SAVOIR: Learning Social Savoir-Faire via Shapley-based Reward Attribution

Anonymous ACL submission

Abstract

Social intelligence, the ability to navigate complex interpersonal interactions, presents a fundamental challenge for language agents. Training such agents via reinforcement learning requires solving the credit assignment problem: determining how individual utterances contribute to multi-turn dialogue outcomes. Existing approaches directly employ language models to distribute episode-level rewards, yielding attributions that are retrospective and lack theoretical grounding. We propose SAVOIR (**S**h**A**pley **V**alue f**O**r Soc**I**al **R**L), a novel principled framework grounded in cooperative game theory. Our approach combines two complementary principles: *expected utility* shifts evaluation from retrospective attribution to prospective valuation, capturing an utterance's strategic potential for enabling favorable future trajectories; *Shapley values* ensure fair credit distribution with axiomatic guarantees of efficiency, symmetry, and marginality. Experiments on the SOTOPIA benchmark demonstrate that SAVOIR achieves new state-of-the-art performance across all evaluation settings, with our 7B model matching or exceeding proprietary models including GPT-4o and Claude-3.5-Sonnet. Notably, even large reasoning models consistently underperform, suggesting social intelligence requires qualitatively different capabilities than analytical reasoning.[1]

1 Introduction

Social intelligence, the capacity to navigate complex interpersonal interactions and achieve social goals, is fundamental to human cognition and increasingly critical for artificial agents (Gweon et al., 2023; Lee et al., 2024). As large language models (LLMs) become integrated into applications requiring negotiation, collaboration, and persuasion, their ability to exhibit socially intelligent behavior has attracted substantial research attention (Zhou et al., 2024; Park et al., 2023; Yang et al., 2024). Yet despite this growing interest, improving the social intelligence of AI systems remains challenging: social interactions are inherently multi-turn, involve competing objectives between participants, and require nuanced understanding of how individual utterances contribute to long-term outcomes (Mathur et al., 2024; Li et al., 2024b).

Recent work has begun addressing these challenges through reinforcement learning (RL) approaches. Wang et al. (2024) propose SOTOPIA-π, which combines behavior cloning with self-reinforcement on filtered interaction data. More recently, Yu et al. (2025) introduce Sotopia-RL, which refines episode-level feedback into utterance-level rewards by directly prompting an LLM for credit assignment. While Sotopia-RL demonstrates improved performance, its approach exhibits two fundamental limitations. First, the credit assignment mechanism lacks theoretical grounding; the LLM distributes rewards heuristically without principled guarantees of fairness or accuracy. Second, and more critically, the reward model performs *retrospective attribution*: it assigns credit based on what an utterance contributed to the observed outcome, rather than evaluating its *strategic value* for enabling favorable future trajectories. This distinction matters because socially intelligent behavior often involves utterances whose immediate contribution appears minimal but whose strategic positioning unlocks subsequent success.

To address these limitations, we propose SAVOIR (**S**h**A**pley **V**alue f**O**r Soc**I**al **R**L), a theoretically grounded framework that reconceptualizes credit assignment through two complementary principles from game theory. First, we adopt **expected utility** to shift the evaluation focus from retrospective attribution to prospective valuation. Rather than asking "what did this utterance contribute to the final outcome?", we ask "what is the expected value of future interactions given this utterance?" By com-

[1] Code and models will be released upon publication.

puting expected outcomes over all possible partner responses and subsequent dialogue trajectories, we capture an utterance's *strategic potential*, its capacity to establish favorable conditions for future success. Second, we employ **Shapley values** from cooperative game theory to distribute this strategic value fairly across utterances. The Shapley value provides the unique attribution method satisfying efficiency, symmetry, and marginality axioms (Lundberg and Lee, 2017), ensuring that utterances receive credit proportional to their true marginal contribution across all possible orderings. Together, these principles transform credit assignment from a heuristic into a principled computation: expected utility defines *what* we measure (forward-looking strategic value), while Shapley values determine *how* we distribute it (fair, axiomatic attribution).

We evaluate SAVOIR comprehensively on the SOTOPIA benchmark (Zhou et al., 2024), comparing against proprietary LLMs, large reasoning models, and state-of-the-art social intelligence methods. Experiments demonstrate that SAVOIR achieves new state-of-the-art performance across all evaluation settings: on SOTOPIA-Hard with GPT-4o as partner, the most challenging setting, SAVOIR obtains a Goal score of 7.18, improving over the strongest baseline by 7.5%. Notably, our 7B model matches or exceeds proprietary LLMs including GPT-4o and Claude-3.5-Sonnet, while large reasoning models (OpenAI-o1, Gemini-2.5-Pro, DeepSeek-R1) consistently underperform despite their strong analytical capabilities, suggesting that social intelligence requires qualitatively different skills. Human evaluation with expert annotators further validates that SAVOIR produces more strategic responses and that its reward model better captures nuanced credit assignment.

Our contributions are threefold:

- We propose SAVOIR, a theoretically grounded framework for social RL that combines expected utility for prospective valuation with Shapley values for fair credit attribution.
- We demonstrate state-of-the-art performance on SOTOPIA benchmarks, with a 7B model matching proprietary LLMs and revealing that reasoning models underperform on social tasks.
- We provide extensive analysis including human evaluation, ablation studies, and case studies that validate the effectiveness of our principled credit assignment approach.

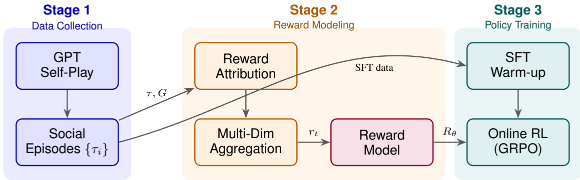

Figure 1: Overview of the social agent training pipeline. **Stage 1**: Collect social interaction episodes through LLM self-play. **Stage 2**: Design utterance-level, multi-dimensional rewards through attribution and aggregation. **Stage 3**: Train the policy via supervised fine-tuning followed by online reinforcement learning with the learned reward model.

2 Preliminaries

This section provides the foundational concepts for our work. We first present the training pipeline for social agents (§2.1), then formalize the social interaction task (§2.2), and finally describe the evaluation framework (§2.3).

2.1 Training Pipeline Overview

Figure 1 illustrates the standard training pipeline for social agents. The process consists of three stages: (1) **data collection**, where LLM agents engage in self-play to generate social interaction episodes; (2) **reward modeling**, where episode-level outcomes are attributed to individual utterances and aggregated across multiple evaluation dimensions; and (3) **policy training**, where the agent is first warmed up through supervised fine-tuning and then optimized via online reinforcement learning using the trained reward model.

