WHAT MAKES A SOCIAL AGENT'S INTERACTIONS EFFECTIVE? LEARNING FROM VERIFIABLE OUTCOMES IN PAIRED WORLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Effective social interaction requires more than fluent responses: agents must reach plans that fit the participants' needs and circumstances. Existing evaluations of social agents mainly rely on LLM judges that rate the conversation, without checking the resulting plan. We introduce HICO-Bench, a benchmark of 180 everyday coordination scenarios, such as resource allocation, scheduling, and authorization, in which two agents with private information converse and then independently record the plan they agreed on. A program checks *Feasibility* (the two records match and satisfy both participants' private conditions), and an LLM judge checks *Integrity* (the conversation supports the recorded plan); passing both is a *Verified Success*. Each scenario has two versions that share the same public context but differ in one private condition, so the correct plan differs between them. Across thirteen LLMs, 36% of conversations rated highly for goal completion fail Verified Success. To learn from outcomes that arrive only at the end of a conversation, we propose Paired-Credit Policy Optimization (PaCPO). After each turn, PaCPO asks the agent for a tentative plan, checks it in both versions of the scenario, and rewards the turns after which the tentative plans fit their respective versions. PaCPO raises Qwen3-8B's Verified Success from 20.4% to 28.3%.

1 INTRODUCTION

LLM agents increasingly help people coordinate everyday matters. Consider Maya asking her colleague Leo, *"Could you give me a hand with tomorrow's client pitch?"* (Figure 1). Whether Leo should review her slides or present in her place depends on facts that only Maya knows: whether she can attend, and what help she wants. A reply such as *"Don't worry, I'll take over"* sounds supportive, yet commits to a plan that fails if Maya can present herself. Splitting a bill, scheduling a meeting, and approving a request share this structure: an acceptable outcome depends on *private conditions*, such as budget, availability, or authority, revealed through dialogue.

Existing evaluations of LLM social agents focus primarily on the conversation. SOTOPIA (Zhou et al., 2024) uses an LLM judge to rate agents on goal completion, believability, and related dimensions, and subsequent work trains agents to improve these ratings (Wang et al., 2024; Yu et al., 2025). Such ratings do not verify whether the agreed plan satisfies the participants' conditions, and model-based judges can overestimate agents' social abilities (Wang et al., 2024). Negotiation and tool-use benchmarks do verify outcomes automatically (Lewis et al., 2017; Yao et al., 2024), but not in everyday coordination, where the decisive facts must be elicited through conversation.

We introduce HICO-Bench (Hidden Information and Checkable Outcomes), a benchmark that evaluates social interaction by its outcome. After each conversation, both agents independently record the agreed plan in a structured form. A program checks *Feasibility*: whether the two records match and satisfy all private conditions. An LLM judge checks *Integrity*: whether the conversation supports the recorded plan, for example without invented authority or a record that contradicts what was said. An interaction is a *Verified Success* only if it passes both checks. Each scenario has two versions, or *worlds*, that differ in a single private condition (e.g., whether Maya can attend) and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **Reassuring words need not lead to appropriate help.** (b) Without clarifying, Leo offers to take over, although Maya can present and only wants a slide review. (c) After clarifying, the appropriate plan depends on whether Maya can attend, which is initially private to her. Feasibility checks that the plan works and both records match; Integrity checks that dialogue supports it.

admit disjoint sets of correct plans; an agent that responds identically in both worlds therefore fails at least one. Across thirteen LLMs, 36% of conversations with Goal $\geq 8/10$ fail Verified Success.

Verified outcomes can also serve as a training signal, but they arrive only after the final turn, leaving the contribution of individual turns unclear. We propose *Paired-Credit Policy Optimization* (PaCPO), which uses the paired worlds for credit assignment. After each turn, we probe the agent, outside the conversation, for a tentative plan it would record at that point, and check this plan in both worlds. Before Leo learns whether Maya can attend, his tentative plan is the same in both worlds and can fit at most one; once she answers, the plan formed in each world should fit that world but not its counterpart. PaCPO rewards the turns at which this occurs, redistributing the terminal reward without changing each conversation's total return. On Qwen3-8B, PaCPO improves Verified Success from 20.4% to 28.3%, compared with 25.0% for PPO and 26.3% for GRPO trained on the same number of dialogues. Our contributions are as follows:

- **HICO-Bench**, a benchmark of 180 paired scenarios across six coordination families, evaluated by rule-based Feasibility and judge-based Integrity alongside SOTOPIA ratings.
- **PaCPO**, a credit-assignment method that compares tentative plans across paired worlds once the private condition that differs between them has been disclosed.
- **An evaluation of thirteen LLMs and five RL baselines**, showing that highly rated conversations can yield failed plans and that outcome training improves Verified Success.

2 Related Work

Social reasoning and interaction. Social IQa and FANToM test social commonsense and reasoning about who knows what in a conversation (Sap et al., 2019; Kim et al., 2023). SOTOPIA evaluates agents in role-played social scenarios, with an LLM judge rating goal completion, believability, and other dimensions (Zhou et al., 2024); SOTOPIA-ToM extends this setting to information seeking and privacy (YS et al., 2026). HICO-Bench focuses on the plan reached through dialogue: whether it fits both participants' private conditions and is supported by what they said. We retain SOTOPIA's ratings as complementary measures of conversation quality.

Task-oriented dialogue and negotiation. Several interactive benchmarks verify outcomes with rules. In Deal or No Deal and CaSiNo, agents divide items and the final split is scored against each agent's private values (Lewis et al., 2017; Chawla et al., 2021); τ-bench and τ^2-bench check whether a customer-service agent leaves a database in the correct state (Yao et al., 2024; Barres et al., 2025);

Social Gym uses games whose winner is decided by rules (He et al., 2026). These settings target fixed, domain-specific goals. HICO-Bench extends outcome verification to everyday coordination and additionally checks that the conversation supports the plan.

Clarification and information seeking. Prior work trains or evaluates models on asking clarifying questions, for example by rewarding questions that reveal user preferences (Andukuri et al., 2024) or by scoring questions at checkpoints (Zhao et al., 2026; Zhang et al., 2024). We do not reward questions directly: in HICO-Bench, the decisive fact may be volunteered, already known, or elicited, and only the correctness of the resulting plan is rewarded.

Credit assignment in reinforcement learning. When feedback arrives only at the end of an episode, a learner must estimate which earlier steps deserve credit. Process supervision labels each reasoning step (Lightman et al., 2024); VinePPO estimates a step's value by sampling continuations from it (Kazemnejad et al., 2025); GiGPO and GraphGPO compare steps that share the same state across rollouts (Feng et al., 2025; Cheng et al., 2026). For social agents, SOTOPIA-RL and SAVOIR distribute LLM-judge ratings over the utterances of a conversation (Yu et al., 2025; Feng et al., 2026). In free-form dialogue, however, identical states rarely recur across rollouts, and LLM ratings are noisy. PaCPO instead compares tentative plans across the two worlds of a scenario and checks them with rules, so that each comparison is verifiable.

3 HICO-BENCH

3.1 INTERACTION PROTOCOL

Each scenario involves two roles, such as Maya and Leo in Figure 1. Both roles see the same *public context* p: the client pitch is tomorrow at 10 a.m., Maya is due to present, and Leo may present only if Maya cannot attend. Each role i also receives a private *role card* v_i that the other role does not see; Maya's card states whether she can attend and what help she wants. The roles then alternate free-form turns, producing a history h. Each role observes only p, its own card, and h; it never sees the partner's card or the set of acceptable plans.

When the conversation ends, each role independently fills in a structured *record* c_i describing the plan it believes was agreed, following a public template $\mathcal{C}$ (e.g., who presents and what Leo will do). Independent records test whether the roles reached a shared understanding (Clark & Brennan, 1991). The *verifier state* z stores the private conditions of both roles. A scenario is thus $x = (p, v_1, v_2, z, \mathcal{C})$.

3.2 PAIRED SCENARIOS

Each scenario has two worlds, w_0 and w_1, that share the public context and differ in exactly one private condition, chosen so that no plan is correct in both. In our example, Maya can attend in w_0, so the correct plan is that she presents and Leo reviews her slide; in w_1 she cannot attend, so the correct plan is that Leo presents. Following contrast sets (Ribeiro et al., 2020; Gardner et al., 2020), an agent that makes the same offer in both worlds must fail one of them.

Table 1 lists the six coordination families. We adapt social situations from SOTOPIA (Zhou et al., 2024), write role cards for both worlds, and encode the acceptance rules as a program, which is tested on correct and incorrect plans. The 180 scenarios are split into 96 training, 24 development, and 60 test scenarios, with both worlds in the same split. Evaluating each role in each world separately yields $60 \times 2 \times 2 = 240$ test tasks (96 development tasks).

3.3 OUTCOME EVALUATION

Feasibility: does the plan work? Each record specifies an outcome *kind* and its *terms*. The kind is *agreement* (a plan is settled), *decline* (no plan satisfies both roles), or *defer* (the request requires approval from someone outside the conversation). In w_0 of our example, a correct record is `{kind: agreement, terms: {presenter: Maya, leo_task: review_budget_slide}}`. A program checks that (1) the two records are identical up to the ordering of fields and items; (2) the kind is correct for the world; and (3) the terms satisfy the family's rules (Table 1), e.g., in resource allocation the shares sum to the $240 bill and respect each person's limit. Any plan satisfying

Table 1: **Interaction families in HICO-Bench.** Each family has 16 training, 4 development, and 10 test scenarios, each with two worlds.

Interaction family	Example	Private conditions and verification
Resource allocation	Neighbors divide a $240 fence-repair bill.	Shares respect both participants' bounds and sum to the fixed total.
Authorization	A consultant handles a request for one extra review hour.	Approval stays within delegated authority; otherwise, the request is referred.
Bounded quantities	A buyer and seller agree on the number of apple crates.	The agreed integer quantity lies within both participants' acceptable intervals.
Joint scheduling	Studio partners arrange a 20-minute celebration.	The shared slot fits both availability windows and has the required duration.
Shared-resource scheduling	Two photographers reserve a tripod for ten minutes each.	Bookings fit both use windows and required durations, without overlapping.
Set-based consent	Two people agree on belief-related discussion topics.	Selected topics match the required set and are all permitted.

the rules is accepted; in scheduling, for instance, any mutually free slot passes. $F_z(c_1, c_2) = 1$ if all three tests pass and 0 otherwise.

