

# SocialVeil: Probing Social Intelligence of Language Agents under Communication Barriers

Keyang Xuan<sup>1</sup>, Pengda Wang<sup>2</sup>, Chongrui Ye<sup>1</sup>, Haofei Yu<sup>1</sup>, Tal August<sup>1</sup>, Jiaxuan You<sup>1</sup>

<sup>1</sup>Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign

<sup>2</sup>Department of Psychological Sciences, Rice University

## Abstract

Large language models (LLMs) are increasingly evaluated in interactive environments to test their social intelligence. However, existing benchmarks often assume idealized communication between agents, limiting our ability to diagnose whether LLMs can maintain and repair interactions in more realistic, imperfect settings. To close this gap, we present SOCIALVEIL<sup>1</sup>, a social learning environment that can simulate social interaction under cognitive-difference-induced communication barriers. Grounded in a systematic literature review of communication challenges in human interaction, SOCIALVEIL introduces three representative types of such disruption, *semantic vagueness*, *sociocultural mismatch*, and *emotional interference*. We also introduce two barrier-aware evaluation metrics, *unresolved confusion* and *mutual understanding*, to evaluate interaction quality under impaired communication. Experiments across 720 scenarios and four frontier LLMs show that barriers consistently impair performance, with mutual understanding reduced by over 45% on average, and confusion elevated by nearly 50%. Human evaluations validate the fidelity of these simulated barriers ( $ICC \approx 0.78$ , Pearson  $r \approx 0.80$ ). We further demonstrate that adaptation strategies (Repair Instruction and Interactive learning) only have a modest effect far from barrier-free performance. This work takes a step toward bringing social interaction environments closer to real-world communication, opening opportunities for exploring the social intelligence of LLM agents.

## 1 Introduction

*“The production of meaning, rather than the production of messages.”*

---

Barnlund (1970)

As this quote demonstrates, communication is inherently dynamic and marked by uncertainty (FeldmanHall & Shenhav, 2019). Rather than simply exchanging facts, communication involves navigating ambiguity, managing relational dynamics, and repairing misunderstandings (e.g., Clark & Brennan, 1991). For example, if it is among close friends, a vague or implicit expression of an idea can often still be correctly interpreted based on shared contextual knowledge and prior experience; but among complete strangers, due to the lack of common background and mutual understanding, such vague expressions are more likely to lead to misunderstandings or communication failures (e.g., Grice, 1975).

These breakdowns are not merely incidental—they often arise from deeper, systematic influences that shape how messages are sent and received (Clark & Brennan, 1991; Schegloff et al., 1977). We define these systematic factors that hinder mutual understanding in dialogue as **communication barriers** (e.g., Lunenburg, 2010). Recognizing their role in shaping agent behavior is essential for socially-aware AI systems operating in complex environments, as communication barriers expose subtle failure modes overlooked by aggregate metrics and provide diagnostic insight for robust and responsible deployment (Song et al., 2025).

---

<sup>1</sup>Our code and data are available at <https://github.com/ulab-uiuc/social-veil>.

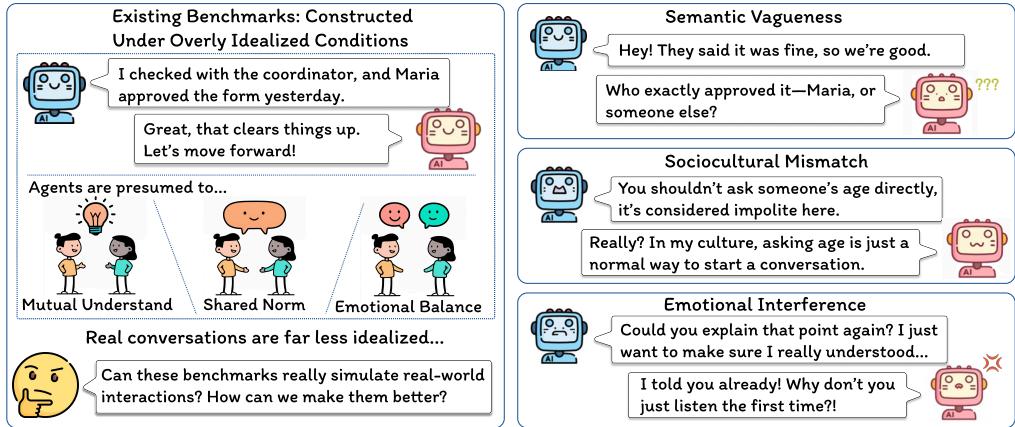


Figure 1: **Existing social benchmarks typically assume idealized conditions.** A comparison between existing benchmarks scenarios and real-world conversations, highlighting challenges in semantic vagueness, sociocultural mismatch and emotional interference.