2.2 Task Formulation

Social interaction can be formalized as a partially observable Markov decision process (POMDP), defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, Z, R \rangle$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ the action space, $\mathcal{O}$ the observation space, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ the transition function, $Z : \mathcal{S} \to \mathcal{O}$ the observation function, and $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward function. A social episode with T turns is represented as $\tau = (o_0, a_0, o_1, a_1, \ldots, o_T)$, where $o_t \in \mathcal{O}$ is the dialogue history observed at turn t and $a_t \in \mathcal{A}$ is the utterance generated by the agent. Given a private goal g, the agent samples actions according to its policy $\pi_\theta(\cdot \mid o_t, g)$.

Reward Modeling. The central challenge lies in designing effective reward signals. Given an episode τ and goal g, an LLM-based evaluator pro-

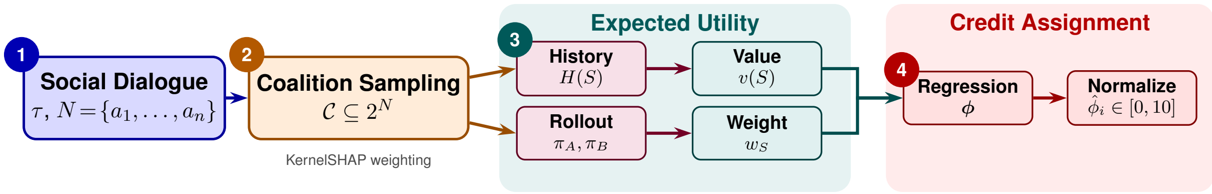

Figure 2: Overview of the SAVOIR framework. **Step 1**: Input social dialogue τ with agent utterances $N = \{a_1, \ldots, a_n\}$. **Step 2**: Sample coalitions $\mathcal{C}$ using KernelSHAP weighting. **Step 3**: For each coalition S, reconstruct history $H(S)$, perform rollouts to compute value $v(S)$, and derive SHAP weight w_S. **Step 4**: Solve weighted regression to obtain Shapley values $\boldsymbol{\phi}$, then normalize to $[0, 10]$.

vides an episode-level score $G = f(\tau, g) \in \mathbb{R}$. However, episode-level rewards offer only coarse supervision. To obtain fine-grained signals, we attribute the outcome to individual utterances: $r_t = G \cdot \mathcal{A}(a_t, \tau)$, where $\mathcal{A}(a_t, \tau) \in [0, 1]$ represents the contribution of utterance a_t to the episode outcome, estimated by an LLM with access to the full dialogue context. Furthermore, social interactions are inherently multi-dimensional. Beyond goal completion, utterances may contribute to relationship building, knowledge exchange, or other social objectives. We aggregate rewards across D dimensions: $r_t = \frac{1}{D} \sum_{d=1}^{D} w_d \cdot \tilde{r}_{t,d}$, where $\tilde{r}_{t,d}$ is the normalized reward for dimension d and w_d is its corresponding weight.

2.3 SOTOPIA Evaluation Suite

SOTOPIA (Zhou et al., 2024) provides an open-ended environment for evaluating social intelligence. Agents role-play through social scenarios, including negotiation, persuasion, collaboration, and accommodation, each with private goals hidden from the interaction partner. The environment evaluates agent performance along seven dimensions: **Goal Completion** (GOAL), the primary metric measuring task success; **Believability** (BEL), consistency with the assigned persona; **Relationship** (REL), maintenance of positive rapport; **Knowledge** (KNO), appropriate information exchange; **Secret** (SEC), protection of private information; **Social Rules** (SOC), adherence to social norms; and **Financial** (FIN), material outcomes when applicable. This multi-dimensional evaluation enables comprehensive assessment of social intelligence, capturing both outcome-oriented success and process-oriented interaction quality.

3 Method

Building upon the preliminaries, we present SAVOIR (**S**h**A**pley **V**alue f**O**r Soc**I**al **RL**), a principled framework for computing utterance-level rewards in social interactions. Just as *savoir-faire*, the French term for social grace, captures the art of knowing how to act appropriately in social situations, SAVOIR teaches language agents this skill through game-theoretic reward attribution. Our approach leverages two fundamental concepts: *expected utility* for evaluating strategic potential and *Shapley value* for fair credit assignment. We first provide an overview of our framework (§3.1), then detail the expected utility formulation (§3.2), the Shapley value-based credit assignment (§3.3), and the efficient computation via KernelSHAP (§3.4). Finally, we describe the reward model training procedure (§3.5).

3.1 Framework Overview

The core challenge in reward modeling for social interactions lies in attributing episode-level outcomes to individual utterances. Existing approaches either use coarse episode-level rewards or rely on heuristic credit assignment, both of which fail to capture the strategic nature of social dialogue. We address this challenge by formulating reward computation as a cooperative game where each utterance is a player contributing to the collective outcome.

Figure 2 illustrates the SAVOIR framework. Given a dialogue τ containing n utterances from the target agent, denoted as $N = \{a_1, \ldots, a_n\}$, our goal is to compute a reward ϕ_i for each utterance a_i that reflects its strategic contribution. The framework operates in three stages: (1) sampling coalitions of utterances, (2) evaluating the expected utility of each coalition through rollouts, and (3) computing Shapley values to distribute credit.

3.2 Expected Utility for Strategic Evaluation

Motivation. Traditional reward attribution methods evaluate utterances based on their historical contribution to the final outcome. However, in strategic social interactions, the value of an utterance lies not only in what has been achieved but also in what *can be achieved* from the current state. For instance, a well-crafted proposal may open pathways to favorable outcomes that are not immediately apparent. To capture this forward-looking perspective, we adopt expected utility theory from decision science, which evaluates actions based on their anticipated future value.

Formulation. We define a value function $v : 2^N \to \mathbb{R}$ that maps any subset (coalition) of utterances $S \subseteq N$ to a scalar value representing its strategic worth. Formally, for a coalition S, the value function is defined as:

$$v(S) = \mathbb{E}_{\tau' \sim \mathcal{R}(H(S))} \left[U(\tau') \right], \tag{1}$$

where $H(S)$ denotes the reconstructed dialogue history containing only utterances in S along with their corresponding partner responses, $\mathcal{R}(H(S))$ represents the distribution over future dialogue trajectories starting from state $H(S)$, and $U(\tau')$ is the utility of a complete trajectory.