Integrity: does the conversation support the plan? A feasible record may still be unsupported by the conversation: both records may state that Maya presents while Leo promised to take over, or Leo may approve extra work by claiming authority he lacks. An LLM judge reads the full conversation and both records and flags four violations, each with quoted evidence: (i) an uncorrected factual error, such as invented authority or resources; (ii) repeated exchanges without progress; (iii) a record that contradicts the spoken agreement; and (iv) output-format text leaking into the dialogue. $I_z(h, c_1, c_2) = 1$ if no violation is found. Corrected proposals and truthful referrals are allowed.

Verified Success. Our primary metric is the fraction of tasks that pass both checks,

$$S_z(h, c_1, c_2) = F_z(c_1, c_2)\, I_z(h, c_1, c_2). \tag{1}$$

We also report each check separately, as they capture different failure modes.

Conversation ratings. We additionally report SOTOPIA's seven LLM-judged ratings (Zhou et al., 2024), focusing on *Goal* (0–10, attainment of the role's social goal) and *Believability* (0–10, naturalness and role consistency); all seven appear in Appendix B. These ratings are never used for training; they test whether high ratings coincide with verified outcomes.

4 Paired-Credit Policy Optimization

PaCPO trains the agent with a terminal reward but redistributes it across turns. After each turn, we ask which plan the agent would record if the conversation ended there, and compare these tentative plans across the two worlds of a scenario. Before Maya reveals whether she can attend, Leo's tentative plan is the same in both worlds and can fit at most one; once she does, the plans formed in the two worlds should diverge so that each fits its own world. PaCPO rewards the turns at which this divergence appears. Figure 2 illustrates the pipeline; Algorithm 1 (Appendix E) details the update.

4.1 Terminal Reward

All RL methods in our experiments optimize the same terminal reward

$$R(h, c_1, c_2, z) = 0.8F_z(c_1, c_2) + 0.2I_z(h, c_1, c_2), \tag{2}$$

which gives partial credit when only one check passes. PPO and GRPO receive R at the final turn; PaCPO redistributes it across turns.

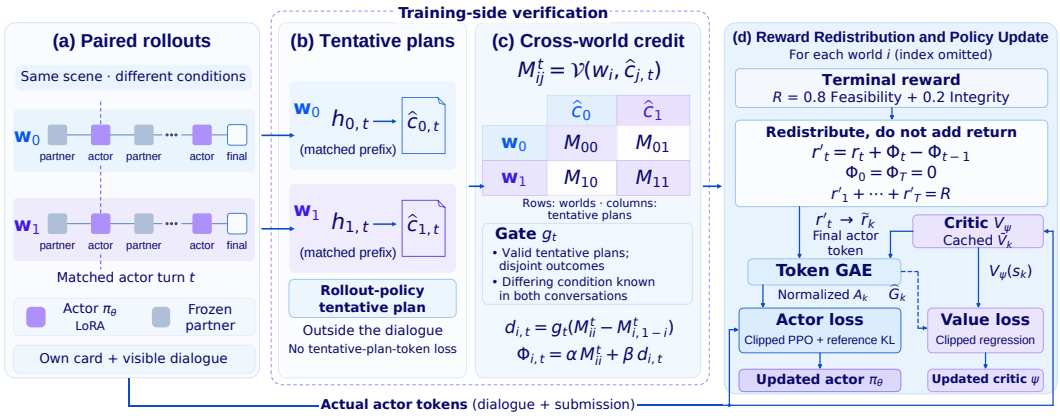

Figure 2: **PaCPO overview.** Rollouts in the two worlds yield a tentative plan after each turn. Checking each plan in both worlds determines turn-level credit, which redistributes the terminal reward before the PPO update. Indices i, j denote worlds, t turns, and k tokens; panel (d) omits i. R is the terminal reward, r_t and r'_t are the original and redistributed rewards, and Φ_t is the potential. The dashed branch carries GAE value targets.

4.2 Tentative Plans and Cross-World Checks

Let $i \in \{0, 1\}$ index the worlds. For each scenario, we roll out the same tested role in both worlds against a fixed partner; the two conversations may differ in wording and length. Let v_i be the tested role's card in world w_i, $h_{i,t}$ the conversation up to its t-th turn, and T_i the turn at which it submits its record. After each turn $t < T_i$, we probe the current policy for a *tentative plan*: the plan it would record if the conversation ended there, given only what the role can see:

$$\widehat{c}_{i,t} \sim \pi_{\theta_{\text{old}}}(\cdot \mid p, v_i, h_{i,t}, \mathcal{C}). \tag{3}$$

The tentative plan is not added to the conversation.

Let $\mathcal{V}(w, c) \in \{0, 1\}$ indicate whether plan c satisfies the Feasibility rules of world w, with $\mathcal{V}(w, \perp) = 0$ for a missing or malformed plan. At turn t, PaCPO checks both tentative plans in both worlds:

$$M^t_{ij} = \mathcal{V}(w_i, \widehat{c}_{j,t}), \qquad i, j \in \{0, 1\}. \tag{4}$$

The diagonal entries indicate whether each world's plan fits its own world; the off-diagonal entries indicate whether the plan from the other world would also fit.

4.3 Paired Credit

The comparison is informative only after the private condition that differs between the worlds has been disclosed. We therefore compare plans at the same turn index and open a gate, $g_t = 1$, only if both tentative plans are valid, the two worlds admit no common plan, and the differing condition is available in both conversations by turn t: it is either on the tested role's own card or has been disclosed by the partner, as marked by a frozen LLM annotator during training. For world i, the paired credit is

$$d_{i,t} = \begin{cases} M^t_{ii} - M^t_{i,1-i}, & g_t = 1, \\ 0, & g_t = 0. \end{cases} \tag{5}$$

It is positive when the plan formed in w_i fits w_i but the plan from the other world does not. In our example, once Maya has answered, proposing that Maya presents in w_0 and that Leo presents in w_1 gives $M^t = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ and hence $d_{0,t} = d_{1,t} = 1$; identical plans in both worlds give zero credit. The potential at turn t combines within-world feasibility and paired credit:

$$\Phi_{i,t} = \alpha M^t_{ii} + \beta d_{i,t}. \tag{6}$$

Table 2: **Thirteen-model HICO-Bench benchmark.** Fixed Qwen3-8B partner, 240 tasks per model; sorted by increasing Verified Success (Ver. Succ.). Size denotes local model size; `api` denotes remote inference. Goal and Believability range from 0 to 10; full ratings are in Appendix D.

Model	Size	Verifiable outcomes (%)			SOTOPIA ratings	
		Feasibility	Integrity	Ver. Succ.	Goal	Believ.
Qwen3-8B (Yang et al., 2025)	8B	49.58	25.83	20.42	7.47	6.31
MiniMax M3 (MiniMax, 2026)	api	58.33	62.50	50.42	9.18	8.68
GPT-4o (OpenAI, 2024)	api	62.08	63.33	51.25	9.12	8.71
GPT-5.6 Luna (OpenAI, 2026a)	api	73.75	56.67	52.08	9.61	8.63
Qwen3.8-27B (Qwen Team, 2026)	27B	64.17	62.92	54.58	9.21	8.72
DeepSeek-V4-Pro (DeepSeek-AI, 2026)	api	68.33	66.25	59.17	9.44	8.88
Qwen3.8-Max (Alibaba Cloud, 2026)	api	72.50	69.58	60.83	9.62	8.95
Gemma 4 31B IT (Google DeepMind, 2026b)	31B	72.08	72.92	63.75	9.38	8.80
Kimi-K3 (Moonshot AI, 2026)	api	75.00	77.92	66.25	9.52	8.94
GLM-5.3 (Z.ai, 2026)	api	73.75	75.83	67.08	9.65	8.96
GPT-5.6 Sol (OpenAI, 2026a)	api	81.25	82.92	76.25	9.63	8.95
GPT-6 Astra (OpenAI, 2026b)	api	**87.50**	86.25	80.83	**9.75**	**9.01**
Gemini 3.7 Flash (Google DeepMind, 2026a)	api	87.08	**86.67**	**82.50**	9.54	8.93

4.4 Reward Redistribution

Let $r_{i,T_i} = R_i$ and $r_{i,t} = 0$ for $t < T_i$, and set $\Phi_{i,0} = \Phi_{i,T_i} = 0$. Following potential-based shaping and return redistribution (Ng et al., 1999; Arjona-Medina et al., 2019), we define

$$r'_{i,t} = r_{i,t} + \Phi_{i,t} - \Phi_{i,t-1}. \tag{7}$$

With $\gamma = 1$, the shaping terms telescope, so the total reward of each conversation is unchanged:

$$\sum_{t=1}^{T_i} r'_{i,t} = R_i + \Phi_{i,T_i} - \Phi_{i,0} = R_i. \tag{8}$$

A turn is thus rewarded when it raises the potential, for example by eliciting the condition that lets the two worlds' plans diverge correctly, and the final turn receives R_i minus the credit already assigned. We optimize the redistributed rewards with PPO (Schulman et al., 2017), placing each $r'_{i,t}$ on the last token of turn t and using a learned value baseline with GAE (Schulman et al., 2015); Appendix E gives the details.

5 Experimental Results

After describing the setup (Section 5.1), we ask what outcome verification reveals beyond conversation ratings (Section 5.2), whether learning from verified outcomes improves interaction (Section 5.3), whether the gains extend to other partners and tasks (Section 5.4), and which components of PaCPO contribute (Section 5.5).

5.1 Experimental Setup

Training. We train Qwen3-8B (Yang et al., 2025) with LoRA (Hu et al., 2022) against a frozen Qwen3-8B partner. Each RL run uses 384 training conversations: one preassigned role per training scenario, balanced across scenarios, with two samples per world. During training, Integrity is scored by a frozen local Qwen3.8-27B judge (Qwen Team, 2026). To keep the number of sampled conversations equal across methods, the group-based baselines (GRPO, Dr.GRPO, GiGPO, and GraphGPO), which draw $G = 4$ conversations per world, cover 48 training scenarios, whereas PPO and PaCPO, which draw two per world, cover all 96. PaCPO uses $\alpha = 0.25$ and $\beta = 0.10$. Appendix C.1 reports an additional SFT-initialization study, and Appendix E.2 reports training costs.

Evaluation. The test set has 240 tasks (60 scenarios, two worlds, two roles), each played against the frozen Qwen3-8B partner. Feasibility is computed by rules; Integrity and the SOTOPIA ratings

Table 3: **Seven-system training comparison.** 240 test tasks per row; 384 training dialogues per RL method. To match this budget, group-based baselines ($G = 4$) cover 48 training scenarios, and PPO and PaCPO cover 96. Goal and Believability (0–10) are not optimized. Best values are bold.