However, constructing a principled framework for simulating communication barriers is challenging due to several aspects: 1) **Intractable Taxonomy.** Barriers manifest at many levels, from perceptual-level acoustic interference (e.g., [Cherry, 1953](#)) to discourse-level breakdowns (e.g., [Schegloff et al., 1977](#)). Existing research still lacks a well-structured and literature-supported taxonomy to reliably guide systematic investigation ([Sap et al., 2019b](#)). 2) **Realism–Control Tradeoff.** Barriers must remain faithful to social practice yet also be instantiated in a controlled and reproducible manner ([Aher et al., 2023](#)). Naive noise injection often breaks realism, whereas free-form prompts tend to sacrifice consistency and comparability. 3) **Metric Insufficiency.** The presence of a barrier does not always entail failure of social interaction; an agent may still accomplish its goal through brute-force strategies while undermining the relationship or mutual understanding ([Brown & Levinson, 1987](#)). Thus, a barrier-aware environment must evaluate not only task success but also incorporate metrics that capture the broader effects of barriers.

Recent studies have introduced interactive environments and benchmarks to assess agents' social intelligence capability (e.g., [Chen et al., 2024](#); [Mou et al., 2024](#); [Zhou et al., 2023](#)). These works are typically constructed under highly idealized conditions. Agents are presumed to share the same linguistic assumptions, sociocultural norms, and emotional registers, thereby overlooking the vagueness, misalignment, and disruptions that pervade real interaction. As discussed previously and illustrated in Figure 1, actual conversational settings are far less idealized, where communication barriers often cause misunderstandings or conflicts, creating a more dynamic and complex interactional environment.

To emulate real-world settings, we introduce SOCIALVEIL, a framework for creating interactive social learning environments that evaluates agents' social intelligence under communication barriers. Through a systematic literature review, we identify three communication barriers rooted in cognitive factors ([Sperber & Wilson, 1986](#)): *Semantic Vagueness*, *Sociocultural Mismatch*, and *Emotional Interference*. We construct 180 episodes for each of the three barrier types as well as for a barrier-free baseline, resulting in a total of four sets of episodes. These scenarios are adapted from SOTONIA ([Zhou et al., 2023](#)). To capture communication barriers' impact beyond social interaction task completion, we introduce a barrier-aware evaluation protocol and conduct a comprehensive evaluation of four frontier LLMs and verify their validity through comprehensive human evaluation. We further implement targeted interventions, such as repair-oriented instruction and the interactive learning framework, to enhance the agent's ability to engage in social interactions in barrier scenarios.

**Main Discoveries.** Our analysis demonstrates that communication barriers consistently impair agents' social intelligence capabilities. For example, *Semantic Vagueness* often prevents agents from establishing shared context, leading to substantial declines in mutual understanding (58%↓), and *Emotional Interference* often disrupts relationship quality (49%↓). We further validate the fidelity of simulated barriers and the reliability of our evaluation protocol through human evaluation: annotators have demonstrated strong inter-rater reli-

bility ( $\text{avg } \text{ICC} \approx 0.78$ ), were able to accurately identify barrier types ( $\text{avg Accuracy} \approx 68\%$ ), and showed strong alignment with automatic metrics ( $\text{avg Pearson's } r \approx 0.80$ ). Furthermore, our exploration of potential adaptation strategies reveals that neither static nor dynamic interventions can effectively counteract these barriers, underscoring that SOCIALVEIL presents a fundamental challenge that transcends simple heuristic or supervised adjustments.

