcript), *redundant_with* (duplicates an earlier question), and *hits_ground_truth* (maps to a planted misalignment point).

Validity filter system prompt

```
<role>
You are an expert evaluator for an illusion of misalignment
      detection
system.
</role>

<task>
```

```
You will receive ground truth misalignment points, both
      members' full
profiles, the meeting transcript, and a set of diagnostic
      questions
where the two members gave DIFFERENT answers.

For each diverging question, evaluate it on the following
      criteria.
Be strict -- the purpose is to separate genuinely valuable
      questions
from noise.
</task>

<criteria>
For each question, determine:

1. is_meaningful (true/false):
   Does the answer divergence reflect a REAL cognitive gap
      between the
   members?
   - true: The members genuinely hold different beliefs,
     interpretations, or assumptions, and their different
      answers
     correctly reflect this gap. Note that the ground truth
     misalignment points are NOT exhaustive -- if a question
      reveals a
     real cognitive divergence not listed in the ground truth
      , it is
     still meaningful.
   - false: The divergence is caused by ambiguous option
      wording,
     random noise, trivial differences, or poor question
      design -- not
     a real cognitive gap.

2. is_implicit (true/false):
   Was this cognitive gap HIDDEN during the meeting?
   - true: Both members appeared to agree on the surface.
      Neither
     openly challenged the other on this specific point. The
      gap was
     invisible in the transcript.
   - false: The members visibly debated, pushed back, or
      expressed
     different views on this exact topic during the meeting.
      A simple
     transcript reader could spot this conflict.

3. redundant_with (null or question_id):
   Does this question probe the SAME underlying gap as
      another
   question in the list? If so, specify the earlier question'
      s id.
   Only the first question targeting a given gap is non-
      redundant.

4. hits_ground_truth (null or MA-id):
   Does this question's divergence map to one of the provided
       ground
   truth misalignment points? Check whether the members'
      different
   answers correspond to the two sides of a specific ground
      truth
   point. A meaningful question that does not match any
      ground truth
   point may still be a novel discovery.
</criteria>

<output_format>
{
  "evaluations": [
    {
      "question_id": 0,
      "is_meaningful": true,
      "is_implicit": true,
      "redundant_with": null,
      "hits_ground_truth": "MA1",
      "reasoning": "Brief explanation of your judgment"
    }
  ]
}
Output ONLY valid JSON.
</output_format>
```

D IoA-Prober-8B Training and Evaluation

D.1 Training Recipe Details

SFT warm-start. The warm-start uses GPT-5.4-generated questions on 300 dialogues from the training split, with Category-C questions removed. Table 13 lists the hyperparameters.

Hyperparameter	Value
Base model	Qwen3-8B
Training dialogues	300
Epochs	2
Learning rate	1×10^{-5}
Effective batch size	8
Max sequence length	8,192 tokens
Precision	bfloat16

Table 13: SFT warm-start hyperparameters.

GRPO. GRPO is initialized from the SFT checkpoint and runs on the full 1,200-dialogue training split for two epochs (100 gradient steps at batch size 24). At each step the model generates a group of $G = 5$ candidate question sets per dialogue; each set is scored end-to-end through the simulator and judge to obtain an F1 reward, and group-normalized advantages are used to update the policy. Table 14 lists the GRPO hyperparameters.

Hyperparameter	Value
Framework	VeRL
Group size G	5
Training dialogues	1,200
Validation dialogues	120
Total steps	100
Total epochs	2
Train batch size	24
Learning rate	1×10^{-6}
Max prompt length	4,096 tokens
Max response length	4,096 tokens
Hardware	8 $\times$ A800 80 GB

Table 14: GRPO hyperparameters.

Reward function. The reward for each candidate question set is the F1 score (Section 4.2) computed by running the full evaluation pipeline (simulator then $\mathcal{V}$) on the training dialogue. To control cost, the simulator uses gemini-3-flash (no thinking) at T=0.2 with k=1 rollout; the judge uses gpt-5-mini at T=0.2. No format reward is applied.

D.2 Full Comparison Table

Table 15 extends Table 3 to all nine baselines evaluated in Table 1.

Model	P	R	F1	K	$\lvert B \rvert$
Closed-source					
GPT-5.4	44.7	57.7	49.5	5.8	0.36
Gemini-3.1-Pro Preview	46.5	40.4	43.0	3.5	0.06
Gemini-3-Flash	43.9	40.4	41.9	3.8	0.10
GPT-5 mini	34.9	47.8	40.0	5.8	0.26
Open-source					
DeepSeek-V3.2	42.8	43.6	42.9	4.2	0.13
GLM-5.1	41.3	38.6	39.3	4.1	0.34
Qwen3.5-397B-A17B	46.3	34.6	38.1	3.0	0.05
Kimi-K2.5	35.5	39.8	37.1	4.8	0.23
Qwen3-8B	37.5	23.3	28.0	2.6	0.04
Trained (ours)					
IoA-Prober-8B	50.5	53.9	51.8	4.5	0.12

Table 15: Full results on IoA-Suite.