Method	Verifiable outcomes (%)			SOTOPIA ratings	
	Feasibility	Integrity	Ver. Succ.	Goal	Believ.
Base (Yang et al., 2025)	49.58	25.83	20.42	7.47	6.31
PPO (Schulman et al., 2017)	50.00	33.75	25.00	7.58	6.51
GRPO (Shao et al., 2024)	**52.92**	31.25	26.25	7.83	6.62
Dr.GRPO (Liu et al., 2025)	50.42	27.92	21.25	7.67	6.39
GiGPO (Feng et al., 2025)	52.08	30.42	24.58	**7.86**	6.60
GraphGPO (Cheng et al., 2026)	52.50	31.67	24.17	7.84	6.60
PaCPO	**52.92**	**35.42**	**28.33**	7.83	**6.70**

are judged by Qwen3.8-Max (Alibaba Cloud, 2026), a different model from the training judge, following rubric-based LLM evaluation (Zheng et al., 2023; Liu et al., 2023). Failed generations count as failures. On 240 re-judged outputs, Qwen3.8-Max and Kimi-K3 agree on 90.0% of Integrity decisions (Cohen's $\kappa = 0.78$; Appendix K.1).

5.2 Multi-Model Benchmark

Table 2 reports all thirteen models. Verified Success ranges from 20.42% to 82.50%, leaving headroom even for the strongest models. The two checks capture different failures: Qwen3-8B passes Feasibility on 49.58% of tasks but reaches only 20.42% Verified Success, mostly due to Integrity failures.

Conversation ratings and verified outcomes can rank models differently. GPT-5.6 Luna scores above Gemini 3.7 Flash on Goal (9.61 vs. 9.54) but far below it on Verified Success (52.08% vs. 82.50%). Pooled over all models, 35.99% of conversations with Goal ≥ 8 fail Verified Success (Appendix G), because the recorded plan is infeasible or the conversation does not support it.

Dialogue itself matters. On 12 development scenarios with hidden conditions, dialogue raises Base Feasibility by 50 points over submitting records directly (Appendix G).

5.3 Learning from Verified Outcomes

Table 3 compares RL methods trained with the same terminal reward and number of dialogues. PaCPO achieves the highest Verified Success, improving Base by 7.92 points to 28.33%. PaCPO and GRPO reach the same Feasibility, but PaCPO has higher Integrity, so its gain comes from conversations that better support feasible plans.

Although SOTOPIA ratings are not optimized, PaCPO also improves Goal and Believability over Base (Appendix C). In a blinded pilot with two expert annotators, human judgments agree with Feasibility on 100% of cases and with Qwen3.8-Max Integrity on 82.22%; the annotators also prefer PaCPO over Base for helpfulness more often than the reverse (8 vs. 5 of 24 pairs; Appendix K.2).

Where outcome learning helps. Gains vary by family (Figure 3). The largest improvements over Base are in bounded quantities (+22.50 points) and authorization (+10.00 points). In bounded quantities, PPO and PaCPO reach the same Feasibility, so PaCPO's additional gain again comes from Integrity. Resource allocation stays at 17.5% Verified Success for Base, PPO, and PaCPO, and both scheduling families remain difficult (Appendix J).

What changes in the conversation? Figure 5a breaks down the aggregate gains into changes across individual scenarios. Improvements and regressions occur within the same family, including the difficult scheduling settings. Figure 5b shows one such improvement. Finley and Harper disclose nonoverlapping windows for using a shared table. Base incorrectly calls them overlapping, and its partner accepts that error. PaCPO instead identifies the four-minute gap and obtains agreement.

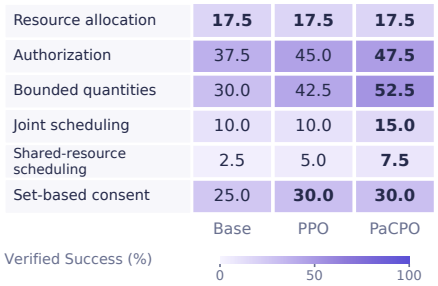

Figure 3: **Verified Success by interaction family.** Forty test tasks per family with a fixed partner; bold marks the best score in each row.

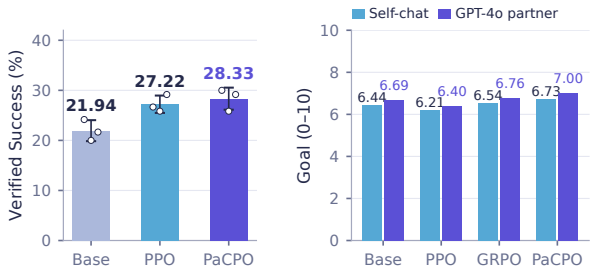

Figure 4: **Self-play and SOTOPIA transfer.** Left: HICO Verified Success, mean ± SD over three generation repeats. Right: SOTOPIA Goal on 70 environment–character pairs per partner setting, judged by GPT-4o.

Table 4: **External interaction tasks.** Task-specific training evaluated on 154 MultiWOZ and 86 CaSiNo test dialogues. All metrics are percentages; Norm. Nash denotes normalized Nash surplus.

	MultiWOZ	CaSiNo	
Method	Comm.	Agreement	Norm. Nash
Base (Yang et al., 2025)	79.87	36.05	27.40
PPO (Schulman et al., 2017)	80.52	60.47	45.86
GRPO (Shao et al., 2024)	79.87	53.49	42.14
PaCPO (ours)	**81.17**	**63.95**	**49.37**

Both runs ultimately submit the same feasible records, but only PaCPO's conversation supports them. This example illustrates why checking the final records alone misses an important part of successful coordination: the participants must also establish the plan in dialogue.

5.4 Self-Play and Interactive Transfer

We also evaluate self-play, transfer to SOTOPIA, and training on other interaction tasks.

Self-play. Self-play replaces the frozen partner with another copy of the evaluated checkpoint, testing robustness when both participants change. Across three generation repeats, PaCPO reaches 28.33% mean Verified Success, exceeding Base by 6.39 points and PPO by 1.11 points (Figure 4, left). Gains persist in this setting, with a smaller margin over PPO than against the fixed partner. Full results appear in Appendix H.

SOTOPIA transfer. SOTOPIA (Zhou et al., 2024) tests direct transfer. Using iStar's partner settings and GPT-4o scoring (Liu et al., 2026), we evaluate HICO checkpoints on 70 environment–character pairs without further training. PaCPO leads in both partner settings (Figure 4, right), connecting outcome training to social performance beyond HICO's checks (Appendix H).

Training on other interaction tasks. MultiWOZ (Budzianowski et al., 2018) constraint communication and CaSiNo (Chawla et al., 2021) negotiation test task-specific training (Table 4; Appendix I). PaCPO gives a small communication gain on MultiWOZ and leads agreement and normalized Nash surplus on CaSiNo. Gains are larger in bargaining than in constraint communication.

5.5 Credit Ablations

Table 5 isolates three design choices. Replacing the other world's plan with a second plan from the same world tests whether the paired worlds matter, and shuffling credit across turns tests whether its timing matters. Both reduce Verified Success to 26.25%, 2.08 points below full PaCPO (Ap-

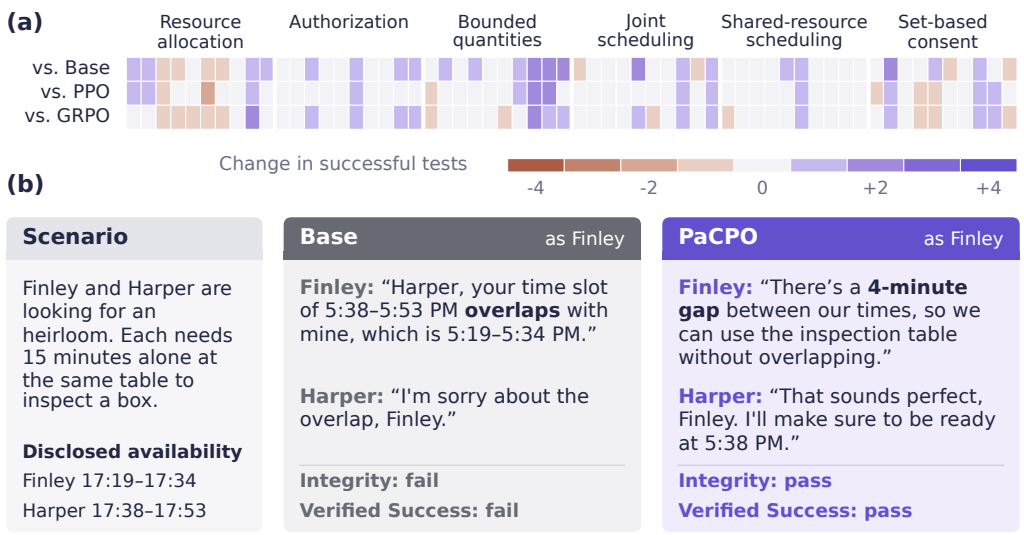

Figure 5: **Scenario-level changes and a dialogue example.** (a) PaCPO's change in successful tests relative to each baseline across 60 scenarios (four tests per column: two worlds and two roles). (b) Verbatim excerpts with the same availability information and Finley as the tested role. Both runs submit the same feasible records, but only PaCPO's dialogue supports them.

Table 5: **PaCPO credit ablations.** 240 test tasks per condition; all values are percentages.

Training condition	Feasibility	Integrity	Ver. Succ.
Same-world candidate control	50.83	32.08	26.25
Shuffled credit positions	50.42	32.92	26.25
Without disclosure gate	49.58	33.33	25.83
PaCPO (full method)	**52.92**	**35.42**	**28.33**

pendix E). Opening the gate before the differing condition is disclosed reduces Verified Success to 25.83%, the largest drop, consistent with assigning credit only after disclosure.

During training, paired credit is nonzero at 55.6% of gate-open turns (Appendix J); otherwise, the two worlds' plans are equally feasible, and only the within-world and terminal terms apply.

6 Conclusion

Effective social interaction should be judged not only by the conversation but also by the plan it produces. HICO-Bench evaluates this with paired scenarios and two checks, Feasibility and Integrity, whose conjunction defines Verified Success. Across thirteen LLMs, high conversation ratings often accompany plans that fail these checks. PaCPO uses the paired worlds to turn end-of-conversation outcomes into turn-level credit, improving Verified Success over the base model and RL baselines. These results suggest that verified outcomes complement conversation ratings, both for evaluation and as a training signal. The benchmark currently covers six coordination families with structured outcomes, and Integrity relies on an LLM judge; extending verification to less structured interactions and expanding human validation are important next steps (Appendices L and K).

AI use statement

Generative AI assisted manuscript preparation and code development. Model-based data generation and evaluation are described in the relevant sections. The authors take full responsibility for the final work.