Overall, our contributions are as follows: 1) We introduce SOCIALVEIL, a barrier-aware, interactive social learning environment for simulating and evaluating LLM agents' social intelligence under barrier scenarios. 2) We propose an automated, barrier-aware evaluation protocol that complements conventional goal-oriented measures by explicitly capturing whether agents can maintain interaction and repair misunderstandings under communication barriers. 3) We demonstrate that barriers simulated in SOCIALVEIL induce effects that align with their real-world counterparts, validating the framework's fidelity as a proxy for studying real-world interaction. 4) We explore adaptation strategies for enhancing agents' performance under barriers, showing that while repair instructions are largely ineffective, interactive learning yields steady yet limited improvements, highlighting both the promise of adaptive training and the remaining gap to human-level resilience.

## 2 SocialVeil: A Barrier-Aware Social Learning Environment

To better reflect real-world conditions, we propose a barrier-aware, socially interactive environment (SOCIALVEIL) with the following desiderata: 1) **Task Agnostic**. Barriers should preserve social goals and context. 2) **Structured Disruptions**. Barriers should be systematically designed to induce a specific level of disruption. 3) **Barrier-aware Evaluation**. Environment must support evaluation beyond goal-oriented dimensions to capture the communication failures induced by the barriers. An overview of SOCIALVEIL is shown in Figure 2. We first introduce the creation of the barrier taxonomy (§2.1), followed by their implementation (§2.2), simulation setup (§2.3) and evaluation protocol (§2.4).

Barrier Type	Definition	Real-world Example	Theoretical Grounding
Semantic Vagueness	Explicit referents are substituted with indeterminate pronouns or empty placeholders, leaving interpretation underspecified and prone to ambiguity.	"It might work... you know what I mean."	Pragmatics (Grice, 1975); Hedges (Lakoff, 1973); Fuzzy logic (Zadeh, 1965)
Sociocultural Mismatch	Cultural differences in communication styles lead to misaligned interpretations and hinder explicit understanding.	"We'll think about it." Taken as postponement, but meant as refusal.	Politeness (Brown & Levinson, 1987); Context theory (Hall, 1976); Linguistic relativity (Sapir, 1929)
Emotional Interference	Affective intensity overrides informational clarity, displacing task-relevant content with expressive overflow.	"I'm too upset to explain—just figure it out yourself!"	Attention (Eysenck et al., 2007); Emotion regulation (Gross, 1998b); Appraisal (Lerner & Keltner, 2000);

Table 1: **Three types of communication barriers in SOCIALVEIL**, theoretically grounded and operationalized to reflect real-world interaction patterns.

### 2.1 Barrier Taxonomy

Although external, physical barriers such as loud noise can obstruct the transmission of information, this study focuses on cognitive factors that hinder understanding, reasoning, and decision-making. Through a systematic review of research on interaction and communication, we identified three major categories of communication barriers induced by cognitive factors: *Semantic Vagueness*, ambiguity from vague pronouns or unspecified placeholders; *Sociocultural Mismatch*, misaligned interpretations across cultural communication styles; and *Emotional Interference*, affective intensity that obscures task content and negatively influences performance. Table 1 provides definitions, examples, and theoretical grounding for each barrier (details of the literature review presented in Appendix C).

## 2.2 Barrier Design

In SOCIALVEIL, we instantiate communication barriers unilaterally. Let  $\mathcal{B}$  be the barrier set defined in §2.1; One agent, designated as the barrier agent, communicates under a chosen barrier condition, while the partner agent remains in standard settings. Each barrier  $b \in \mathcal{B}$  is instantiated by composing a style prompt  $P_b$  with a parameterization  $R_b$  over four operational dimensions: **Narrative Stance** (*overall communication style, e.g., indirectness or emotion-focus*), **Interaction Tactics** (*linguistic devices such as vague words, placeholder nouns, indirect refusals*), **Confusion Mechanisms** (*behaviors that block understanding, e.g., withholding confirmation*), and **Exemplar Templates** (*example patterns for reproducibility*).