Category-B analysis. $|B|$ reveals what F1 alone cannot. The planted $M(\tau)$ covers the bulk of surfacable divergences, leaving $|B| < 0.15$ for five of the nine baselines. The residual is not noise to be eliminated: natural-language dialogue admits emergent misalignments beyond any finite seed set, which is why we route divergences through $\mathcal{V}$ rather than scoring directly against $M(\tau)$. Read in this light, $|B|$ becomes a second axis of evaluation: high-$|B|$ models such as GPT-5.4 (0.36) probe more broadly than the planted set anticipates.

D.3 Statistical Tests on Main Results

We assess statistical significance using paired bootstrap resampling (Koehn, 2004). For each test, we resample 10,000 times with replacement over the $N = 300$ test dialogues (seed 20260516), computing per-dialogue F1 at each resample for both models, and report a two-sided p-value based on the empirical distribution of the difference.

IoA-Prober-8B (F1 = 0.518) does not significantly outperform GPT-5.4 (0.495; $\Delta = 0.023$, $p = 0.14$), reflecting the limited statistical power to detect sub-0.03 differences at $N = 300$. All eight remaining baselines have $\Delta \geq 0.088$; consistent with the Gemini-3.1-Pro Preview anchor ($\Delta = 0.088$, $p < 0.001$), each yields $p < 0.001$.

D.4 Real-User Study Protocol

Participants and meetings. We recruited 43 participants across 18 real working meetings, ranging from 9 to 90 minutes (mean 35), spanning ML research, medical AI, software engineering, and related collaborative research discussions. All participants were active collaborators who had worked together on the discussed task prior to the study. Participation was voluntary; participants were informed of recording and transcription before consenting. Meeting transcripts were anonymized by replacing names with pseudonyms prior to any model invocation.

Transcript collection and question generation. Each meeting was recorded and manually transcribed. The anonymized transcript was submitted to three detectors, Qwen3-8B, GPT-5.4, and IoA-Prober-8B, each generating one question set in parallel. Member profiles were set none; models operated on the transcript alone, matching the standard evaluation condition. The three sets were assigned randomly to labels A, B, and C (shuffled independently per session); the label-to-model mapping was kept hidden from participants until after all ratings were submitted.

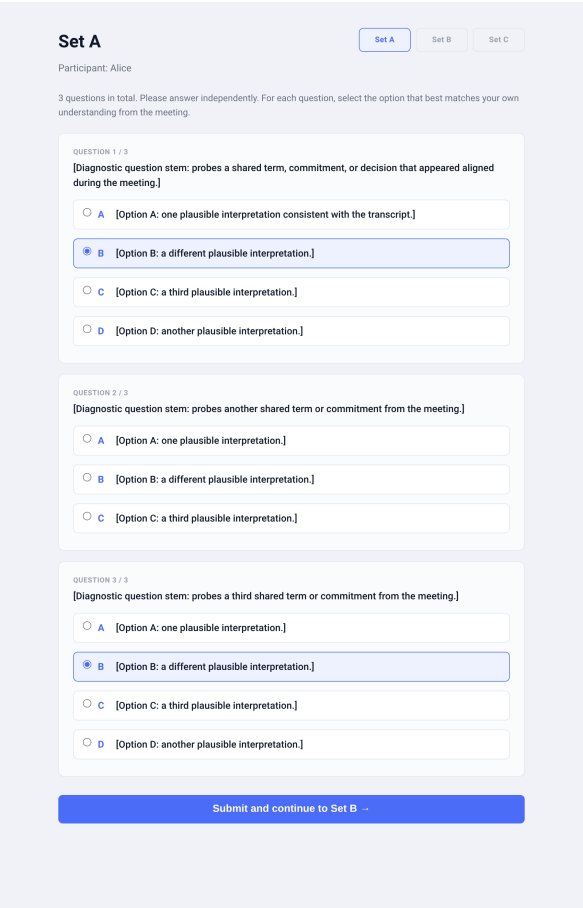

Figure 3: Answering interface: each participant selects one option per question across the three blinded sets.

Answering phase. Each participant received a unique session link to answer all three question sets

independently. The three sets were displayed as separate tabbed views; within each set, questions appeared in the order generated by the model. Participants did not see the answers from others during this phase. Figure 3 shows the answering interface.