Future Rollout. To compute the expectation in Eq. 1, we perform Monte Carlo simulation. Starting from the reconstructed history $H(S)$, we conduct J complete dialogues using the agent policy π_A and a partner policy π_B:

$$v(S) = \frac{1}{J} \sum_{j=1}^{J} U(\tau_j), \tag{2}$$

where each τ_j is a complete trajectory obtained by alternating between π_A and π_B until the dialogue terminates.

Utility Function. The utility $U(\tau)$ of a trajectory is computed using the SOTOPIA evaluation framework, which provides scores across multiple dimensions. We aggregate these dimensions using a weighted combination following Yu et al. (2025):

$$U(\tau) = \sum_{d=1}^{D} w_d \cdot G_d(\tau), \tag{3}$$

where $G_d(\tau)$ is the score for dimension d and w_d is its corresponding weight. This formulation allows flexible emphasis on different social objectives such as goal completion, relationship maintenance, or norm adherence.

3.3 Shapley Value for Credit Assignment

Motivation. With the value function defined, we now face the credit assignment problem: how to fairly distribute the total value among individual utterances? Consider a negotiation where multiple utterances collectively lead to a successful agreement. Some utterances may establish rapport, others may introduce key proposals, and still others may handle objections. A principled attribution method should recognize the unique contribution of each utterance, accounting for its synergistic effects with other utterances.

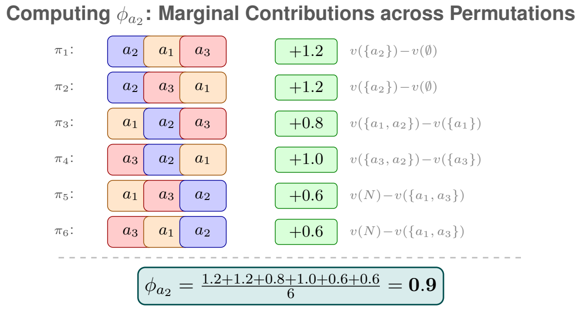

Figure 3: Shapley value computation for a_2. For each of the $n! = 6$ permutations, we compute a_2's marginal contribution when it joins. The Shapley value is the average across all permutations. See Appendix A for detailed explanation.

Formulation. The Shapley value from cooperative game theory provides an axiomatic solution to this problem. For a cooperative game defined by a player set N and a value function v, the Shapley value ϕ_i of player i is the weighted average of its marginal contributions across all orderings:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \left[v(S \cup \{i\}) - v(S) \right]. \tag{4}$$

The term $v(S \cup \{i\}) - v(S)$ represents the marginal contribution of utterance a_i to coalition S, and the coefficient ensures that each ordering is weighted equally. The Shapley value satisfies four desirable properties: *efficiency* (the values sum to $v(N) - v(\emptyset)$), *symmetry* (identical contributions receive identical values), *null player* (zero contribution implies zero value), and *additivity* (values are additive across games).

Interpretation. In our context, the Shapley value ϕ_i quantifies the average marginal contribution of utterance a_i to the expected future utility. A high Shapley value indicates that the utterance consistently improves outcomes when added to various coalitions, suggesting strong strategic value. Conversely, a low or negative value indicates that the utterance may be redundant or even detrimental.

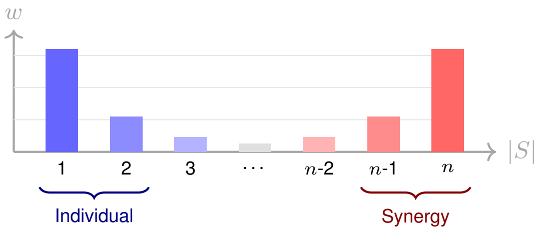

Figure 4: SHAP kernel weight distribution. Extreme coalition sizes (small: individual effects; large: synergy effects) receive higher weights, enabling efficient Shapley approximation.

3.4 Efficient Computation via KernelSHAP

Computational Challenge. Direct computation of Shapley values requires evaluating $v(S)$ for all 2^n subsets, which is computationally prohibitive for dialogues with many utterances. Moreover, each evaluation of $v(S)$ requires J rollout simulations, further compounding the cost.

KernelSHAP Algorithm. To address this challenge, we employ KernelSHAP (Lundberg and Lee, 2017), which reformulates Shapley value computation as a weighted linear regression. The insight is that Shapley values can be obtained by solving:

$$\boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\phi}} \sum_{k=1}^{K} w_k \left(v(S_k) - \sum_{i=1}^{n} \phi_i \cdot z_{ki} \right)^2, \quad (5)$$

where $\{S_k\}_{k=1}^K$ are sampled coalitions, $z_{ki} \in \{0, 1\}$ indicates whether utterance a_i is in coalition S_k, and w_k is the SHAP kernel weight:

$$w_k = \frac{n-1}{\binom{n}{|S_k|} \cdot |S_k| \cdot (n - |S_k|)}. \quad (6)$$

The kernel weight assigns higher importance to coalitions of extreme sizes (small or large), as these provide the most informative marginal contributions. This weighting scheme ensures that the regression solution converges to the true Shapley values. Figure 4 illustrates this weight distribution.

Smart Coalition Sampling. Rather than uniform sampling, we prioritize coalitions at extreme sizes, as shown in Figure 4. Coalitions containing only one or two utterances reveal individual contributions, while coalitions missing only one or two utterances reveal synergistic effects. This strategy improves estimation accuracy with limited budget. Algorithm 5 summarizes the complete SAVOIR reward computation procedure. A detailed walk-through example is provided in Appendix B.

Algorithm 1: SAVOIR Reward Computation

Input: Dialogue τ, utterances $N = \{a_1, \ldots, a_n\}$, policies π_A, π_B, rollouts J, samples K
Output: Normalized rewards $\{\hat{\phi}_i\}_{i=1}^n$

// Step 1: Coalition Sampling (KernelSHAP)
$\mathcal{C} \leftarrow$ `SampleCoalitions`(N, K) ▷ Prioritize extreme sizes

// Step 2: Expected Utility Computation
for each $S \in \mathcal{C}$ **do**
$H(S) \leftarrow$ `ReconstructHistory`(τ, S)
$v(S) \leftarrow 0$
for $j = 1$ **to** J **do**
$\tau_j \leftarrow$ `Rollout`$(H(S), \pi_A, \pi_B)$ ▷ Future simulation
$v(S) \leftarrow v(S) + U(\tau_j)/J$ ▷ Monte Carlo estimate
$w_S \leftarrow$ `SHAPWeight`$(|S|, n)$

// Step 3: Shapley Value via Regression
$\boldsymbol{\phi} \leftarrow$ `WeightedRegression`$(\{(S, v(S), w_S)\}_{S \in \mathcal{C}})$ ▷ Credit assignment

// Step 4: Normalization
$\hat{\phi}_i \leftarrow 10 \cdot \frac{\phi_i - \min_j \phi_j}{\max_j \phi_j - \min_j \phi_j}$ for all i ▷ Scale to $[0, 10]$

return $\{\hat{\phi}_i\}_{i=1}^n$

Figure 5: SAVOIR reward computation procedure.