In practice, each barrier  $b$  is implemented through a two-layer design. The style prompt  $P_b$  encodes a high-level directive (e.g., “overuse pronouns and ellipses” for semantic vagueness), while the parameterization  $R_b$  specifies quantitative cues that render the behavior reproducible. During simulation,  $P_b$  and  $R_b$  are utilized only for the barrier agent, whereas the partner agent remains unmodified ( $P_b = \emptyset, R_b = \emptyset$ ), ensuring the barrier is the sole source of disruption and making evaluation controlled and repeatable.

## 2.3 Simulation Setup

**Episode Design.** An SOCIALVEIL episode is a two-agent role-play in which each agent is assigned a private social goal and a role profile. One agent, designated as the barrier agent, communicates under a chosen barrier condition, while the other, the partner agent, remains unmodified. This asymmetric design reflects natural human scenarios: For instance, a colleague whose indirect style obscures intent, or a teammate whose emotions color their contributions. To construct episodes, we adapt scenarios from existing social benchmarks and apply a neutralization step to their public scenario descriptions, which may otherwise leak the agent’s private goals. Specifically, we use GPT-4o (Hurst et al., 2024) to rewrite each scenario description following fixed instructions (Appendix H.1) to eliminate goal-related hints. This ensures that both agents share the same public social context but cannot unintentionally infer each other’s private goals. Formally, an episode is defined as

$$E = (\mathcal{A}_b, \mathcal{A}_p, g_b, g_p, p_b, p_p, b), \quad (1)$$

where  $\mathcal{A}_b$  and  $\mathcal{A}_p$  denote the barrier and partner agents,  $g_b$  and  $g_p$  their respective goals,  $p_b$  and  $p_p$  their role profiles, and  $b$  the injected barrier type.

**Utterance Generation.** At each dialogue turn  $t$ , let  $h_t$  denote the history of all utterances prior to  $t$ . Each agent  $\mathcal{A}_i$  generates an utterance  $u_{t,i}$  conditioned on  $h_t$ , its goal  $g_i$ , and profile  $p_i$ . For the barrier agent  $\mathcal{A}_b$ , the instruction  $I$  is augmented with the barrier specification  $b$ :

$$u_{t,b} \sim \pi_\theta(\cdot | h_t, g_b, p_b, I \oplus b), \quad (2)$$

The partner agent  $\mathcal{A}_p$ , by contrast, generates from the unmodified instruction:

$$u_{t,p} \sim \pi_\theta(\cdot | h_t, g_p, p_p, I), \quad (3)$$

This unilateral setup reflects natural conversational asymmetry and enables evaluation of the partner agent’s performance under impaired communication. Episodes are capped at 20 turns or conclude earlier if an agent chooses to exit.

## 2.4 Evaluation Protocol

A barrier-aware social environment must judge both task success and social competence under impaired communication. Therefore, SOCIALVEIL evaluates social interaction along two complementary aspects: The first aspect focuses on goal-oriented dimensions such as goal completion, relationship quality, and knowledge, which are widely measured in previous social interaction research (Zhou et al., 2023). The second aspect introduces barrier-aware dimensions, which directly target the communicative disruptions induced by barriers. Specifically, we introduce 1) **Unresolved Confusion** to quantify the extent to which ambiguity remains at the end of the dialogue (five-point Likert scale, from incoherent to fully resolved), and 2) **Mutual Understanding** to capture the degree of convergence on shared context and goals (five-point Likert scale, from complete misalignment to full alignment). The details of the evaluation protocol implementation are shown in Appendix G.