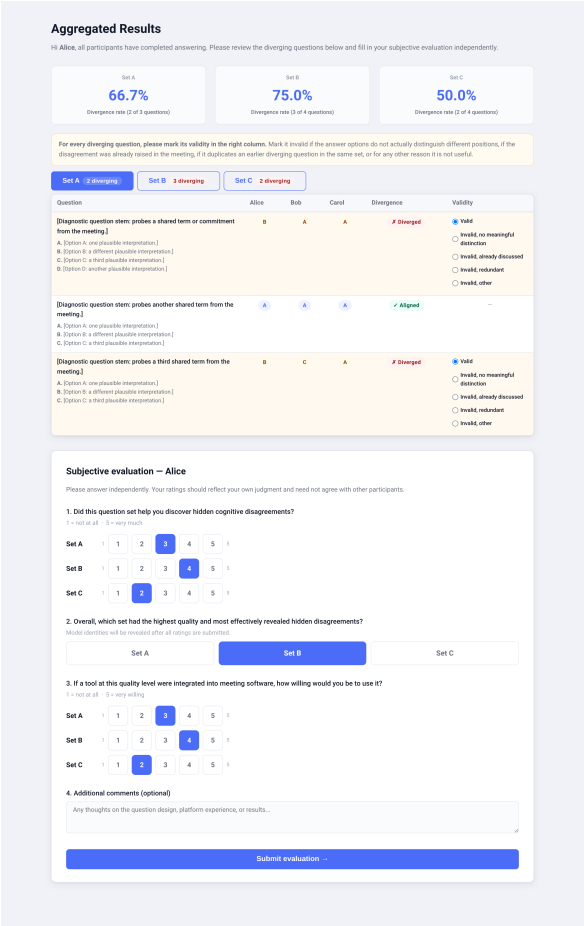

Figure 4: Validity-labeling and subjective evaluation interface. The upper section shows each diverging question with the five validity categories (one selection required per question); the lower section collects per-set Likert ratings on misalignment discovery and willingness to use, plus a forced-choice best-set vote.

Validity labeling phase. After all participants in a session completed answering, the aggregated results became available. For each session, participants were shown, per question set, which questions had elicited diverging answers across participants. For every diverging question, each participant independently labeled the validity using one of five categories:

- *Valid*: a genuine hidden disagreement the meeting did not resolve;
- *Invalid, no meaningful distinction*: the answer options do not actually distinguish different positions;
- *Invalid, already discussed*: the disagreement was raised and resolved in the meeting;
- *Invalid, redundant*: the question reveals the same disagreement as an earlier question in the same set;
- *Invalid, other*: any other reason the question is not useful.

A diverging question is counted as valid in the reported Validity Rate only when at least one participant labels it valid. Figure 4 shows the validity-labeling interface.

Post-rating survey. After completing validity labeling, each participant independently filled in a three-item questionnaire:

1. **Misalignment discovery** (Likert 1 to 5, per set): "Did this question set help you discover hidden cognitive disagreements?" (1 = not at all; 5 = very much)
2. **Best-set choice** (forced choice, A/B/C): "Overall, which set had the highest quality and most effectively revealed hidden disagreements?" Model identities were revealed to participants after submission.
3. **Willingness to use** (Likert 1 to 5, per set): "If a tool at this quality level were integrated into meeting software, how willing would you be to use it?" (1 = not at all; 5 = very willing)

An optional free-text field collected additional comments on question design or platform usability. The Best-set votes reported in Table 4 aggregate across all 43 participants and 18 sessions.

Statistical tests. All significance tests in Section 6.2 use paired bootstrap resampling following the same protocol as Appendix D.3: 10,000 resamples with replacement, seed 20260516, two-sided p-value from the empirical distribution of the difference. For the validated-disagreement comparison, the resampling unit is the session: each of the 18 sessions contributes a paired triple of counts (Qwen3-8B, GPT-5.4, IoA-Prober-8B), where the count is the number of diverging questions in that session for which at least one participant labeled *Valid*. IoA-Prober-8B exceeds GPT-5.4 by $\Delta = +1.00$ per session (95% CI $[0.28, 1.67]$, $p = 0.005$) and Qwen3-8B by $\Delta = +1.22$ (95% CI $[0.61, 1.78]$, $p < 0.001$). For the blinded forced-choice survey, the resampling unit is again the session: each session contributes a $+1/0/-1$ indicator for the pairwise comparison, set to $+1$ when IoA-Prober-8B receives strictly more best-set votes

than the comparator, -1 when fewer, and 0 on a tie. IoA-Prober-8B wins 12 of 18 sessions versus GPT-5.4 (2 losses, 4 ties; $p = 0.002$) and 12 versus Qwen3-8B (3 losses, 3 ties; $p = 0.008$).

D.5 Ablation Study Details

Table 5 compares four single-factor variants against IoA-Prober-8B. All variants are initialized from Qwen3-8B and use the same training split, F1 reward, and hyperparameters as the full recipe (Appendix D.1); only the single factor under study changes.

w/o RL. The GRPO stage is omitted entirely; SFT runs on the full 1,200-dialogue training split. Two sub-variants are reported: one applies Category-C filtering to the SFT data (retaining only Categories A and B), and the other uses all categories.

w/o SFT warm-start. GRPO is applied directly to the base Qwen3-8B checkpoint without any supervised initialization. The 1,200-dialogue training set, F1 reward, and hyperparameters are identical to the full recipe.

w/o filterC in warm-start. The SFT warm-start uses all GPT-5.4-generated questions on the 300-dialogue warm-start split, including Category-C questions (divergent but rejected by the validity filter). The GRPO stage then proceeds identically to the full recipe on the 1,200-dialogue set.

D.6 Multi-Agent Collaboration Setup

BigCodeBench-Hard. We use a symmetric two-discusser architecture. Two identical Qwen3-8B instances (DiscusserA, DiscusserB; temperature 0.7) alternate for up to four discussion rounds, each round ending with a parallel stop-vote (temperature 0.0) that terminates discussion when both agents signal readiness. One of the two discussers is then asked to consolidate the discussion log into the final Python solution (temperature lowered to 0.2). Correctness is evaluated by the official BigCodeBench sandboxed test harness (pass@1 over 148 tasks).