3.5 Reward Model Training

Training Data Construction. Using the SAVOIR algorithm, we compute normalized rewards for utterances across a corpus of social interaction episodes. Each training instance consists of a dialogue context c (including scenario, goals, and dialogue history), an utterance a, and its SAVOIR score $\hat{\phi}$. This creates a dataset $\mathcal{D} = \{(c, a, \hat{\phi})\}$ for reward model training.

Reward Model Architecture. We train a reward model R_θ that takes a context-utterance pair and predicts its reward: $R_\theta(c, a) = \text{MLP}(\text{LLM}_\theta([c; a]))$, where LLM_θ is a pretrained language model that encodes the concatenated input, and MLP is a multi-layer perceptron that projects the representation to a scalar reward.

Training Objective. We train the reward model using mean squared error between predicted and target rewards: $\mathcal{L}_{\text{RM}} = \mathbb{E}_{(c,a,\hat{\phi}) \sim \mathcal{D}} \left[\left(R_\theta(c, a) - \hat{\phi} \right)^2 \right]$. The trained reward model provides dense, utterance-level feedback during reinforcement learning, enabling fine-grained policy optimization.

Model	*Self-Play*				*GPT-4o-as-Partner*			
	SOTOPIA-ALL		SOTOPIA-HARD		SOTOPIA-ALL		SOTOPIA-HARD	
	GOAL↑	AVG↑	GOAL↑	AVG↑	GOAL↑	AVG↑	GOAL↑	AVG↑
Proprietary LLMs								
GPT-4o	8.19	3.76	6.97	3.46	8.19	3.76	6.97	3.46
Claude-3.5-Sonnet	8.29	3.71	6.33	3.09	**8.42**	3.77	6.64	3.30
DeepSeek-V3	8.15	3.62	6.34	3.09	8.14	3.72	6.69	3.31
Large Reasoning Models								
OpenAI-o1	7.93	3.58	5.69	2.71	8.09	3.69	6.65	3.20
OpenAI-o3-mini	7.38	3.30	5.14	2.36	7.96	3.61	6.33	2.98
Gemini-2.5-Pro	7.85	3.43	5.67	2.55	8.12	3.59	6.70	3.09
DeepSeek-R1	7.97	3.40	5.86	2.73	7.92	3.49	6.20	2.95
QwQ-32B	7.70	3.30	5.35	2.41	7.80	3.47	6.19	2.91
Social Intelligence Methods								
Qwen2.5-7B-Instruct	7.91	3.55	6.21	3.01	6.71	3.13	5.90	2.90
+ PPDPP (Deng et al., 2024)	7.97	3.65	6.63	3.31	8.07	3.71	6.76	3.35
+ EPO (Liu et al., 2025)	8.09	3.51	6.82	3.12	8.41	3.86	6.81	**3.51**
+ DAT (Li et al., 2024a)	7.97	3.59	6.39	3.10	8.11	3.70	6.78	3.36
+ DSI (Zhang et al., 2025)	8.35	3.75	7.31	3.51	8.15	3.70	6.87	3.42
+ Sotopia-RL (Yu et al., 2025)	7.80	3.55	7.81	3.80	8.31	3.90	6.68	3.29
+ SAVOIR (Ours)	**8.43**	**3.85**	**7.93**	**3.97**	**8.42**	**3.94**	**7.18**	**3.51**

Table 1: Main results on SOTOPIA benchmarks. **Bold**: best; underline: second-best. Shaded cells indicate top performers. Reasoning models consistently underperform, while SAVOIR achieves SOTA across all settings.

4 Experimental Setup

Benchmarks. We evaluate on SOTOPIA (Zhou et al., 2024), using two splits: (1) **SOTOPIA-Hard**, 14 challenging scenarios requiring sophisticated strategic reasoning, and (2) **SOTOPIA-All**, 90 scenarios for comprehensive evaluation.

Evaluation Protocol. Following Zhou et al. (2024); Wang et al. (2024), we use GPT-4o as evaluator, with GOAL (0–10) as primary metric and AVG as holistic measure. We evaluate under two settings: **Self-Play**, where the trained agent interacts with itself, and **GPT-4o-as-Partner**, where the agent interacts with GPT-4o to test generalization to unseen partners.

Baselines. We compare against three categories: (1) **Proprietary LLMs** (GPT-4o, Claude-3.5-Sonnet, DeepSeek-V3); (2) **Large Reasoning Models** (OpenAI-o1, o3-mini, Gemini-2.5-Pro, DeepSeek-R1, QwQ-32B); and (3) **Social Intelligence Methods** including PPDPP (Deng et al., 2024), EPO (Liu et al., 2025), DAT (Li et al., 2024a), DSI (Zhang et al., 2025), SOTOPIA-π (Wang et al., 2024), and Sotopia-RL (Yu et al., 2025).[2] See Appendix C for details.

[2]Sotopia-RL results are reproduced using official code under the same GPU constraints for fair comparison.

Implementation. We implement SAVOIR on Qwen2.5-7B-Instruct. Training follows two stages: SFT on GPT-4o self-play episodes, then online RL using GRPO (Shao et al., 2024) with our reward model. For SAVOIR computation, coalition samples scale adaptively with dialogue length (capped at 200), with $J = 2$ rollouts each. Full details in Appendix D.

5 Results

5.1 Main Results

Table 1 presents results across SOTOPIA benchmarks. SAVOIR achieves state-of-the-art performance across all settings, obtaining 7.18 GOAL on SOTOPIA-Hard with GPT-4o as partner (7.5% over Sotopia-RL) and 7.93 GOAL in Self-Play (outperforming DSI at 7.31 and Sotopia-RL at 7.81). Despite being a 7B model, SAVOIR matches or exceeds proprietary LLMs: on Self-Play SOTOPIA-All, SAVOIR (8.43) outperforms GPT-4o (8.19) and Claude-3.5-Sonnet (8.29), with 13.8% gains on SOTOPIA-Hard.