Ethics statement

The scenarios use fictional or adapted role-play material. Their operational conditions concern feasible coordination and do not establish judgments about real individuals. Broader social quality remains distinct from feasibility under these formal conditions. Release preparation includes provenance and applicable licenses; released artifacts should exclude identifying information.

Reproducibility statement

Appendices A and B describe the scenario families and evaluation criteria; Appendix E gives the training configuration and algorithm, and Appendices H and I document the external evaluation protocols. Tables and figures derive from recorded results. The intended release includes scenario lineage, role cards, verifier tests, prompts, rubrics, training settings, and checkpoint IDs.

References

Alibaba Cloud. Alibaba unveils Qwen3.8-Max: Its largest and most capable flagship model to date. Official release announcement, 2026. URL https://www.alibabacloud.com/en/press-room/alibaba-unveils-qwen3-8-max. Accessed September 24, 2026.

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

Jose A. Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. RUDDER: Return decomposition for delayed rewards. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1806.07857.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*, 2025. URL https://arxiv.org/abs/2506.07982.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ—a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of EMNLP*, pp. 5016–5026, 2018. doi: 10.18653/v1/D18-1547. URL https://aclanthology.org/D18-1547/.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of NAACL-HLT*, pp. 3167–3185, 2021. doi: 10.18653/v1/2021.naacl-main.254. URL https://aclanthology.org/2021.naacl-main.254/.

Xin Cheng, Shuo He, Lang Feng, HaiYang Xu, Ming Yan, Lei Feng, and Bo An. Beyond trajectory-level attribution: Graph-based credit assignment for agentic reinforcement learning. *arXiv preprint arXiv:2605.26684*, 2026. URL https://arxiv.org/abs/2605.26684.

Herbert H. Clark and Susan E. Brennan. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley (eds.), *Perspectives on Socially Shared Cognition*, pp. 127–149. American Psychological Association, 1991. doi: 10.1037/10096-006. URL https://doi.org/10.1037/10096-006.

DeepSeek-AI. DeepSeek-V4-Pro model card. Official model repository, 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro. Accessed September 24, 2026.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 422–428, 2020. URL https://aclanthology.org/2020.lrec-1.53/.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *Advances in Neural Information Processing Systems*, volume 38, pp. 46375–46408, 2025. doi: 10.52202/085713-1544. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/420c9f777c0b4f78d515e53cf74d58b2-Abstract-Conference.html.

Xiachong Feng, Yi Jiang, Xiaocheng Feng, Deyi Yin, Libo Qin, Yangfan Ye, Lei Huang, Weitao Ma, Yuxuan Gu, Chonghan Qin, Bing Qin, and Lingpeng Kong. SAVOIR: Learning social savoir-faire via shapley-based reward attribution. *arXiv preprint arXiv:2604.18982*, 2026. URL https://arxiv.org/abs/2604.18982.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

Google DeepMind. Gemini 3.7 Flash model card, 2026a. URL https://deepmind.google/models/model-cards/gemini-3-7-flash/. Accessed September 24, 2026.

Google DeepMind. Gemma 4 31B IT model card. Official model repository, 2026b. URL https://huggingface.co/google/gemma-4-31B-it. Accessed September 24, 2026.

Keyu He, Xuhui Zhou, and Maarten Sap. Social Gym and SPaRTan: Benchmarking and improving LLM social reasoning via multi-agent game tournaments. *arXiv preprint arXiv:2608.09128*, 2026. URL https://arxiv.org/abs/2608.09128.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Shengyi Huang, Rousslan Fernand Julien Dossa, Antonin Raffin, Anssi Kanervisto, and Weixun Wang. The 37 implementation details of proximal policy optimization. ICLR Blog Track, 2022. URL https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 29557–29590, 2025. URL https://proceedings.mlr.press/v267/kazemnejad25a.html.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 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*, pp. 14397–14413, 2023. doi: 10.18653/v1/2023.emnlp-main.890. URL https://aclanthology.org/2023.emnlp-main.890/.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of EMNLP*, pp. 2443–2453, 2017. doi: 10.18653/v1/D17-1259. URL https://aclanthology.org/D17-1259/.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Xiaoqian Liu, Ke Wang, Yuchuan Wu, Fei Huang, Yongbin Li, Jianbin Jiao, and Junge Zhang. Agentic reinforcement learning with implicit step rewards. In *International Conference on Learning Representations*, volume 2026, pp. 129271–129291, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d1e6739f319be4ba8e748cc05f23bba1-Abstract-Conference.html.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-Like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025. URL https://arxiv.org/abs/2503.20783.

MiniMax. MiniMax-M3 model card. Official model repository, 2026. URL https://huggingface.co/MiniMaxAI/MiniMax-M3. Accessed September 24, 2026.

Moonshot AI. Kimi-K3 model card. Official model repository, 2026. URL https://huggingface.co/moonshotai/Kimi-K3. Accessed September 24, 2026.

John F. Nash, Jr. The bargaining problem. *Econometrica*, 18(2):155–162, 1950. doi: 10.2307/1907266. URL https://www.jstor.org/stable/1907266.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the International Conference on Machine Learning*, pp. 278–287, 1999. URL https://ai.stanford.edu/~ang/papers/shaping-icml99.pdf.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276.

OpenAI. GPT-5.6 system card, 2026a. URL https://deploymentsafety.openai.com/gpt-5-6. Accessed September 24, 2026.

OpenAI. GPT-6 Astra system card, 2026b. URL https://deploymentsafety.openai.com/gpt-6-astra. Accessed September 24, 2026.

Qwen Team. Qwen3.8-27B model card. Official model repository, 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B. Accessed September 24, 2026.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of ACL*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of EMNLP-IJCNLP*, pp. 4463–4473, 2019. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454/.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015. URL https://arxiv.org/abs/1506.02438.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. SOTOPIA-π: Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12912–12940, 2024. doi: 10.18653/v1/2024.acl-long.698. URL https://aclanthology.org/2024.acl-long.698/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024. URL https://arxiv.org/abs/2406.12045.

Yashwanth YS, Ruichen Wang, Shihua Zeng, Xuhui Zhou, Koichi Onoue, Vasudha Varadarajan, and Maarten Sap. SOTOPIA-TOM: Evaluating privacy and information management in multi-agent interaction with theory of mind. *arXiv preprint arXiv:2605.02307*, 2026. URL https://arxiv.org/abs/2605.02307.

Haofei Yu, Zhengyang Qi, Yining Zhao, Kolby Nottingham, Keyang Xuan, Bodhisattwa Prasad Majumder, Hao Zhu, Paul Pu Liang, and Jiaxuan You. Sotopia-RL: Reward design for social intelligence. *arXiv preprint arXiv:2508.03905*, 2025. URL https://arxiv.org/abs/2508.03905.

Z.ai. GLM-5.3 model card. Official model repository, 2026. URL https://huggingface.co/zai-org/GLM-5.3. Accessed September 24, 2026.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. MultiWOZ 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 109–117, 2020. doi: 10.18653/v1/2020.nlp4convai-1.13. URL https://aclanthology.org/2020.nlp4convai-1.13/.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10746–10766, 2024. doi: 10.18653/v1/2024.acl-long.578. URL https://aclanthology.org/2024.acl-long.578/.

Jiale Zhao, Ke Fang, and Lu Cheng. When and what to ask: AskBench and rubric-guided RLVR for LLM clarification. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 17120–17140, 2026. doi: 10.18653/v1/2026.findings-acl.845. URL https://aclanthology.org/2026.findings-acl.845/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.05685.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, volume 2024, pp. 40975–41019, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b3075b88e583a0e98d8b24338a613060-Abstract-Conference.html.

APPENDIX CONTENTS

A Scenario Families . . . 15

B Evaluation Metrics . . . 15

C Additional Training Results . . . 16

D Model Benchmark . . . 17

E Training Details . . . 17

F A Paired Scenario . . . 22

G Outcome Reliability and Information Diagnostics . . . 23

H Interactive Transfer and Self-Play . . . 23

I External-Task Training . . . 24

J Family-Level Analysis . . . 25

K Robustness and Human Validation . . . 26

L Limitations and Future Work . . . 27

A Scenario Families

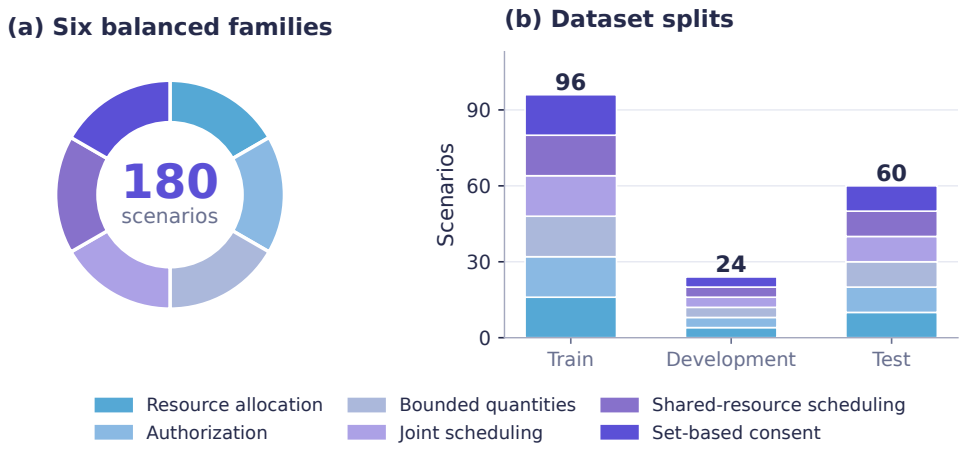

Figure 6: **Dataset composition.** Each family contributes 30 scenarios: 16 training, 4 development, and 10 test. Two worlds per scenario yield 360 worlds; role-wise evaluation yields 96 development and 240 test tasks.

HICO-Bench groups scenarios by the coordination requirement that determines which arrangements are feasible. These requirements apply across settings such as work, family, and commerce.

Table 1 gives examples and verification rules; Figure 6 shows the balanced distribution. Each family has 30 scenarios: 16 training, 4 development, and 10 test, with two worlds per scenario. Shared-resource scheduling verifies two non-overlapping reservations. Set-based consent checks explicit required and allowed item sets.

Feasibility requires matching independent records and the correct outcome kind: agreement when a feasible plan exists, decline when none exists, or defer when the request requires authorization that neither role has.

B Evaluation Metrics

In addition to the primary outcomes and SOTOPIA ratings, we report three exploratory diagnostics: *Information-gap Management* (IM), *Grounded Understanding Accuracy* (GU), and *Information-informed Action Appropriateness* (IA). These use the frozen task-specific rubric and are reported separately, outside the primary metrics and training reward.