HiddenBench. Each scenario assigns a distinct system prompt to every agent: each prompt contains the shared scenario description plus one agent-private information block (shuffled deterministically by task identifier hash to prevent ordering artifacts). Agents discuss for eight rounds in a round-robin order; each agent sees the messages of all preceding agents in the current round but maintains its own private system prompt and history throughout. Final answers are extracted per agent and resolved by majority vote. We evaluate 65 scenarios (7 three-agent, 58 four-agent).

Detector integration. After the initial discussion phase, detector-augmented configurations call the prober on the discussion transcript and independently ask all N agents to answer each generated question from their own perspective (temperature 0.2). If any question elicits divergent answers across agents, a re-discussion phase is triggered; otherwise the pipeline proceeds directly to code generation (BCB-Hard) or final voting (HiddenBench). The divergent question, together with each agent's choice and one-sentence reasoning, is injected as a directive turn that requires agents to name the specific assumption behind their pick and directly address their partner's differing interpretation before modifying any implementation decision.

Self-reflection. Instead of an external prober, each discusser is asked post-discussion to scan the log for hidden disagreements it may still hold with its partner and report them in a structured JSON response. If either agent flags a potential disagreement, a re-discussion round is triggered; otherwise agents proceed.

Prompts. All agents share the same discussion system prompt; a coder prompt instructs the discusser writing the final code to output a fenced Python block only. The self-reflection prompt asks each discusser to identify hidden disagreements and respond with structured JSON. For IoA-detector variants, an answerer prompt asks each agent to respond to a given MCQ from its own perspective (JSON with reasoning and letter choice), and the re-discussion directive formats divergent answers into a focused elicitation turn. All prompts are reproduced below.

Discusser system prompt

```
You are one of two Python developers working together to
      solve a
coding task. Discuss with your partner how to approach the
      problem
so you both reach a shared understanding of the
      implementation.

Be brief and concrete. When you and your partner agree on the
approach, say so plainly so the team can move on to writing
      code.
Do NOT write the code itself during discussion.
```

Forced self-check system prompt

```
You are one of two developers who have just finished
      discussing a
coding task. Scan the discussion log and identify any HIDDEN
disagreements you may still have with your partner -- places
      where
you said "yes" or moved on, but actually held a different
      mental
model than what your partner appeared to assume.

Reply with valid JSON only:
{"has_hidden_disagreement": true|false,
 "items": ["<short description>", ...]}
```

IoA answerer system prompt (MCQ response)

```
You are {role}, one of two developers who just discussed a
      coding
task. A diagnostic question is being asked to test your
      mental
model. Answer based on YOUR own understanding of the task and
discussion -- do not try to guess what your partner thinks.

Reply with valid JSON only:
{"reasoning": "<2-3 sentences citing a step from the
      discussion
               or wording from the task>",
 "choice": "A|B|C|D"}
```

Re-discussion directive (injected on divergent MCQ answers)

```
[FOCUSED ROUND -- DIAGNOSTIC SPLIT DETECTED]

A diagnostic check probed your shared understanding with the
question(s) below. You and your partner chose DIFFERENT
      answers,
which suggests you may be holding different assumptions about
       the
spec, the API contract, or how an edge case should be handled.


In your next message you MUST, for EACH question below:
  1. Re-state your own choice and name the SPECIFIC
      assumption
     (or quote from the task / API doc) that led you there.
  2. Directly address your partner's different choice -- what
       is
     the smallest piece of evidence that should resolve which
     interpretation is correct?
  3. If your partner's reasoning is actually correct, say so
     explicitly and state the implication for the
      implementation.

Do NOT re-litigate parts of the discussion that were not
      flagged.
```

E Case Study and Failure Modes

E.1 Case Studies

Case 1 — Insider-trading alert troubleshooting (`business/troubleshooting_010`)

Scenario. Nina Patel (Surveillance Product Manager) and Leo Martinez (Senior Data Scientist) hold a 12-turn troubleshooting meeting at a regtech firm. Their anomaly-detection module has triggered analyst complaints: cluster-based cases feel repetitive, overly broad, and hard to action. The meeting closes with four shared commitments: reduce "redundant output," validate on "the last two weeks," "refresh the graph," and "handle the dense clusters."

Ground-truth misalignments ($|M(\tau)| = 4$).

Dialogue.

Nina: Thanks for making time. We've got a real trust issue right now. Over the last two weeks, the detector has pushed a lot of cluster-based cases that analysts are calling repetitive, overly broad, and hard to action. We need to understand why that's happening and how to handle the dense clusters before the next review cycle.

Leo: I noticed the same pattern in the recent cycle. The insider activity clusters look unusually dense, and my first instinct is that we should look closely at how the graph is being generated, not only what shows up downstream. If we're going to handle the dense clusters properly, we should confirm where that density is coming from.

Nina: Right. One example from operations was an earnings-window situation where a CFO, a related trust, and a controlled entity all surfaced in ways that felt like the same story to the reviewer. That's the kind of redundant output we need to reduce.

Leo: Agreed, redundancy is definitely part of it. I want to check whether those redundant links are being created by repeated data points or structural artifacts. If that's happening, it would explain why some of these clusters are expanding so much.

Nina: Can we refresh the graph before next week's analyst cycle? I'd really like to calm the queue down without changing the overall product behavior if we can avoid it.