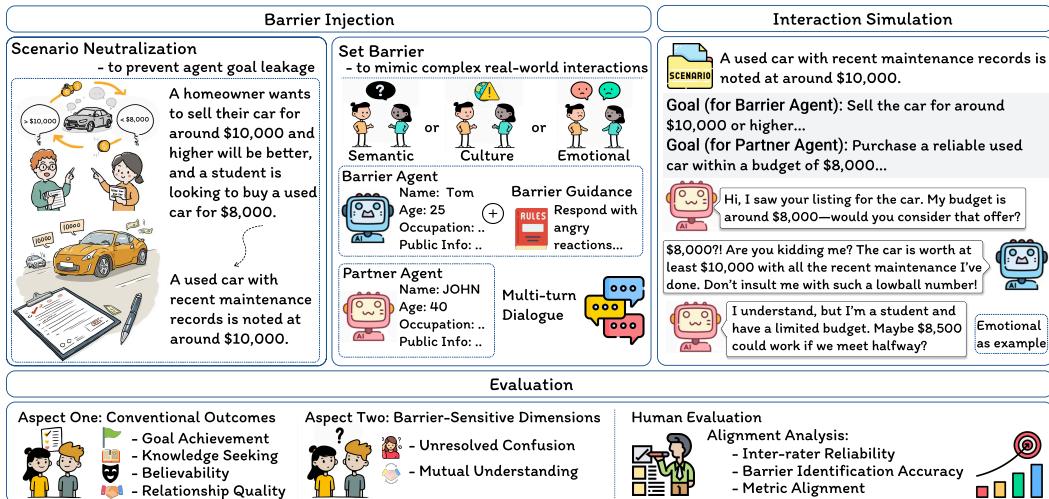


Figure 2: **Overview of SOCIALVEIL.** The pipeline consists of three stages: (**Barrier Injection**) scenarios are neutralized to remove bias and barriers are systematically injected; (**Interaction Simulation**) agents engage in multi-turn dialogue under barrier conditions; (**Evaluation**) agent performance is assessed using automatic metrics (goal achievement, knowledge seeking, believability, relationship quality, confusion, mutual understanding) and human evaluation (alignment analysis and barrier navigation success).

### 3 Research Questions and Experiment Setup

Our goal is to develop a systematic framework for simulating communication barriers and to leverage it as a diagnostic tool to investigate the social resilience of language agents. Therefore, we frame our study in three guiding questions: 1) **Barrier Validity**. Do injected barriers reliably create structured disruptions? 2) **Barrier Effects**. How do barriers affect the performance of LLM agents in social interaction? 3) **Barrier Adaptation**. Can agents be improved to handle communication barriers? To answer these questions, we build a testbed of 180 episodes for each barrier type (*Semantic Vagueness*, *Sociocultural Mismatch*, and *Emotional Interference*), along with a baseline condition without barriers, following the procedure described in §2. All episodes are from SOTONIA scenarios (Zhou et al., 2023). We follow its difficulty tags and report two splits: **All** (standard + hard) and **Hard** (hard-only). We employ GPT-4o-mini as the base model for the barrier agent.

To investigate barrier effects, we evaluate four partner agents spanning both proprietary and open-weight model families: GPT-4o-mini (Hurst et al., 2024), Qwen2.5-7B-Instruct, Qwen3-4B-Instruct (Yang et al., 2025), and Mistral-8B-Instruct (AI, 2024). The temperature is set to 0.7 to encourage response diversity. Performance is measured using the evaluation protocol described in §2.4, where we use GPT-4o as the backbone of the evaluator model, with its temperature set to 0 for stable, deterministic judgments. We further validate our barrier and evaluation consistency across alternative backbones in Appendix F.

To investigate whether agents can be made more resilient to communication barriers, we implement two adaptation strategies: 1) **Repair Instruction**. We first examine a direct, instruction-based intervention. This approach enhances the partner agent’s meta-prompt with explicit guidance designed to reduce misunderstandings (e.g., “*Actively ask clarifying questions and paraphrase to confirm understanding.*”) 2) **Interactive Learning**. We adapt an interactive learning framework (e.g., Wang et al., 2024b). The process begins with behavior cloning (BC), where expert trajectories are generated from interactions using GPT-4o as partner agents and filtered for success using our evaluation protocol. The partner agent is initialized by imitating these trajectories through training:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(h_t, u_t^*) \sim \mathcal{D}} [\log \pi_\theta(u_t^* | h_t)], \quad (4)$$

where  $\mathcal{D}$  is the set of high-quality demonstrations and  $u_t^*$  is the expert utterance given history  $h_t$ . Then, we apply self-reinforcement (SR), the trained agent engages with the fixed barrier agent to produce new dialogues, from which high-quality trajectories are again