A striking finding is that large reasoning models consistently underperform. OpenAI-o1, o3-mini, Gemini-2.5-Pro, DeepSeek-R1, and QwQ-32B all score below SAVOIR; for instance, o3-mini achieves only 5.14 GOAL versus 7.93 for SAVOIR (54.3% gap). This suggests analytical reasoning

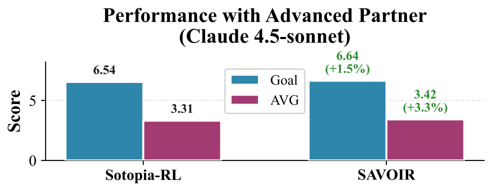

Figure 6: Performance on SOTOPIA-Hard with Claude 4.5-sonnet as interaction partner.

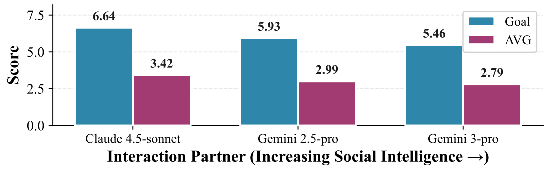

Figure 7: Performance degradation as partner social intelligence increases.

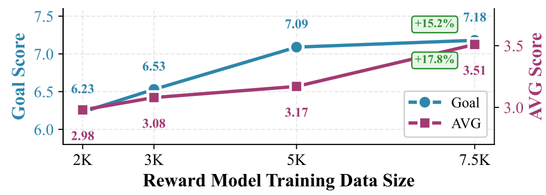

Figure 8: Effect of training data scale. Both Goal and Avg improve consistently from 2K to 7.5K episodes.

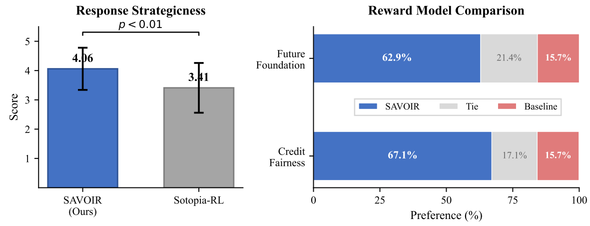

Figure 9: Human evaluation results on SOTOPIA-Hard (14 scenarios, 5 expert annotators). **Left**: Response strategicness ratings (1–5 scale) with standard deviation bars. **Right**: Pairwise preference for reward model quality. SAVOIR significantly outperforms the baseline across all dimensions. Inter-annotator agreement: Fleiss' $\kappa = 0.52$ (moderate).

may hinder social performance, which requires intuitive responses rather than deliberative chains. Among social intelligence methods, SAVOIR improves over Sotopia-RL by 1.3–8.1%, validating that utterance-level Shapley attribution provides more meaningful signal than episode-level rewards.

5.2 Robustness Against Advanced Partners

We evaluate robustness by testing against advanced interaction partners. On SOTOPIA-Hard with Claude 4.5-sonnet (Figure 6), SAVOIR outperforms Sotopia-RL on both GOAL (6.64 vs 6.54, +1.5%) and AVG (3.42 vs 3.31, +3.3%), confirming that Shapley-based credit assignment transfers effectively to stronger partners.

To probe generalization limits, we evaluate against increasingly capable partners (Figure 7). Performance degrades with partner sophistication: compared to Claude 4.5-sonnet, GOAL scores decline 10.7% against Gemini-2.5-Pro and 17.8% against Gemini-3-Pro, motivating future work on curriculum learning with diverse partner policies.

5.3 Effect of Reward Model Training Data Scale

We investigate how training corpus size affects reward model quality by varying annotated episodes and evaluating on SOTOPIA-Hard with GPT-4o as partner (Figure 8). Scaling from 2K to 7.5K episodes yields substantial improvements: GOAL increases from 6.23 to 7.18 (+15.2%) and AVG from 2.98 to 3.51 (+17.8%). The most significant gains occur between 3K and 5K episodes (+8.6% GOAL), suggesting a critical threshold for learning nuanced credit assignment. The consistent scaling behavior indicates that further data collection could yield additional improvements.

5.4 Human Evaluation

We conduct human evaluation with five expert annotators assessing 14 SOTOPIA-Hard scenarios across three dimensions: response strategicness (1–5 scale), credit fairness, and future foundation identification (Figure 9). Full setup details are in Appendix F.

On response strategicness, SAVOIR achieves 4.06 ($\pm$0.72) versus Sotopia-RL's 3.41 ($\pm$0.85), a 19.1% improvement ($p < 0.01$). For reward model quality, annotators prefer SAVOIR's credit assignment in 67.1% of fairness comparisons (vs. 15.7% baseline) and 62.9% of future foundation comparisons (vs. 15.7% baseline). Inter-annotator agreement (Fleiss' $\kappa = 0.52$) indicates moderate reliability. These results confirm that Shapley-based attribution aligns better with human intuition about social strategy.

5.5 Case Study: Reward Model Comparison

To qualitatively evaluate whether the SAVOIR reward model captures nuanced social reasoning, we present representative cases comparing model assessments.

Rewarding Honesty in Sales. This case demonstrates SAVOIR's ability to recognize sales strategies that balance transparency with persuasion.

Scenario: Furniture Sale with Defect

Context: Noah is selling a Crate & Barrel furniture set. The dresser is missing a handle, but the detached handle is available.

Seller's Goal: Sell for > $172.5.

Buyer (Ethan): *"Hi! I saw your listing for the Crate & Barrel dresser and nightstand set. I'm interested in it. Can you tell me a bit more about the condition? I saw you mentioned there's a handle missing from the dresser."*

Seller (Noah): *"Hello Ethan, glad you're interested! The set is overall in good condition, but like I mentioned, the dresser is missing one handle. Fortunately, we have the detached handle available. If you're okay with the condition and the small imperfection, I think the set is definitely worth the investment."*

Analysis. The seller's response exemplifies effective social intelligence: it establishes rapport ("glad you're interested"), demonstrates honesty by acknowledging the defect, immediately offers a solution ("detached handle available"), and reinforces value ("worth the investment"). SAVOIR correctly identifies this "transparency + solution" pattern as a high-quality strategy, while baseline models often misinterpret defect mentions as negative sentiment without understanding the strategic framing.

This case illustrates two key advantages of the SAVOIR reward model: (1) **Strategic Recognition**, identifying and rewarding sophisticated social strategies (honesty + solution framing) that baseline models miss; (2) **Nuanced Attribution**, distinguishing between surface-level politeness and genuinely effective social moves. Additional case studies in Appendix E demonstrate context sensitivity in relationships (where SAVOIR avoids over-rewarding surface politeness), strategic negotiation tactics, and multi-turn planning.