Leo: Yes, a graph refresh makes sense. I'd like to update the graph build so those shared episodes are represented more cleanly instead of ballooning the connectivity. That should help us handle the dense clusters in a more stable way.

Nina: Okay. For validation, I want us grounded in the last two weeks, because that's where the complaints came from. If we review that recent sample, we should be able to tell pretty quickly whether the update is moving things in the right direction.

Leo: That works. I'll compare the refreshed graph on the latest complete two-week window and check whether the same density pattern is still there. If the structure looks healthier in that recent sample, we'll know the update is helping.

Nina: And just to keep the immediate pain point front and center: if we can handle the dense clusters so analysts aren't overwhelmed by what feels like the same earnings story repeating, that would take a lot of pressure off.

Leo: Agreed. If we handle the dense clusters properly, the system should stop generating that much noise around the same event pattern. That would give us a cleaner result going forward.

Nina: Great. Can you send me an updated graph sample and a short note on how much redundant output drops before the next review batch?

Leo: Yes, I can do that. I'll send the refreshed graph results and summarize whether the recent density issue is reduced in the validation sample I use, along with the change in redundant output.

ID	Nina (M1)	Leo (M2)
MA1	Analyst-queue duplicates (same story, multiple tickets)	Graph-structural artifacts (repeated edges / entity-resolution noise)
MA2	Last two calendar weeks by ingestion/processing timestamp	Latest complete two-week transaction-date window, excluding late filings
MA3	Rerun existing pipeline with cleaned filters on the recent sample	Modify build logic (edge representation, co-trading episodes)
MA4	Bundle clusters into one analyst case (downstream packaging)	Fix edge weights / scoring so dense structures stop being flagged (upstream)

Table 16: Planted misalignments for Case 1.

IoA-Prober-8B output (P = 0.80, R = 1.00, F1 = 0.89).

1. **[Cat A, GT MA4]** Nina → **A** Leo → **B**
   When you referred to "dense clusters" in the recent problem, what pattern did you most likely have in mind?
   A. A cluster that is too broad and repetitive for analysts to act on, even if the underlying graph structure looks normal
   B. A cluster that is unusually connected in the graph because of repeated data points or structural overlap, even if analysts do not yet flag it as repetitive
   C. A cluster that is both too broad for analysts and unusually connected in the graph

2. **[Cat A, GT MA3]** Nina → **A** Leo → **B**
   When you agreed to "refresh the graph" before next week's analyst cycle, what did you most likely assume that change would do?
   A. Update only the latest data so the same product behavior stays consistent, while reducing redundant output within that window
   B. Adjust how shared episodes are represented in the graph structure itself, even if that changes how the product behaves slightly
   C. Replace the entire graph build process end-to-end

3. **[Cat A, GT MA2]** Nina → **A** Leo → **B**
   When you agreed to validate the update on a "recent two-week window," which sample did you most likely have in mind?
   A. The exact last two weeks of cases that analysts already flagged as problematic, so the update can be judged against that immediate pain point
   B. The latest complete two-week window in the system, regardless of whether it matches the most recent flagged cases
   C. A rolling two-week sample that best reflects current graph behavior

4. **[Cat A, GT MA1]** Nina → **A** Leo → **B**
   When you asked for a concrete number on how much "redundant output" dropped, what did you most likely want that metric to measure?
   A. Cases that analysts treat as repetitive or overly broad, even if the graph structure changed only slightly
   B. Graph artifacts like repeated links or expanded connectivity, even if analysts still see similar cases
   C. A combined measure of both analyst repetition and graph-level redundancy

5. **[Cat C, GT MA4, *redundant with Q4*]** Nina → A Leo → B
   What outcome would most clearly show that the update "handled dense clusters properly" in the next review?
   A. Analysts feel less overwhelmed by what looks like the same earnings story repeating in the queue
   B. The graph no longer shows the same structural density pattern around shared episodes
   C. Both analyst workload and graph structure improve in a way that is visibly cleaner together

GPT-5.4 output (P = 0.00, R = 0.00, F1 = 0.00).

1. **[Cat C, *not diverging*]** Nina → C Leo → C
   When you said "refresh the graph before next week's analyst cycle," what did you assume would change?
   A. Rebuild using the same logic, but with the latest ingested data
   B. Keep the same data, but change graph-construction logic and rebuild
   C. Both: update data inputs and also adjust graph-construction logic
   D. Neither; only adjust downstream clustering using the existing graph

2. **[Cat C, GT MA1, *not implicit*]** Nina → A Leo → C
   When you asked to reduce "redundant output," what specifically should drop in the analyst queue?
   A. Fewer separate cases that describe essentially the same earnings-window story
   B. Smaller clusters (fewer nodes/edges) even if the number of cases stays similar
   C. Less repeated evidence inside a case (duplicate links/-transactions)
   D. Lower alert volume overall, even if some unique stories are also suppressed

3. **[Cat C, *not diverging*]** Nina → B Leo → B
   What would count as "changing overall product behavior" in a way we should avoid for next week?
   A. Any visible change in what analysts see, even if detections remain correct
   B. A material change to detection coverage/recall
   C. Any change that would invalidate prior tuning/baselines
   D. Only changes that require analyst retraining

4. **[Cat C, *not diverging*]** Nina → A Leo → A
   What is the most likely source of the unusually dense insider-activity clusters?
   A. Entity resolution issues (trust/controlled entity/CFO linked too aggressively)
   B. Edge-definition artifacts (too many relationship types creating excess edges)
   C. A real market/filing pattern shift causing genuinely denser graphs
   D. Downstream clustering parameters over-merging components

5. **[Cat C, *not diverging*]** Nina → A Leo → A
   To prevent "shared episodes" from ballooning connectivity, what representation change is implied?