IM does not require a question when the necessary information is already available. GU distinguishes stated facts from assumptions and tracks correction. IA includes proposals, conditional commitments, deferral, and refusal rather than physical action. IA diagnoses information-conditioned response quality.

Each rating includes a short explanation linked to dialogue turn IDs. The anchors assess how the tested role manages information, interprets disclosures, and develops a response. Scores of 9–10 require concrete dialogue evidence that the relevant information has been handled sufficiently and accurately.

Table 6: **Evaluation metrics.** Binary outcomes and social dimensions with their scoring mechanisms. The original seven dimensions retain SOTOPIA ranges (Zhou et al., 2024).

Metric	Range	Mechanism	Interpretation
Feasibility	$\{0, 1\}$	Rules	Compatible records with terms satisfying the modeled conditions.
Integrity	$\{0, 1\}$	Language judge	No unresolved factual error, loop, commitment mismatch, or protocol leakage.
Verified Success	$\{0, 1\}$	Product	Feasibility and Integrity both pass.
Goal	$[0, 10]$	Language judge	Attainment of the tested role's goal.
Knowledge	$[0, 10]$	Language judge	Acquisition of relevant information.
Believability	$[0, 10]$	Language judge	Naturalness and role consistency.
Relationship	$[-5, 5]$	Language judge	Effect on the relationship.
Secret	$[-10, 0]$	Language judge	Preservation of protected information.
Social Rules	$[-10, 0]$	Language judge	Social or legal violations.
Material Benefits	$[-5, 5]$	Language judge	Financial and material effects.
IM	$[0, 10]$	Language judge	Information gaps resolved or appropriately kept conditional.
GU	$[0, 10]$	Language judge	Accurate interpretation supported by available evidence.
IA	$[0, 10]$	Language judge	Sufficiently specified responses using available information.

Table 7: **Anchors for the three diagnostic ratings.** Each dimension evaluates the tested role over the complete dialogue using the frozen rubric.

Score	IM: information gaps	GU: interpretation	IA: response appropriateness
0–2	Commits despite a critical manageable unknown, or inquiry seriously blocks progress.	Material fabrication or persistent misunderstanding of facts, roles, or consent.	An unretracted response violates a critical known condition or invents authority.
3–5	Useful exchange occurs, but a consequential gap remains unmanaged.	Only generic acknowledgment, or an unresolved material interpretive error.	A useful direction, but an essential term is missing or unsupported despite an opportunity.
6–8	Main uncertainty handled, with a specific omission, redundancy, or delay.	Main facts understood, with a limited ambiguity or relevant corrected error.	Mainly feasible and concrete, with a limited coordination or scope weakness.
9–10	Demonstrated sufficient information, or a genuinely conditional step managing the dependency.	Concrete responses demonstrate sufficient, accurate understanding and preserve uncertainty.	A sufficiently specified, grounded response; justified refusal or referral can qualify.

C Additional Training Results

C.1 SFT Initialization

Table 8 compares Base, SFT, PaCPO, and SFT followed by PaCPO (SFT + PaCPO) to examine how demonstration-based initialization complements outcome learning. This four-condition study supplements the Base-initialized algorithm comparison in Table 3.

Applying PaCPO after SFT raises Verified Success from 25.83% (SFT alone) to 31.67%, the highest in our experiments (Table 8). The gain comes mainly from Integrity: the Feasibility of SFT-initialized PaCPO (48.75%) remains slightly below that of Base, and its Goal is lower than that of Base-initialized PaCPO.

Table 8: **SFT initialization and subsequent PaCPO.** 240 test tasks per condition. SFT uses 209 accepted teacher dialogues (991 examples, 64 updates); each RL stage uses 384 dialogues. All generated failures remain in the outcome denominator.

Method	Feasibility	Integrity	Ver. Succ.
Base (Yang et al., 2025)	49.58	25.83	20.42
SFT	46.25	29.58	25.83
PaCPO	**52.92**	35.42	28.33
SFT + PaCPO	48.75	**36.25**	**31.67**

C.2 Complete Social-Quality Scores

Table 9 brings together all training conditions. The first seven score columns are SOTOPIA quality checks; the final three are exploratory process diagnostics. Dimensions retain their original scales; zero on Social Rules denotes no penalty. PaCPO improves Goal and Believability over Base, while SFT-initialized PaCPO obtains the highest Believability. The differing profiles complement the outcome comparison in Table 3.

Table 9: **Complete SOTOPIA ratings for training methods.** Seven SOTOPIA dimensions and three process diagnostics, reported individually on their original scales. Know., Believ., Rel., and Mat. denote Knowledge, Believability, Relationship, and Material Benefits; Rules denotes Social Rules. All columns are higher-is-better; bold marks column-best values. SFT uses 239 scored dialogues; the other conditions use 240.

Method	SOTOPIA ratings							Diagnostics		
	Goal	Know.	Believ.	Rel.	Secret	Rules	Mat.	IM	GU	IA
Range	$[0, 10]$	$[0, 10]$	$[0, 10]$	$[-5, 5]$	$[-10, 0]$	$[-10, 0]$	$[-5, 5]$	$[0, 10]$	$[0, 10]$	$[0, 10]$
Base	7.47	7.41	6.31	2.15	-1.43	**0.00**	1.08	7.87	7.56	7.09
PPO	7.58	7.61	6.51	**2.28**	**-1.24**	**0.00**	1.20	8.12	7.93	7.51
GRPO, $G = 4$	7.83	7.65	6.62	2.24	-1.38	**0.00**	1.28	8.13	7.89	7.50
Dr.GRPO, $G = 4$	7.67	7.67	6.39	2.25	-1.35	**0.00**	1.28	8.16	7.97	7.49
GiGPO, $G = 4$	**7.86**	**7.72**	6.60	2.25	-1.37	-0.03	1.23	8.15	7.89	7.54
GraphGPO, $G = 4$	7.84	7.70	6.60	2.18	-1.45	-0.02	**1.31**	**8.25**	**8.01**	**7.61**
PaCPO	7.83	7.69	6.70	2.19	-1.42	-0.02	1.21	8.18	7.95	7.57
Demonstration-based initialization										
SFT	7.35	7.57	6.62	2.12	-1.48	-0.09	1.14	7.66	7.10	6.89
SFT + PaCPO	7.29	7.42	**7.03**	1.97	-1.41	-0.08	1.18	7.72	7.47	7.17

D Model Benchmark

Figure 7 compares Feasibility, Integrity, and Verified Success for every model, and Table 10 reports their complete SOTOPIA ratings. Goal scores are high for most API models, while the outcome rates remain more widely separated. This contrast motivates evaluating conversation ratings and verified outcomes together.

E Training Details

Algorithm 1 specifies the complete update corresponding to Figure 2.

Training procedure. Both worlds of a scenario are sampled under the same behavior-policy version. The two trajectories are not forced to have identical language or identical length. After each turn of the tested role, PaCPO generates one tentative plan from the visible prefix. Tentative plans at matching turn indices are checked in both worlds. When one conversation lacks a corresponding turn or a valid tentative plan, the paired credit is zero, while within-world feasibility remains available.

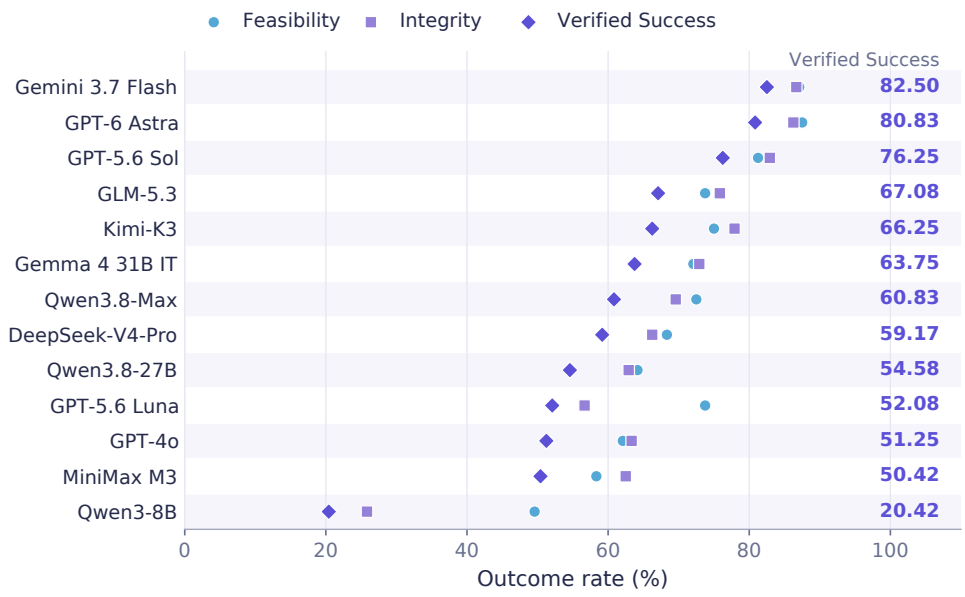

Figure 7: **Verified outcomes across thirteen models.** Each point summarizes 240 tasks; right-hand labels give Verified Success. Complete SOTOPIA ratings appear in Table 10.

Table 10: **Complete SOTOPIA ratings for thirteen models.** Model order follows Table 2; abbreviations and score ranges follow Table 9. Bold marks column-best values. Luna and Sol each have 239 scored dialogues; the other models have 240. All 240 attempts enter the binary-outcome denominators.

Model	SOTOPIA ratings							Diagnostics		
	Goal	Know.	Believ.	Rel.	Secret	Rules	Mat.	IM	GU	IA
Range	$[0, 10]$	$[0, 10]$	$[0, 10]$	$[-5, 5]$	$[-10, 0]$	$[-10, 0]$	$[-5, 5]$	$[0, 10]$	$[0, 10]$	$[0, 10]$
Qwen3-8B	7.47	7.41	6.31	2.15	-1.43	**0.00**	1.08	7.87	7.56	7.09
MiniMax M3	9.18	8.31	8.68	2.80	-0.80	-0.01	1.40	9.31	9.28	9.19
GPT-4o	9.12	8.22	8.71	2.68	-0.76	**0.00**	1.32	9.06	8.99	8.85
GPT-5.6 Luna	9.61	8.47	8.63	2.38	-0.67	**0.00**	1.31	9.55	9.62	9.52
Qwen3.8-27B	9.21	8.30	8.72	2.60	-0.84	**0.00**	1.28	9.12	9.07	9.03
DeepSeek-V4-Pro	9.44	8.46	8.88	2.85	-0.70	**0.00**	1.39	9.50	9.47	9.43
Qwen3.8-Max	9.62	8.53	8.95	**3.10**	-0.74	**0.00**	1.33	9.59	9.72	9.61
Gemma 4 31B IT	9.38	8.21	8.80	2.65	**-0.49**	**0.00**	1.43	9.13	9.22	9.10
Kimi-K3	9.52	8.51	8.94	2.61	-0.75	**0.00**	1.34	9.52	9.56	9.49
GLM-5.3	9.65	8.44	8.96	2.84	-0.50	**0.00**	**1.45**	9.59	9.66	9.61
GPT-5.6 Sol	9.63	**8.60**	8.95	1.99	-0.65	**0.00**	1.33	9.74	9.81	9.79
GPT-6 Astra	**9.75**	**8.60**	**9.01**	2.17	-0.71	**0.00**	1.30	**9.76**	**9.84**	**9.84**
Gemini 3.7 Flash	9.54	8.54	8.93	2.22	-0.66	**0.00**	1.26	9.67	9.74	9.75

Algorithm 1 gives the complete rollout update. Implementation details below specify value learning, token masking, and information boundaries.