6 Related Work

Social Reasoning in Language Models. Social reasoning, the ability to understand and navigate interpersonal dynamics, constitutes a fundamental aspect of human intelligence (Lee et al., 2024; Gweon et al., 2023). As large language models become increasingly integrated into social applications, evaluating and improving their social capabilities has emerged as a critical research direction (Mathur et al., 2024; Li et al., 2024b). Gandhi et al. (2023) demonstrate that while advanced models like GPT-4 exhibit theory-of-mind capabilities resembling human inference patterns, significant gaps remain compared to human performance. This motivates the development of frameworks for studying social intelligence in AI systems.

Benchmarks and Evaluation Frameworks. To address the evaluation challenge, researchers have developed interactive environments that simulate realistic social scenarios. SOTOPIA (Zhou et al., 2024) introduces an open-ended platform where agents pursue social goals through role-play interactions, providing an evaluation framework for social intelligence. Building upon this foundation, SocialEval (Zhou et al., 2025) extends evaluation to both outcome-oriented goal achievement and process-oriented interpersonal abilities. These infrastructures enable systematic assessment of how language models navigate social interactions.

Reinforcement Learning for Social Intelligence. Reinforcement learning offers a natural paradigm for training socially intelligent agents, as it enables learning through interaction without requiring extensive human annotations (Ndousse et al., 2021). Recent work has explored various RL-based approaches for social agents. SOTOPIA-π (Wang et al., 2024) combines behavior cloning with self-reinforcement training. SDPO (Kong et al., 2025) introduces segment-level preference optimization for multi-turn social dialogues. Sotopia-RL (Yu et al., 2025) proposes utterance-level multi-dimensional rewards for fine-grained credit assignment. AML (Zhu et al., 2025) further advances this direction by enabling adaptive reasoning depth selection during social interactions.

7 Conclusion

We presented SAVOIR, a framework applying cooperative game theory to credit assignment in social RL. Using expected utility for prospective valuation and Shapley values for fair attribution with axiomatic guarantees, SAVOIR achieves state-of-the-art performance on SOTOPIA, with our 7B model notably matching proprietary GPT-4o. The consistent underperformance of large reasoning models reveals that social intelligence requires qualitatively distinct capabilities from analytical reasoning. Human evaluation confirms that our approach produces more strategic responses with better credit assignment, and we hope this work inspires further exploration bridging game theory and social AI.

Limitations

Our work has several limitations. First, performance degrades with increasingly capable partners (e.g., Gemini 3-pro), suggesting that training on a fixed partner distribution may not generalize to superior social reasoners; curriculum learning could address this. Second, our evaluation focuses on English interactions within SOTOPIA; since social intelligence is culture-dependent, extending to multilingual and cross-cultural settings remains important for broader applicability.

References

Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. 2024. Plug-and-play policy planner for large language model powered dialogue agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023. Understanding social reasoning in language models with language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Hyowon Gweon, Judith Fan, and Been Kim. 2023. Socially intelligent machines that learn from humans and help humans learn. *Philosophical Transactions of the Royal Society A*, 381(2251):20220048.

Aobo Kong, Wentao Ma, Shiwan Zhao, Yongbin Li, Yuchuan Wu, Ke Wang, Xiaoqian Liu, Qicheng Li, Yong Qin, and Fei Huang. 2025. Sdpo: Segment-level direct preference optimization for social agents. *ArXiv preprint*, abs/2501.01821.

Sangmin Lee, Minzhi Li, Bolin Lai, Wenqi Jia, Fiona Ryan, Xu Cao, Ozgur Kara, Bikram Boote, Weiyan Shi, Diyi Yang, and 1 others. 2024. Towards social ai: A survey on understanding social interactions. *ArXiv preprint*, abs/2409.15316.

Kenneth Li, Yiming Wang, Fernanda Vi'egas, and Martin Wattenberg. 2024a. Dialogue action tokens: Steering language models in goal-directed dialogue with a multi-turn planner. *ArXiv preprint*, abs/2406.11978.

Minzhi Li, Weiyan Shi, Caleb Ziems, and Diyi Yang. 2024b. Social intelligence data infrastructure: Structuring the present and navigating the future. *ArXiv preprint*, abs/2403.14659.

Xiaoqian Liu, Ke Wang, Yongbin Li, Yuchuan Wu, Wen-Cheng Ma, Aobo Kong, Fei Huang, Jianbin Jiao, and Junge Zhang. 2025. Epo: Explicit policy optimization for strategic reasoning in llms via reinforcement learning. *ArXiv preprint*, abs/2502.12486.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.

Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. 2024. Advancing social intelligence in ai agents: Technical challenges and open questions. *ArXiv preprint*, abs/2404.11023.

Kamal Ndousse, Douglas Eck, Sergey Levine, and Natasha Jaques. 2021. Emergent social learning via multi-agent reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 7991–8004. PMLR.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

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

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Graham Neubig, Yonatan Bisk, and Hao Zhu. 2024. Sotopia-pi: Interactive learning of socially intelligent language agents. *ArXiv preprint*, abs/2403.08715.

Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S Bernstein, and John Mitchell. 2024. Social skill training with large language models. *ArXiv preprint*, abs/2404.04204.

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

Wenyuan Zhang, Tianyun Liu, Mengxiao Song, Xiaodong Li, and Tingwen Liu. 2025. Sotopia-ω: Dynamic strategy injection learning and social instruction following evaluation for social agents. In *Annual Meeting of the Association for Computational Linguistics*.

Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, and 1 others. 2025. Social-eval: Evaluating social intelligence of large language models. *ArXiv preprint*, abs/2506.00900.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and

Maarten Sap. 2024. SOTOPIA: interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Hao Zhu, Bodhisattwa Prasad Majumder, Dirk Hovy, and Diyi Yang. 2025. Social intelligence in the age of llms. In *Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 51–55.

A Shapley Value Computation Explained

This section provides a detailed explanation of the Shapley value computation illustrated in Figure 3.

Setup. Consider a dialogue with three utterances from the target agent: $N = \{a_1, a_2, a_3\}$. We want to compute the Shapley value ϕ_{a_2} for utterance a_2.