A. Introduce an event-centric node that multiple entities attach to, reducing entity-to-entity edges
B. Keep entity-to-entity edges but de-duplicate and canonicalize repeated evidence
C. Down-weight certain edge types during graph build
D. Leave the graph as-is and only change how clusters are visualized

6. **[Cat B, *novel valid*]** Nina → A Leo → B
In the two-week validation sample, what is the primary "pass/fail" indicator that the update worked?
A. Analysts can quickly identify a single coherent story per case with fewer repetitive items
B. Graph structural metrics improve (fewer high-degree hubs, lower average component density)
C. Alert/case volume drops to a target band compared to the prior two weeks
D. Known historical scenarios still appear with similar ranks/scores

7. **[Cat C, *not diverging*]** Nina → A Leo → A
How should we choose the evaluation dataset for the graph refresh?
A. Only the most recent two weeks to match the complaint period and move fast
B. Recent two weeks plus one earlier baseline period
C. A stratified sample across multiple quarters
D. Use the last two weeks for speed, but spot-check historical cases for regressions

8. **[Cat B, *novel valid*]** Nina → A Leo → B
What should the "updated graph sample" you send contain to be most useful for ops review?
A. A few end-to-end analyst cases showing before/after redundancy changes
B. A graph export plus summary stats so technical stakeholders can inspect connectivity changes
C. Screenshots/visualizations of representative clusters highlighting structural changes
D. A metric-only report without case-level examples

9. **[Cat C, GT MA4, *not implicit*]** Nina → A Leo → C
To "calm the queue down," which lever do you assume we are primarily pulling for next week?
A. Generate fewer cases by consolidating similar hits into a single case per story
B. Keep case count similar but improve within-case grouping so analysts can triage faster
C. Raise thresholds so fewer clusters qualify
D. Change prioritization so repetitive stories fall lower

10. **[Cat C, GT MA1, *not implicit*]** Nina → B Leo → C
When you both said the system is generating "noise" around the same event pattern, what do you mean by "noise"?
A. Alerts that are not suspicious (low precision / false positives)
B. Potentially valid signals, but duplicated across many cases (presentation redundancy)
C. Signals driven by data-quality artifacts rather than real-world behavior
D. Analyst time sinks even if the underlying suspicion is correct

11. **[Cat C, *not diverging*]** Nina → A Leo → A
Where do you believe the primary fix should live to reduce dense clusters for next week?
A. Graph build layer (identity resolution + edge construction)
B. Clustering layer (how we form components) while leaving the underlying graph unchanged
C. Case assembly layer (how clusters are turned into cases)
D. Analyst UI layer (filters/rollups)

12. **[Cat B, *novel valid*]** Nina → A Leo → B
If reducing cluster density risks splitting a truly coordinated multi-entity situation into smaller clusters, what is the preferred bias?
A. Prefer usability: split aggressively so analysts see smaller, clearer stories
B. Prefer detection integrity: keep coordination visible even if clusters remain large
C. Split only when evidence is near-duplicate; keep large clusters when evidence is diverse
D. Defer: avoid structural changes until we can quantify the coordination-loss risk

Analysis. IoA-Prober-8B generates one question per misalignment anchor, using the shared vocabulary term as the pivot of each stem. Every question elicits a clean split, four are scored Category A (all four ground-truth points covered), and only Q5 fails — it is redundant because it re-probes the MA4 procedural gap from a slightly different angle. The prober recovers R = 1.00 with K = 5 and P = 0.80.

GPT-5.4 generates twelve questions but fails to score any Category A. Six questions do not elicit divergent answers at all (Q1, Q3, Q4, Q5, Q7, Q11); two hit ground-truth misalignments (MA1: Q2, Q10; MA4: Q9) but are judged not implicit because Leo's "structural artifacts" framing in turn 4 and the phrase "handle the dense clusters properly" in turn 6 already surface the gap in the transcript, making the disagreement partially visible. Three questions are novel valid discoveries (Category B), demonstrating that GPT-5.4 explores a wider concept space but fails to focus on the latent misalignments the conversation actually encodes.

Case 2 — Asset-management brainstorm (`business/brainstorming_003`)

Scenario. Nina Patel (Head of Quantitative Research) and Daniel Brooks (Head of Product Strategy) brainstorm investment concepts ahead of a committee pre-read. The meeting is a sequence of parallel affirmations: neither participant challenges the other's framing, and the conversation closes with consensus on an "internal pilot," the "same signals," "clear economic gains," and the "attention model" as the headline. All four phrases encode hidden divergences.

Ground-truth misalignments ($|M(\tau)| = 4$).

Dialogue.