Value learning. The policy has a separately optimized linear value head. Its input hidden states are detached from the policy backbone. Proposal verification supplies process rewards, while the value head predicts return for the PPO baseline. Each turn's redistributed reward is placed on the last actor token, and advantages are normalized over the rollout batch. For a three-output illustration, let $R_i = 1$ and $(\Phi_{i,0}, \Phi_{i,1}, \Phi_{i,2}, \Phi_{i,3}) = (0, 0.25, 0.35, 0)$. The original reward sequence $(0, 0, 1)$ becomes $(0.25, 0.10, 0.65)$, retaining a total of 1. This example illustrates the transformation; actual potentials depend on the tentative-plan checks and the gate.

Algorithm 1 Paired-Credit Policy Optimization (PaCPO)

Require: Paired scenarios $\mathcal{D}$; actor π_θ; fixed partner b and reference π_{ref}; value head V_ψ; checks $F, I, \mathcal{V}$; scales $\alpha = 0.25$, $\beta = 0.10$.

for each rollout batch **do**
  Freeze rollout policy $\pi_{\text{old}} \leftarrow \pi_\theta$ and value baseline $\bar{V}$; sample pairs from $\mathcal{D}$.
  Generate independent paired trajectories with π_{old} and b, including final records.
  for each trajectory pair (τ_0, τ_1) with the same tested role **do**
    Compute $R_i \leftarrow 0.8F_i + 0.2I_i$ for each world $i \in \{0, 1\}$.
    for each world i and actor dialogue prefix $t < T_i$ **do**
      Sample $\widehat{c}_{i,t} \sim \pi_{\text{old}}(\cdot \mid p, v_i^{\text{act}}, h_{i,t}, \mathcal{C})$, outside the dialogue.
      Set $\Phi_{i,t} \leftarrow \alpha\mathcal{V}(w_i, \widehat{c}_{i,t})$; use zero for an invalid tentative plan.
    end for
    for each matched actor-turn index t with two valid tentative plans **do**
      Compute $M_{ij}^t \leftarrow \mathcal{V}(w_i, \widehat{c}_{j,t})$ for $i, j \in \{0, 1\}$.
      $g_t \leftarrow \text{Disjoint}(w_0, w_1) \wedge \text{Known}(0, t) \wedge \text{Known}(1, t)$.
      $\Phi_{i,t} \leftarrow \Phi_{i,t} + \beta g_t (M_{ii}^t - M_{i,1-i}^t)$ for both worlds.
    end for
    for each world i **do**
      Set $\Phi_{i,0} = \Phi_{i,T_i} = 0$; $r_{i,T_i} = R_i$ and $r_{i,t<T_i} = 0$.
      Redistribute $r'_{i,t} \leftarrow r_{i,t} + \Phi_{i,t} - \Phi_{i,t-1}$; hence $\sum_t r'_{i,t} = R_i$.
      Place $r'_{i,t}$ on each output's last token; compute raw GAE $\widehat{A}$ using $\bar{V}$.
      Cache value targets $\widehat{G} \leftarrow \widehat{A} + \bar{V}$ before advantage normalization.
    end for
  end for
  Normalize $\widehat{A}$ over rollout actor tokens to obtain A.
  Minimize the clipped actor loss with reference KL and the separate clipped value loss.
end for
return π_θ

T_i is the turn of the final record. PPO includes the tested role's dialogue and record tokens, not partner or tentative-plan tokens. $\text{Known}(i, t)$ uses the tested role's card or an annotated disclosure at or before t; all cross-world checks remain training-side.

Table 11: **Principal PaCPO configuration.**

Parameter	Value
Policy / fixed partner	Qwen3-8B LoRA / frozen Qwen3-8B
Initialization	Base; no SFT adapter
Train scenarios / worlds / conversations	96 / 192 / 384
Rollout batch / PPO epochs	16 conversations / 2
Actor / value-head learning rate	2×10^{-6} / 2×10^{-4}
GAE discount / trace	$\gamma = 1$ / $\lambda = 0.95$
PPO ratio clip / value clip	0.2 / 0.2
Reference KL / value-loss coefficient	0.02 / 0.5
Training Integrity judge / reasoning	Qwen3.8-27B / low
Within-world / paired scale	$\alpha = 0.25$ / $\beta = 0.10$
Maximum output / alternating turns	512 tokens / 20
Training / Test generation seed	917 / 317917
Eligible / nonzero paired positions	1,378 / 766
Accepted optimizer updates	192

Token-level advantages and targets. Within one trajectory, let $k = 1, \ldots, K$ index the tested role's tokens, including the final record. The state s_k contains the role-visible prefix and preceding tokens of the current output. Across outputs, the next actor state includes any intervening partner observation. Partner and tentative-plan tokens do not occupy positions in this reward sequence. Set $\widetilde{r}_k = r'_{i,t}$ on the final token of turn t, and zero elsewhere.

Let $\bar{V}_k$ be the cached, pre-update value prediction at s_k, with $\bar{V}_{K+1} = 0$. Token-level GAE (Schulman et al., 2015) and the value target are

$$\begin{aligned} \delta_k &= \widetilde{r}_k + \gamma \bar{V}_{k+1} - \bar{V}_k, \\ \widehat{A}_k &= \sum_{\ell=0}^{K-k} (\gamma\lambda)^\ell \delta_{k+\ell}, \qquad \widehat{G}_k = \widehat{A}_k + \bar{V}_k. \end{aligned} \tag{9}$$

We use $\gamma = 1$ and $\lambda = 0.95$. Cached values, raw advantages, and λ-return targets $\widehat{G}_k$ remain fixed during the two PPO epochs. The actor uses normalized advantages

$$A_k = \frac{\widehat{A}_k - \mu_{\mathcal{R}}}{\sigma_{\mathcal{R}} + 10^{-8}}, \tag{10}$$

where $\mu_{\mathcal{R}}$ and $\sigma_{\mathcal{R}}$ are the mean and population standard deviation over all actor tokens in the 16-dialogue rollout batch $\mathcal{R}$. Value targets use raw, not normalized, advantages.

Actor loss and reference regularization. The implementation averages tokens within each dialogue, then averages dialogues within each four-dialogue optimizer minibatch $\mathcal{B}$. For a token quantity $f_{n,k}$ in dialogue n, define

$$\langle f \rangle_{\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} \frac{1}{K_n} \sum_{k=1}^{K_n} f_{n,k}. \tag{11}$$

Suppressing the dialogue index, the PPO ratio and clipped actor loss (Schulman et al., 2017) to *minimize* are

$$\begin{aligned} \rho_k &= \frac{\pi_\theta(a_k \mid s_k)}{\pi_{\text{old}}(a_k \mid s_k)}, \\ \mathcal{L}_{\text{actor}} &= -\left\langle \min(\rho_k A_k, \text{clip}(\rho_k, 1-\epsilon, 1+\epsilon) A_k) \right\rangle_{\mathcal{B}} + \eta \mathcal{L}_{\text{ref}}, \end{aligned} \tag{12}$$

with $\epsilon = 0.2$ and $\eta = 0.02$. We use the nonnegative sampled reference penalty from DeepSeekMath (Shao et al., 2024, Eq. 4):

$$\mathcal{L}_{\text{ref}} = \left\langle e^{u_k} - u_k - 1 \right\rangle_{\mathcal{B}}, \qquad u_k = \log \pi_{\text{ref}}(a_k \mid s_k) - \log \pi_\theta(a_k \mid s_k). \tag{13}$$

These quantities are evaluated on the stored rollout tokens. Current, behavior, and reference token probabilities are normalized over the same public-schema grammar support used for generation, including the end-of-sequence decision. KL regularization is added to the loss, not to $\widetilde{r}_k$ or the value target.

Clipped value regression. Let $V_k = V_\psi(s_k)$ be the current value prediction. Following the clipped value-regression implementation described by Huang et al. (2022), with $\epsilon_V = 0.2$, the value loss is

$$\begin{aligned} V_k^{\text{clip}} &= \bar{V}_k + \text{clip}(V_k - \bar{V}_k, -\epsilon_V, \epsilon_V), \\ \mathcal{L}_V &= \frac{1}{2} \left\langle \max\left[(V_k - \widehat{G}_k)^2, (V_k^{\text{clip}} - \widehat{G}_k)^2 \right] \right\rangle_{\mathcal{B}}. \end{aligned} \tag{14}$$

The minimized loss is $\mathcal{L} = \mathcal{L}_{\text{actor}} + 0.5\mathcal{L}_V$. The value head receives detached policy hidden states, so value regression updates its own parameters but does not backpropagate through the policy backbone. The critic predicts λ-returns; the rule verifier supplies tentative-plan feasibility.

Information boundaries. Alternate-world state and undisclosed partner-private facts remain training-side. They are never injected into actor or reference-policy prompts. Facts disclosed by the partner become ordinary dialogue context. The hypothetical record is not treated as a demonstrated response. When the differing condition is on the tested role's own card, it is available from the first turn. Otherwise, the frozen local judge receives both partner cards and the recorded dialogue to identify the earliest partner disclosure; turn indices determine which earlier tentative plans remain gated off. This annotation approximates when the differing condition becomes available to the tested role.

Comparability. The main methods share the backbone, partner, terminal reward, generation limits, and test protocol. They use 384 training dialogues; group-based baselines use $G = 4$ and sample 48 scenarios, while PPO and PaCPO sample 96. PaCPO additionally generates tentative plans and applies cross-world rule checks. Comparisons match dialogue counts, not scenario coverage or total compute. Main comparisons use one training seed (917) and one test generation seed (317917).