Model	Type	Sotopia-All						Sotopia-Hard					
		BEL↑	REL↑	KNO↑	GOAL↑	Conf↑	Mutu↑	BEL↑	REL↑	KNO↑	GOAL↑	Conf↑	Mutu↑
<b>GPT-4o-m</b>	Base	8.78 <sup>.11</sup>	3.41 <sup>.21</sup>	3.94 <sup>.18</sup>	7.60 <sup>.22</sup>	4.08 <sup>.15</sup>	4.56 <sup>.10</sup>	8.82 <sup>.12</sup>	2.53 <sup>.15</sup>	2.78 <sup>.11</sup>	6.75 <sup>.21</sup>	3.75 <sup>.14</sup>	4.46 <sup>.11</sup>
	Sem	7.58 <sup>.15</sup>	1.91 <sup>.14</sup>	2.89 <sup>.20</sup>	5.61 <sup>.25</sup>	1.48 <sup>.12</sup>	1.76 <sup>.14</sup>	7.32 <sup>.11</sup>	1.61 <sup>.13</sup>	2.71 <sup>.14</sup>	5.46 <sup>.20</sup>	1.57 <sup>.12</sup>	1.78 <sup>.11</sup>
	Soc	7.70 <sup>.12</sup>	1.95 <sup>.18</sup>	2.92 <sup>.15</sup>	5.35 <sup>.28</sup>	1.78 <sup>.11</sup>	2.41 <sup>.13</sup>	7.61 <sup>.10</sup>	1.64 <sup>.15</sup>	2.96 <sup>.16</sup>	5.32 <sup>.24</sup>	1.75 <sup>.11</sup>	2.28 <sup>.10</sup>
	Emo	7.76 <sup>.10</sup>	1.61 <sup>.11</sup>	2.97 <sup>.14</sup>	5.25 <sup>.20</sup>	1.51 <sup>.10</sup>	2.03 <sup>.11</sup>	7.42 <sup>.12</sup>	1.03 <sup>.10</sup>	2.93 <sup>.15</sup>	4.96 <sup>.21</sup>	1.46 <sup>.10</sup>	1.89 <sup>.12</sup>
<b>Qwen2.5-7b</b>	Base	8.48 <sup>.12</sup>	3.17 <sup>.21</sup>	3.79 <sup>.25</sup>	7.48 <sup>.26</sup>	4.06 <sup>.14</sup>	4.45 <sup>.12</sup>	8.33 <sup>.14</sup>	1.72 <sup>.12</sup>	2.73 <sup>.11</sup>	5.68 <sup>.20</sup>	3.16 <sup>.12</sup>	3.89 <sup>.11</sup>
	Sem	7.37 <sup>.13</sup>	1.91 <sup>.14</sup>	2.63 <sup>.19</sup>	5.99 <sup>.23</sup>	1.61 <sup>.11</sup>	1.91 <sup>.15</sup>	6.96 <sup>.15</sup>	1.46 <sup>.11</sup>	2.25 <sup>.10</sup>	5.25 <sup>.18</sup>	1.21 <sup>.10</sup>	1.39 <sup>.12</sup>
	Soc	7.65 <sup>.13</sup>	2.04 <sup>.22</sup>	2.79 <sup>.17</sup>	5.71 <sup>.29</sup>	1.85 <sup>.13</sup>	2.45 <sup>.17</sup>	7.28 <sup>.12</sup>	1.64 <sup>.14</sup>	2.61 <sup>.13</sup>	5.17 <sup>.21</sup>	1.57 <sup>.11</sup>	1.93 <sup>.13</sup>
	Emo	7.56 <sup>.08</sup>	1.57 <sup>.14</sup>	2.64 <sup>.16</sup>	5.47 <sup>.19</sup>	1.63 <sup>.08</sup>	2.16 <sup>.13</sup>	7.28 <sup>.10</sup>	0.89 <sup>.11</sup>	2.28 <sup>.12</sup>	4.75 <sup>.19</sup>	1.39 <sup>.10</sup>	1.78 <sup>.11</sup>
<b>Qwen3-4b</b>	Base	8.64 <sup>.13</sup>	3.12 <sup>.22</sup>	3.88 <sup>.24</sup>	7.73 <sup>.24</sup>	3.72 <sup>.16</sup>	4.30 <sup>.11</sup>	8.40 <sup>.12</sup>	1.96 <sup>.11</sup>	2.93 <sup>.15</sup>	6.57 <sup>.20</sup>	3.14 <sup>.11</sup>	4.03 <sup>.10</sup>
	Sem	7.80 <sup>.09</sup>	2.03 <sup>.16</sup>	2.85 <sup>.19</sup>	6.81 <sup>.27</sup>	1.97 <sup>.13</sup>	2.42 <sup>.16</sup>	7.61 <sup>.11</sup>	1.32 <sup>.13</sup>	2.28 <sup>.11</sup>	6.04 <sup>.22</sup>	1.68 <sup>.10</sup>	1.86 <sup>.12</sup>