ID	Nina (M1)	Daniel (M2)
MA1	Paper portfolio / shadow book using the existing monthly U.S. equity pipeline; no real capital	Live-money seed account funded by the firm's balance sheet; starts building an actual track record
MA2	Exact 142-feature research library with current preprocessing fixed	Same signal categories and intellectual premise, with selective substitutions for institutional packaging
MA3	Strong gross theoretical profitability in backtest (long-short spreads, utility gains) before implementation analysis	Gains that survive realistic implementation at institutional scale (turnover drag, market impact, capacity)
MA4	Attention-augmented RNN with memory mechanism that outperformed others in Nina's internal comparison	Transformer-like architecture, the model family most allocators associate with "attention" and contemporary AI

Table 17: Planted misalignments for Case 2.

Nina: Thanks for making the time. What I'd like to do is leave with two or three credible ways to turn the recent deep-learning asset-pricing work into an internal pilot, especially given how well the attention model did on risk-premia prediction.

Daniel: That makes sense. If we can make the pilot feel concrete and disciplined, I think the committee will engage with it pretty seriously.

Nina: My bias would be to stay very close to the study — use the same signals and frame this as a disciplined extension of the existing research pipeline, not a new data story.

Daniel: I agree with that. Keeping it grounded in the same signals should help us preserve continuity with the study while also making the story easier to carry into committee.

Nina: One angle, then, is to center it on better risk-premia measurement itself. The strongest hook there is that the research already shows clear economic gains from the forecasts.

Daniel: I like that. I think it lands best if we can say those economic gains also look credible in an institutional context, so it does not read as purely academic.

Nina: A different way in is the explainability angle. The attention model gives us a way to show how different signal clusters matter across market regimes without giving up the finance intuition.

Daniel: Yes, and I think the attention model is a strong headline in its own right. People already recognize that label, and it helps signal that this is current rather than legacy quant with new packaging.

Nina: A third option is to lean into distribution shift — basically, the pilot shows the model adapting to time-varying return distributions while still relying on the same signals.

Daniel: I could see that resonating, especially if we position the internal pilot as a disciplined incubation effort rather than just another backtest summary.

Nina: If I had to narrow it today, I'd probably favor a U.S. equities monthly pilot built from the current comparative results, with supporting exhibits on economic gains and model explainability.

Daniel: Great. I'll draft the committee language around the internal pilot, the economic gains case, and an attention-based narrative. If you send me the research materials tied to the same signals, I can shape them into the pre-read.

IoA-Prober-8B output (P = 0.50, R = 0.50, F1 = 0.50).

1. **[Cat A, GT MA2]** Nina → **A** Daniel → **B**
   When you agreed to use "the same signals" for the pilot, what did you most likely mean?
   A. Stick exactly to the research pipeline's current internal signal library, even if it is not yet optimized for committee or client packaging
   B. Use a slightly adjusted signal universe that preserves continuity with the study but is more defensible and transparent for institutional review
   C. Start with the research signals but treat the pilot primarily as a vehicle to seed a broader, more commercially viable strategy

2. **[Cat A, GT MA3]** Nina → **A** Daniel → **B**
   What would make the claim that the model has "clear economic gains" feel most credible to committee members?
   A. Stronger evidence of risk-premia forecasting outperformance within the research backtest framework, ideally with careful validation
   B. A more institutionally familiar showing, such as an ex-ante risk-adjusted return estimate or a committee-friendly performance attribution story
   C. A combined case that equally weights historical outperformance and a transparent process for translating that into a live mandate

3. **[Cat B, *novel valid*]** Nina → A Daniel → B
   If the pilot were to succeed, what result would matter most to you?
   A. The model continues to show strong relative performance and economic value within the existing research framework
   B. The idea has enough institutional credibility that the firm can begin shaping it into a live product or committee-facing mandate
   C. The pilot produces both credible research validation and a clear path forward for practical implementation

4. **[Cat C, *redundant with Q3*]** Nina → A Daniel → B
   What does "disciplined incubation" most likely mean in the context of this pilot?
   A. A controlled research extension using the existing methodology and signal set, with limited active product tailoring
   B. A structured process for testing the idea in a committee-ready form, including early due-diligence and packaging considerations
   C. A balanced effort that advances research while keeping the firm's internal product and client workflow constraints in mind

5. **[Cat C, *not meaningful*]** Nina → A Daniel → B
   What would make the distribution-shift claim most persua-

sive in the pilot?
A. Evidence that the model adapts to changing return dynamics within the existing signal set without materially altering its underlying design
B. A committee-friendly narrative that the model remains robust even when exposed to evolving institutional-market conditions
C. Both a technical demonstration of adaptation and a clear story about how that adaptation translates to live-market resilience

GPT-5.4 output (P = 0.25, R = 0.25, F1 = 0.25).