Mechanism controls. The same-world control draws comparison candidates under the same world conditions. Shuffled credit positions preserve the computed credit values while changing their temporal placement. The disclosure-gate control retains the same disclosure annotations, validity checks, and disjointness condition, and applies paired credit regardless of whether the differing condition has appeared in both conversations. Table 5 reports all three comparisons.

Evaluation protocol. The main comparison, mechanism controls, and self-play use the same frozen Integrity rubric with Qwen3.8-Max. The seven SOTOPIA and three diagnostic rubrics are unchanged.

E.1 GiGPO and GraphGPO Adaptation

The GiGPO and GraphGPO rows in Table 3 use their released advantage-computation routines within the HICO dialogue trainer (Feng et al., 2025; Cheng et al., 2026). We load the numerical function bodies unchanged from a pinned `verl-agent` checkout; the environment interface, observation representation, rollout schedule, and policy-update loop use our HICO implementation.[1]

Rollout groups and task feedback. Each group contains $G = 4$ independent dialogues for the same scenario, world, and tested role, against the frozen Qwen3-8B partner. Each turn of the tested role, including the final record, is one step. The terminal reward R is assigned to the final step; earlier rewards are zero. The collection budget is 384 dialogues: 48 scenarios $\times$ two worlds $\times$ four samples, with balanced interaction families and preassigned roles. Both methods are critic-free and generate no tentative plans.

Anchor-state representation. An anchor key hashes the exact actor-input token sequence together with a flag distinguishing dialogue turns from final records. It therefore contains the actor's role card, public instructions, and full visible history. Matching is restricted to the same scenario, world, and role. We disable approximate string matching and semantic state merging; partner-private fields and verifier state are not added to the key. This keeps decision-relevant history intact while making shared anchors depend on identical visible prefixes.

GiGPO advantages. For a dialogue ending at step T, GiGPO computes $U_t = \sum_{u=t}^{T} 0.95^{u-t} r_u$, then compares these discounted returns among occurrences of the same anchor. The resulting step advantage is added to an episode-level advantage computed from R, with both coefficients equal to one. We use the released `mean_std_norm` mode: subtract the group mean and divide by the sample standard deviation plus 10^{-6}. Episode normalization retains the upstream cross-step default, repeating each dialogue's R at its steps when computing the statistics. Step normalization instead uses the corresponding anchor group; a singleton anchor has zero centered step advantage.

GraphGPO advantages. GraphGPO builds a directed graph separately for each rollout group. Nodes are the same exact-history anchors, and unit-weight edges connect successive actor steps. A dialogue with $R = 1$ connects to a shared success node; other terminal nodes remain trajectory-specific. The released routine computes shortest-path distances to success. With distance normalization disabled, the step score is $10(0.10)^{D(s_{t+1})}$, where D is the successor's distance to success. Unreachable nodes receive the upstream fallback distance, one greater than the largest finite distance. These scores are normalized within anchor groups and added to the same episode-level advantage, again with unit coefficients. This path score determines credit; it does not replace the terminal reward.

[1] Source revision: `20bd331bdbc9`. GiGPO uses `gigpo/core_gigpo.py`; GraphGPO uses `recipe/GraphGPO/core_graph.py`.

Policy update and scope. Each computed step advantage is broadcast to the tokens of that turn. Both adapters use the clipped actor loss and reference penalty above, with clipping 0.2, reference coefficient 0.02, actor learning rate 2×10^{-6}, and two optimization epochs. Tokens are averaged within dialogues before dialogues are averaged within four-dialogue optimizer minibatches. No further rollout-wide advantage normalization is applied, and partner tokens remain context only.

Exact-history grouping supports local comparisons only at repeated anchors. Once dialogue histories diverge, unmatched anchors retain the episode-level component without a local step comparison. The reported results characterize this exact-history HICO adaptation; the grouping and state representation are part of the comparison, alongside the advantage estimator.

E.2 Training Cost

Table 12 accounts for both dialogue participants, tentative-plan generation, disclosure annotation, and optimizer updates. Counts include generated outputs later excluded from training. All three runs sample 384 dialogues; PPO and PaCPO retain 384, while GRPO retains 376 after two four-dialogue groups encounter partner-generation failures. These correspond to the $G = 4$ GRPO and PaCPO checkpoints in the main comparison.

Table 12: **Training-generation and annotation cost.** Total output includes tested-role, partner, and tentative-plan tokens. Tentative-plan output is a subset of this total; disclosure requests are additional to terminal Integrity scoring.

Method	Total output tokens	Probe output tokens	Disclosure requests	Optimizer updates
PPO	276,293	0	0	192
GRPO ($G = 4$)	282,431	0	0	188
PaCPO	310,863	57,405	216	192

PaCPO generates 1,869 tentative plans, contributing 57,405 output tokens and 2,024,795 input tokens. Including both dialogue participants, its generated output is 12.51% higher than PPO's. Input counts sum unpadded request lengths, including repeatedly processed prefixes; judge tokens and policy optimization are separate costs. The 216 disclosure annotations complement 168 dialogues in which the differing condition is on the tested role's own card.

The local training scorer records 391, 386, and 603 scoring or annotation attempts for PPO, GRPO, and PaCPO, respectively, including validation retries. PaCPO's total includes 29 cached attempts from its first 16 dialogues, reused from an interrupted run before policy updates. These are counted once, yielding 574 new scorer attempts in the resumed run. Test-time Qwen3.8-Max evaluation is outside this training account.

F A Paired Scenario

The training scenario `branding_dilemma` illustrates authorization. Avery, a new business owner uncertain about a brand change, requests an extra concept-review block from consultant Blair. The public context states that complimentary hours require delegated authority and that a pending referral is not approval. It does not disclose Avery's requested duration or Blair's authority limit.

Table 13: **Paired authorization scenario.** Only Blair's private authorization limit changes.

Field	World w_0	World w_1
Avery's request	1 hour of extra review	1 hour of extra review
Blair's delegated limit	At most 2 hours	At most 0 hours
Referral authority	Agency lead	Agency lead
Appropriate record kind	Agreement	Defer
Approval / referral	True / none	False / agency lead

Both records must retain the requested one-hour parameter and the operation `approve_extra_review`. The first world admits current authorization; the second requires truthful referral rather than invented approval. Integrity additionally checks whether the conversation supports the recorded plan. The changed authority boundary determines whether approval or referral is feasible.

G Outcome Reliability and Information Diagnostics

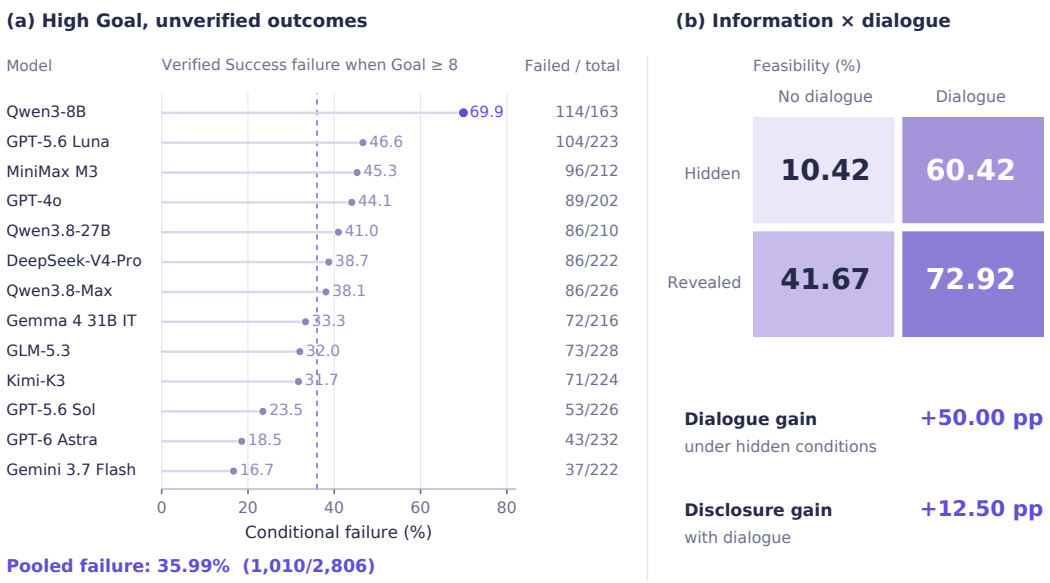

Figure 8: **Why verify outcomes?** (a) Verified Success failure rate conditioned on Goal ≥ 8, with failed/eligible counts for each model. The dashed line is the pooled rate: 1,010/2,806 model–task observations (35.99%). (b) Base under a 2×2 information-access and dialogue design: 12 development scenarios, 48 tasks per condition. Cells show rule-based Feasibility success; gains are percentage-point differences. Dialogue includes information exchange and negotiation.

High Goal does not certify valid outcomes. Among 163 Qwen3-8B conversations with Goal ≥ 8, 114 (69.94%) fail Verified Success. Across all thirteen models, 1,010 of 2,806 high-Goal model–task observations fail (35.99%; Figure 8a). These are repeated model–task observations, not distinct scenarios. A high role-goal rating does not certify a feasible, dialogue-supported outcome.

Information access and interaction. On 12 fixed development scenarios, Figure 8(b) separates information access from interaction. Without dialogue, revealing conditions raises Base Feasibility from 10.42% to 41.67%; with dialogue, it rises from 60.42% to 72.92%. Under hidden conditions, dialogue adds 50.00 points, combining information exchange and negotiation.

The diagnostic uses 12 fixed development scenarios, both worlds, and both tested roles, yielding 48 tasks per condition. No-dialogue conditions have no text on which to judge Integrity or SOTOPIA ratings, so those fields are marked not applicable rather than zero.

No dialogue removes information exchange and coordination. Revealing conditions restores the information while leaving agents to submit independently.

H Interactive Transfer and Self-Play

SOTOPIA protocol. Figure 4 (right) and Section 5.4 report direct transfer of HICO checkpoints without further training. We evaluate 70 environment–character pairs per partner setting, sampled

Table 14: **Information-access and dialogue diagnostic.** Conditions in Figure 8(b); all values are percentages. Own-record feasibility checks the tested role's record against the full scenario constraints, including both participants' conditions; Feasibility additionally requires both records to match. Both checks are rule-based. Verified Success here uses the original local Qwen3.8-27B judge, not Qwen3.8-Max.