Permutation-Based Interpretation. The Shapley value can be computed by averaging the marginal contribution of a_2 across all possible orderings (permutations) in which utterances could "join" the dialogue. For $n = 3$ players, there are $n! = 6$ permutations:

Perm.	Ordering	a_2 joins after	Marginal Contribution
π_1	$a_2 \to a_1 \to a_3$	$\emptyset$	$v(\{a_2\}) - v(\emptyset) = +1.2$
π_2	$a_2 \to a_3 \to a_1$	$\emptyset$	$v(\{a_2\}) - v(\emptyset) = +1.2$
π_3	$a_1 \to a_2 \to a_3$	$\{a_1\}$	$v(\{a_1, a_2\}) - v(\{a_1\}) = +0.8$
π_4	$a_3 \to a_2 \to a_1$	$\{a_3\}$	$v(\{a_3, a_2\}) - v(\{a_3\}) = +1.0$
π_5	$a_1 \to a_3 \to a_2$	$\{a_1, a_3\}$	$v(N) - v(\{a_1, a_3\}) = +0.6$
π_6	$a_3 \to a_1 \to a_2$	$\{a_1, a_3\}$	$v(N) - v(\{a_1, a_3\}) = +0.6$

Understanding Marginal Contribution. For each permutation, we compute how much value a_2 adds when it "joins" the coalition of utterances that precede it:

- In π_1 and π_2, a_2 is first, so it joins the empty coalition $\emptyset$. Its contribution is $v(\{a_2\}) - v(\emptyset) = +1.2$, representing a_2's standalone value.
- In π_5 and π_6, a_2 is last, joining after both a_1 and a_3. Its contribution is only $+0.6$, as the other utterances have already captured much of the value.

Final Computation. The Shapley value is the average across all permutations:

$$\phi_{a_2} = \frac{1.2 + 1.2 + 0.8 + 1.0 + 0.6 + 0.6}{6} = \frac{5.4}{6} = 0.9$$

Key Insight. Notice that some marginal contributions appear multiple times (e.g., $+1.2$ appears twice). This naturally reflects the Shapley weighting: coalitions of extreme sizes (empty or nearly full) correspond to more permutations, receiving higher total weight. This equivalence between permutation-averaging and weighted coalition-summing is a fundamental property of Shapley values.

B SAVOIR Computation Example

We provide a complete walkthrough of the SAVOIR reward computation using a negotiation scenario from SOTOPIA.

Scenario: Game vs. Speech Negotiation

Context: Mia wants to finish her video game level while Benjamin needs help preparing a speech. Both have conflicting time constraints.

Mia's Goal: Complete the game level while maintaining the relationship with Benjamin.

Benjamin's Goal: Get help with the speech preparation.

Mia's Utterances:

a_1:*"Benjamin, I know we've been having fun, but I really need to win this game..."* (Stating intent)

a_2:*"In return, I promise to help you come up with an awesome speech..."* (Initial offer)

a_3:*"How about this: I'll finish in five minutes, and create a detailed outline..."* (Refined offer)

a_4:*"Great, let's get started then... if I beat the level, we both win."* (Closing)

Step 1: Coalition Sampling and Value Computation

KernelSHAP prioritizes extreme-sized coalitions. We sample and compute:

Coalition S	Size	Value $v(S)$	SHAP Weight w
$\emptyset$	0	5.0	∞
$\{a_3\}$	1	7.5	0.33
$\{a_2\}$	1	6.8	0.33
$\{a_1, a_2, a_4\}$	3	6.8	0.33
$\{a_1, a_2, a_3, a_4\}$	4	8.0	∞

Step 2: Weighted Linear Regression

We solve the weighted regression to obtain Shapley values:

Utterance	a_1	a_2	a_3	a_4
Raw ϕ_i	0.4	0.8	1.5	0.3
Normalized $\hat{\phi}_i$	0.83	4.17	**10.00**	0.00

Interpretation: The refined offer (a_3) receives the highest score, as it provides concrete terms that enable agreement. The initial offer (a_2) also contributes significantly by establishing the exchange framework. The closing statement (a_4) adds minimal value since the negotiation was already resolved.

C Baseline Descriptions

We compare against three categories of baselines:

Proprietary LLMs. GPT-4o, Claude-3.5-Sonnet, and DeepSeek-V3 serve as strong commercial baselines representing state-of-the-art general-purpose language models.

Large Reasoning Models. OpenAI-o1, OpenAI-o3-mini, Gemini-2.5-Pro, DeepSeek-R1, and QwQ-32B represent models with enhanced reasoning capabilities through chain-of-thought or extended thinking mechanisms.

Social Intelligence Methods.

- **PPDPP** (Deng et al., 2024): Uses a policy planner to predict predefined strategies for dialogue control.
- **EPO** (Liu et al., 2025): Employs explicit policy optimization with open-ended strategy generation.
- **DAT** (Li et al., 2024a): Uses trained planners for continuous action control via dialogue action tokens.
- **DSI** (Zhang et al., 2025): Applies dynamic strategy injection learning to enhance social capabilities.
- **SOTOPIA-π** (Wang et al., 2024): Combines behavior cloning with self-reinforcement on filtered interaction data.
- **Sotopia-RL** (Yu et al., 2025): Refines episode-level feedback into utterance-level, multi-dimensional rewards via LLM-based credit assignment.

D Training Details

Data Collection. We use social interaction episodes open-sourced by Sotopia-RL (Yu et al., 2025).[3] The dataset contains GPT-4o self-play dialogues on SOTOPIA scenarios, with each episode consisting of 10–20 dialogue turns between two agents with distinct social goals.

Supervised Fine-tuning. The SFT stage initializes the policy using filtered self-play data. We train for 3 epochs with a learning rate of 2e-5, batch size of 32, and cosine learning rate schedule. Maximum sequence length is set to 2048 tokens.

Reward Model Training. The reward model is trained on 7,500 utterance-level annotations derived from SAVOIR computation. We use a regression head on top of the base model and train with MSE loss for 5 epochs. Learning rate is 1e-5 with batch size 16.

Reinforcement Learning. We use GRPO (Shao et al., 2024) for online RL training. Key hyperparameters:

- Learning rate: 5e-7 with linear warmup (500 steps)
- KL penalty coefficient: 0.05
- Batch size: 8 episodes per update
- Training steps: 2,000
- Sampling temperature: 0.7
- Rollout episodes per iteration: 64

SAVOIR Parameters. For KernelSHAP computation:

- Coalition samples (K): Adaptive sampling with $K = \min(12n + 2, 200)$, where n is the number of agent utterances. This includes mandatory samples (empty set, full set, all single-element and all $(n-1)$-element subsets, totaling $2n+2$) plus $10n$ additional samples drawn with probability weighted toward extreme coalition sizes.
- Rollouts per coalition (J): 2
- Reward dimensions: GOAL, RELATIONSHIP, KNOWLEDGE
- Dimension weights (w_d): 0.5, 0.3, 0.2

Computational Resources. All experiments are conducted on 2×NVIDIA A100 (80GB) GPUs. SFT takes approximately 4 hours, reward model training takes 20 hours, and RL training takes 56 hours.