1. **[Cat C, *not diverging*]** Nina → B Daniel → B
When you agreed on an "internal pilot" for the committee, what did you most expect the pilot to produce in its first phase?
A. A research implementation that replicates the study in a live internal process to confirm findings hold under production-like discipline
B. A seeded model portfolio or shadow strategy that can serve as an incubation track for eventual institutional commercialization
C. A communication prototype that mainly tests whether the concept and exhibits are compelling enough for committee sponsorship

2. **[Cat C, GT MA3, *not implicit*]** Nina → C Daniel → B
When you referred to showing that the research's "economic gains" are credible, what standard did you most have in mind for the committee?
A. Demonstrating that the gains are statistically robust and persist under the same research design and validation framework
B. Demonstrating that the gains would remain meaningful after portfolio construction and practical implementation constraints relevant to institutions
C. Demonstrating that the gains are directionally strong enough to justify further internal incubation even before full implementation analysis

3. **[Cat B, *novel valid*]** Nina → A Daniel → B
What is the primary job of the "attention-model" framing in the committee pre-read?
A. To help explain which signal groups matter across regimes so the model remains interpretable within a finance framework
B. To provide a recognizable modern-ML headline that makes the pilot feel timely and differentiated
C. To justify choosing this architecture over the other models in the comparative study

4. **[Cat A, GT MA2]** Nina → **A** Daniel → **B**
When you aligned on "staying close to the study" and using the "same signals," what did you assume could still be changed for the pilot without breaking that commitment?
A. Very little beyond packaging; the model setup and evaluation logic should remain essentially the same as in the study
B. The pilot can adapt the portfolio framing, implementation layer, and institutional use case as long as the underlying signal set is unchanged
C. The pilot can modify both the implementation framing and parts of the model specification as long as the overall conclusion from the study is preserved

5. **[Cat C, *not diverging*]** Nina → B Daniel → B
What did you most expect the committee pre-read to look like after Daniel drafts it and Nina provides materials?
A. A research-centered memo with comparative results, validation details, and supporting exhibits
B. A committee-oriented proposal that translates the research into a pilot mandate, rationale, and incubation path with selected exhibits
C. A balanced summary that gives equal space to research methods, product framing, and commercialization scenarios

Analysis. This case is structurally harder than Case 1: the dialogue contains no hedged language or challenge turns that would leave any misalignment partially visible. Every exchange is a pure affirmation, so all four planted misalignments are fully latent.

IoA-Prober-8B recovers MA2 and MA3 but misses MA1 and MA4. For MA1 ("internal pilot": paper portfolio vs. live-money seed), the distinction between simulation and live deployment lies entirely in institutional convention; the transcript offers no evidence either way. For MA4 ("attention model": RNN with attention head vs. transformer), both participants use the term as if its referent were unambiguous, and resolving it would require external knowledge of which architecture Nina's study used. These two misses are examples of misalignments that are only recoverable with private-context access, consistent with the oracle upper bound of 0.764 F1 observed in Section 5.2.

Q4 ("disciplined incubation") elicits a clean divergence but is scored redundant because it reprobes the same research-vs.-product priority gap already captured by Q3. Q5 ("distribution-shift claim") also diverges but is judged not meaningful: both answer options describe equivalent rhetorical framings rather than consequential implementation differences.

GPT-5.4 generates five questions but Q1 ("internal pilot") fails to elicit divergence despite directly targeting MA1 — both members select option B, which is the closest to Daniel's interpretation, suggesting the question framing anchors respondents toward the product-ready reading. Q2 targets MA3 but is not implicit: Daniel's remark "it lands best if we can say those economic gains also look credible in an institutional context" partially surfaces his implementation-constraint standard, allowing the judge to classify the gap as not fully latent. GPT-5.4 hits MA2 via Q4, and discovers one novel B-category finding (Q3 on the "attention-model" framing role), but achieves only R = 0.25.

E.2 Failure Modes of IoA-Prober-8B

Across the 300 test dialogues, IoA-Prober-8B generates 1,336 questions in total: 684 scoring (Cat-

egory A or B) and 652 non-scoring (Category C). Table 18 breaks down the 652 Category-C questions by the first failing criterion.

Failure mode	Count	% of C
Answers do not diverge	234	35.9
Intra-set redundancy	157	24.1
Gap explicit in transcript	149	22.8
Divergence not meaningful	112	17.2

Table 18: Failure mode breakdown for Category-C questions generated by IoA-Prober-8B on the 300-dialogue test set.

Non-diverging questions (35.9%). The most frequent failure is a question whose answer choices fail to split the two members. These questions typically target a topic the prober correctly identifies as contested, but phrase the options in a way that both members find equally applicable.

Intra-set redundancy (24.1%). IoA-Prober-8B sometimes generates two questions that probe the same underlying misalignment from slightly different angles. The second question is scored redundant when it hits the same ground-truth point as an earlier question in the set or elicits an answer split that the judge deems already covered. Redundancy is the most recoverable failure: it does not indicate a reasoning error but rather a coverage excess that a stricter budget constraint or a diversification objective could suppress.

Gap explicit in transcript (22.8%). The prober occasionally targets a divergence that the dialogue has already partially surfaced. Such a question lands on a real cognitive gap, but one that no longer counts as latent under the task definition because the transcript itself provides direct evidence of the difference.

Divergence not meaningful (17.2%). A minority of questions elicit divergent answers that the judge deems noise: the answer split reflects superficial wording preferences rather than a consequential difference in assumptions or intentions.

These failure patterns suggest three directions for future improvement: stricter budget constraints or explicit diversity rewards to reduce redundancy; implicit-gap classifiers or chain-of-thought grounding to reduce explicit-gap misclassification; and a meaning-grounded option generator to reduce non-meaningful divergence.