Condition	Feasibility	Own-record feasibility	Ver. Succ.
No dialogue	10.42	25.00	–
Revealed + no dialogue	41.67	52.08	–
Hidden + dialogue	60.42	62.50	31.25
Revealed + dialogue	72.92	75.00	39.58

from 14 environments with a fixed selection seed. The two settings use the same checkpoint for both roles or GPT-4o as the partner. SOTOPIA (Zhou et al., 2024) uses iStar-style public prompts (Liu et al., 2026) and GPT-4o seven-dimensional scoring; the reported Goal is the tested Agent-1's score on the original 0–10 scale. Each partner setting uses one generation seed.

Complete self-play results. Table 15 supplements the Verified Success comparison in Figure 4 (left) with Feasibility, Integrity, and standard deviations across three generation seeds. Both participants use the same checkpoint; each seed comprises 120 dialogues and 240 scored role views.

Table 15: **Full self-play results.** Both participants use the same checkpoint. Mean ± sample SD over three generation seeds; 120 dialogues and 240 scored role views per seed. Rates are percentages; generation failures remain in the denominator.

Method	Feasibility	Integrity	Ver. Succ.
Base (Yang et al., 2025)	48.89 ± 2.68	29.17 ± 2.20	21.94 ± 2.10
PPO (Schulman et al., 2017)	48.89 ± 2.55	31.67 ± 2.20	27.22 ± 1.73
PaCPO	51.94 ± 2.68	35.83 ± 2.50	28.33 ± 2.20

I External-Task Training

We evaluate task-specific training in two settings: database-grounded recommendation and bilateral negotiation. Table 4 reports their results.

MultiWOZ: eliciting constraints for a recommendation. We select hotel and restaurant goals from MultiWOZ 2.2 dialogue-state annotations (Zang et al., 2020) and use the corresponding MultiWOZ 2.1 entity databases (Eric et al., 2020). Each source contributes one domain-specific recommendation task with explicit user constraints, such as area, price range, cuisine, or hotel amenities. We select goals with a database match and omit booking requests and requests for a preselected named entity. Train, development, and test retain their original source splits; repeated constraint profiles are removed across splits.

Each pair retains the public request and database while changing one private user constraint. Both worlds have valid recommendations, and their valid entity sets are disjoint. The visitor holds the requirements privately; the assistant asks questions, searches the database, and recommends one entity by name. A successful episode requires that the selected entity satisfy every private constraint and be unambiguously identified in the final message. This *constraint communication* endpoint tests whether elicited requirements reach the actual recommendation. The test set contains 77 paired scenarios and 154 dialogues.

CaSiNo: negotiating under private preferences. We retain CaSiNo's campsite setting (Chawla et al., 2021), with three units each of food, water, and firewood. Each participant has private high-,

medium-, and low-priority items, valued at 5, 4, and 3 points per unit, together with reasons for those preferences. To construct paired worlds, we keep the tested participant's card fixed and replace the partner's complete preference-and-reason card with one from the same original split. This changes which allocations benefit the partner while retaining the public negotiation setting.

The participants negotiate a complete allocation that both accept. We report agreement rate and normalized Nash surplus, based on the Nash bargaining product (Nash, 1950). We take the product of both participants' nonnegative utility gains above the five-point walk-away value and divide it by the maximum attainable product in that world. Episodes without agreement receive zero surplus. These measures distinguish reaching an agreement from finding a mutually beneficial allocation. The test set contains 43 paired scenarios and 86 dialogues.

J Family-Level Analysis

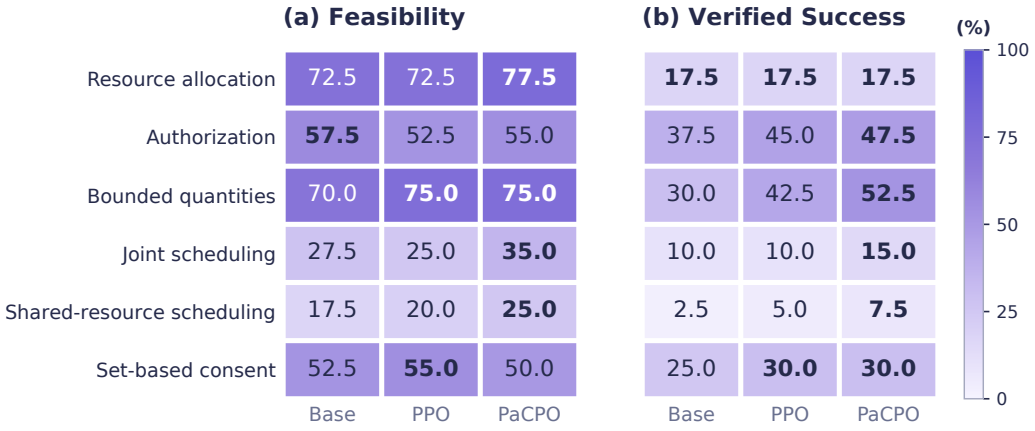

Figure 9: **Results by interaction family.** Each cell contains 40 test tasks (ten scenarios, both worlds and roles). Both panels use the same color scale; bold marks the best method within each row and panel. Scheduling remains difficult across methods.

Figure 9 reports Feasibility and Verified Success on 40 tasks per family: ten scenarios, two worlds, and two roles. All 60 test scenarios have disjoint valid outcome sets across their paired worlds.

Where outcome learning helps. PaCPO's largest Verified Success gain over Base occurs for bounded quantities, rising from 30.0% to 52.5%. It also improves authorization from 37.5% to 47.5%. For bounded quantities, PPO and PaCPO have identical Feasibility (75.0%), but PaCPO has higher Verified Success (52.5% versus 42.5%). The additional gain therefore lies in jointly satisfying feasibility and dialogue integrity, rather than increasing feasibility alone.

What remains difficult. Resource allocation illustrates the remaining gap: PaCPO reaches 77.5% Feasibility, yet Verified Success stays at 17.5%, the same as Base and PPO. Producing a feasible split is thus easier than maintaining dialogue support for it. Joint scheduling and shared-resource scheduling remain the lowest-scoring PaCPO families, at 15.0% and 7.5% Verified Success. These tasks require compatible time commitments across participants, and their low Feasibility rates indicate that feasible coordination itself remains difficult.

Table 16 complements these outcomes with the frequency of paired-credit activation during training.

Activation is lowest for scheduling: both tentative plans are frequently infeasible, producing a tie ($M_{ii} = M_{i,1-i} = 0$) and zero paired contrast.

Table 16: **Paired-credit activation.** Eligible positions have an open gate and two valid tentative plans; nonzero positions satisfy $M_{ii} \neq M_{i,1-i}$. Training uses 384 dialogues over 24 rounds.

Family	Eligible	Nonzero	Nonzero rate
Resource allocation	272	194	71.3%
Authorization	180	118	65.6%
Bounded quantities	264	194	73.5%
Joint scheduling	222	95	42.8%
Shared-resource scheduling	244	85	34.8%
Set-based consent	196	80	40.8%
All families	1378	766	55.6%

K Robustness and Human Validation

Generation and partner variation. The self-play experiment varies both participants and repeats generation with three seeds, holding the scenario set fixed. Figure 4 (left) reports Verified Success, and Table 15 gives the corresponding Feasibility and Integrity means and standard deviations.

K.1 Cross-Judge Agreement

We re-evaluate fixed model outputs with Kimi-K3 (Moonshot AI, 2026), retaining the Qwen3.8-Max Integrity messages, rubric, and evidence checks. Agreement between Qwen3.8-Max and Kimi-K3 is computed on the final binary Integrity decision for each model–task evaluation.

Subset selection. A task position identifies a scenario, world, and tested role. The eligible pool is the intersection of positions with an available frozen Integrity request for Base, PPO, PaCPO, and GRPO ($G = 4$): 239 positions. We stratify by the six interaction families, sort family names and then task keys lexicographically, and shuffle each family's keys using one Python random-number generator initialized with seed 20260924. Selecting the first ten keys per family without replacement yields 60 shared positions spanning 43 scenarios. Selection uses task identifiers and family labels, not Feasibility or Integrity scores, and is fixed before Kimi-K3 is queried. We reuse each method's output at these same positions without regenerating dialogues, giving 60 evaluations per method and 240 in total. Identical requests are scored once and reused, producing 203 unique Kimi-K3 requests; reported statistics give equal weight to all 240 model–task evaluations.

Table 17: **Qwen3.8-Max–Kimi-K3 Integrity agreement.** Both judges evaluate the same 240 model–task cases. Agreement compares final pass/fail decisions; pass rates show each judge's marginal acceptance rate.

Judge pair	Agreement (%)	Cohen's κ	Qwen3.8-Max pass (%)	Kimi-K3 pass (%)
Qwen3.8-Max / Kimi-K3	**90.00**	**0.783**	30.00	40.00

Qwen3.8-Max and Kimi-K3 agree on 216 of 240 decisions, with $\kappa = 0.783$ (Table 17). Kimi-K3 accepts 24 additional cases and rejects none that Qwen3.8-Max accepts.

Scoring interface. The recorded model identifiers are `qwen3.8-max` and `kimi-k3`. Qwen3.8-Max uses strict JSON-schema output with thinking disabled. Kimi-K3 uses JSON-object output with high reasoning effort and local validation against the same frozen schema. Field normalization renames `explanation` to `reason`, removes redundant speaker fields from evidence, and fills omitted empty fields for negative flags. Format-repair prompts instruct the judge to retain substantive judgments and cited evidence. All 240 evaluations per judge yield valid final decisions; original responses and repairs are retained.

K.2 Human Review

Protocol. Two expert annotators review 48 dialogues from Base and PaCPO, covering 12 scenarios and both of their worlds, which form 24 matched Base–PaCPO pairs. For each dialogue, annotators see the public context, both role cards, the full conversation, and both records, with model identities hidden. They judge whether the final plan satisfies the stated conditions (Feasibility) and whether the participants' utterances support it (Integrity). For each matched pair, presented in random order (Zheng et al., 2023), they also judge which model provides more adequate help to the partner, citing turn-level evidence.

Results. Human Feasibility judgments agree with the executable checks on 100% of dialogues, and human Integrity judgments agree with Qwen3.8-Max on 82.22%. In the helpfulness comparison, annotators prefer PaCPO in 8 pairs and Base in 5, rate the two as similar in 3, and judge both inadequate in 8. PaCPO thus receives more favorable judgments, while one third of the pairs reveal cases in which neither model helps adequately.

L Limitations and Future Work

Scope of outcome verification. HICO-Bench covers six coordination families whose arrangements can be represented as structured records and checked against explicit conditions. This supports reproducible feasibility tests, but does not cover the full range of social interaction, such as emotional support or relationships without a well-defined final outcome. Extending the benchmark requires identifying verifiable outcomes without reducing these interactions to inappropriate formal objectives. Feasibility success is feasibility under the specified conditions, not a general judgment of social appropriateness.