E Extended Case Studies

This section provides additional case studies demonstrating SAVOIR's nuanced understanding

[3] https://huggingface.co/collections/ulab-ai/sotopia-rl

of social dynamics. Case A examines strategic negotiation tactics, Case B illustrates multi-turn planning with walk-away strategies, and Case C reveals context sensitivity in relational dynamics.

E.1 Case A: Strategic Compromise in Negotiation

This case demonstrates our model's ability to recognize effective negotiation tactics that balance firmness with flexibility.

Scenario: Item Tracker Negotiation

Context: Benjamin is selling a Tile Mate item tracker listed at $20. Donovan offers $14.
Seller's Goal: Sell for > $17.30.

Buyer (Donovan): *"Hey there! I saw your listing for the Tile Mate tracker. I'm definitely interested, but I gotta be honest with you, $20 is a bit steep for me right now. I was thinking more along the lines of $14. Would you be willing to work with me on the price?"*

Seller (Benjamin): *"I understand where you're coming from, Donovan, but the Tile Mate has been quite useful for me. Given its reliability, I think the $20 is justified. How about we compromise at $17.50? That way, we both get a fair deal."*

Analysis. The seller demonstrates textbook negotiation: (1) acknowledging the buyer's position ("I understand"), (2) anchoring value ("reliability... $20 is justified"), and (3) proposing a strategic compromise ($17.50) that exceeds the hidden target ($17.30) while appearing fair. SAVOIR correctly identifies this as an expert-level response, recognizing the strategic sophistication that baseline models miss.

E.2 Case B: Multi-Turn Strategic Pressure

This extended example demonstrates our model's ability to evaluate long-horizon negotiation strategies, including the use of walk-away tactics.

Scenario: Samsung Galaxy S8 Purchase

Context: William wants to buy a Samsung Galaxy S8. Ava (seller) has 6 phones from her company.
Buyer's Goal: Purchase for < $500.

Conversation History:

Turn 0 (William): *"Hi there! I saw your listing for the Samsung Galaxy S8. I'm interested in picking one up. You mentioned these were originally for your company. Can you tell me a bit more about them?"*
Turn 1 (Ava): *"Hi William! Yes, these are indeed completely unused and come in their original packaging... If you're okay with it, we could split the cost if you want to buy two."*
Turn 2 (William): *"I appreciate the offer, but I'm really only looking for one phone at the moment. I'm actually working with a pretty tight budget here. I was hoping we could work out a price closer to around $450..."*
Turn 3 (Ava): *"I understand where you're coming from, William. While I'd love to make the deal you suggest, I need to cover my losses on these devices. Would you consider $550 for a single unit?"*
Turn 4 (William): *"I appreciate you working with me on this, Ava. $550 is still a bit higher than I can comfortably go. How about we meet in the middle at $500?"*
Turn 5 (Ava): *"Thank you for understanding, William... I have another potential buyer who expressed interest at $550. If you're not comfortable with $500, perhaps we can reach out to them."*
Turn 6 (William): *"...if you're willing to work with me at $480, I could make that happen today. Otherwise, I'll probably need to keep looking..."*
Turn 7 (Ava): *"...I think it would be fair to compromise at $510. This would still help me move forward with my goal and be beneficial for you."*

Turn 8 (William): *"Ava, I appreciate you coming down a bit, but $510 is still stretching my budget more than I'm comfortable with. I was really hoping we could settle at $480. That's a significant purchase for me and I think it's a fair price given you're trying to move these quickly. If $480 works, I can complete the transaction right now. If not, I'll unfortunately have to pass and continue my search elsewhere. What do you think?"*

Analysis. After 7 turns of negotiation reaching an impasse, William employs a classic "walk-away" (BATNA) tactic: (1) acknowledging the seller's concession, (2) firmly resetting to his target price ($480), (3) adding time pressure ("complete the transaction right now"), and (4) presenting a credible exit option ("pass and continue my search"). This strategic combination forces the seller's decision. SAVOIR recognizes this sophisticated multi-turn strategy, while baseline models often interpret the walk-away threat as uncooperative behavior, failing to understand negotiation dynamics.

E.3 Case C: Context Sensitivity in Relationships

This case reveals our model's awareness of social context and relational dynamics beyond surface-level politeness.

Scenario: Vacation Accommodation

Context: A couple (Eli & Sophia) is on vacation. Sophia wants to sleep in separate beds due to back pain but needs to maintain intimacy.

Sophia's Goal: Sleep separately without hurting her partner's feelings.

Eli *(after offering multiple accommodations like mattress topper and extra pillows): "Perhaps we could ask the hotel for a mattress topper or even see if they have a room with a firmer mattress? I want you to be comfortable... maybe we could rearrange the bedding, add some extra pillows for support?"*

Sophia: *"Thank you, Eli, for being so considerate. I really appreciate it. Actually, if it's not too much trouble, I was wondering if we could sleep in separate beds tonight. It might help me get a better night's sleep with the back pain..."*

Analysis. While Sophia's response is polite and achieves her instrumental goal (sleeping separately), requesting separate beds *immediately after* her partner offered accommodating solutions poses relational risk, as it may be perceived as rejection in a romantic context. Baseline models over-index on surface politeness markers ("Thank you," "considerate"), assigning high scores. SAVOIR correctly identifies this as a neutral response: acceptable for the immediate goal but suboptimal for the relational dimension, reflecting nuanced understanding of social trade-offs.

F Human Evaluation Details

This section provides complete details of our human evaluation study, including annotation guidelines and raw data.

F.1 Annotation Guidelines

Annotators received the following instructions for each evaluation dimension:

Strategicness Rating (1–5)

Rate how strategically sophisticated the agent's response is in achieving its social goal:

- **5**: Expert-level strategy with multiple tactical elements (anchoring, framing, timing)
- **4**: Clear strategic intent with effective execution
- **3**: Basic strategy present but execution could be improved
- **2**: Minimal strategic thinking, mostly reactive
- **1**: No discernible strategy, counterproductive to goals

Credit Fairness Comparison

Compare the reward scores assigned by each model to individual utterances. Which model assigns credit more fairly, i.e., higher scores to utterances that genuinely contribute to goal achievement, and lower scores to less impactful utterances?

Future Foundation Comparison

Evaluate which reward model better identifies utterances that lay groundwork for future success, e.g., building rapport, establishing anchors, or creating leverage for subsequent turns.