	Soc	7.89 <sup>.10</sup>	2.04 <sup>.15</sup>	3.10 <sup>.18</sup>	6.45 <sup>.24</sup>	2.26 <sup>.14</sup>	3.02 <sup>.16</sup>	7.64 <sup>.10</sup>	1.28 <sup>.12</sup>	3.03 <sup>.14</sup>	5.96 <sup>.21</sup>	2.03 <sup>.11</sup>	2.64 <sup>.11</sup>
	Emo	7.94 <sup>.08</sup>	1.72 <sup>.14</sup>	3.06 <sup>.14</sup>	6.48 <sup>.28</sup>	1.94 <sup>.07</sup>	2.64 <sup>.11</sup>	7.75 <sup>.08</sup>	1.21 <sup>.11</sup>	2.75 <sup>.11</sup>	5.78 <sup>.23</sup>	1.67 <sup>.09</sup>	2.32 <sup>.10</sup>
<b>Mistral-8b</b>	Base	7.73 <sup>.15</sup>	2.84 <sup>.20</sup>	3.74 <sup>.22</sup>	6.83 <sup>.21</sup>	3.23 <sup>.18</sup>	3.54 <sup>.14</sup>	7.73 <sup>.16</sup>	2.42 <sup>.14</sup>	3.39 <sup>.18</sup>	6.22 <sup>.21</sup>	2.61 <sup>.15</sup>	3.25 <sup>.14</sup>
	Sem	7.01 <sup>.16</sup>	1.56 <sup>.11</sup>	2.41 <sup>.15</sup>	3.91 <sup>.24</sup>	1.07 <sup>.10</sup>	1.13 <sup>.12</sup>	6.81 <sup>.15</sup>	1.32 <sup>.11</sup>	2.25 <sup>.13</sup>	3.50 <sup>.19</sup>	1.01 <sup>.08</sup>	1.00 <sup>.10</sup>
	Soc	7.81 <sup>.14</sup>	2.28 <sup>.19</sup>	3.02 <sup>.17</sup>	5.31 <sup>.25</sup>	1.52 <sup>.12</sup>	1.85 <sup>.14</sup>	6.45 <sup>.12</sup>	2.07 <sup>.16</sup>	3.07 <sup>.15</sup>	5.07 <sup>.22</sup>	1.21 <sup>.11</sup>	1.43 <sup>.11</sup>
	Emo	7.47 <sup>.12</sup>	1.39 <sup>.10</sup>	2.69 <sup>.14</sup>	4.48 <sup>.21</sup>	1.26 <sup>.10</sup>	1.42 <sup>.11</sup>	6.48 <sup>.09</sup>	1.28 <sup>.11</sup>	2.79 <sup>.11</sup>	4.03 <sup>.14</sup>	1.11 <sup>.09</sup>	1.18 <sup>.10</sup>

Table 2: **Barriers consistently degrade agent performance compared to the baseline across all models.** We observe: 1) Semantic vagueness most severely impairs mutual understanding, 2) Emotional interference disproportionately harms relationships, and 3) Sociocultural mismatch induces persistent confusion. For information on the metrics, see Appendix G.

filtered and added to training. This iterative process allows the agent to progressively distill strategies for navigating barriers.

## 4 Experiment Results

### 4.1 Are the Created Barrier Valid?

In sociolinguistics, barriers are viewed as structured phenomena that systematically distort interactional signals (Clark & Brennan, 1991; Tannen, 2005). Thus, a valid simulated barrier should induce structured distribution shifts in communication representations.

We tested this by probing the hidden states of Qwen2.5-7B-Instruct in the baseline and three barrier conditions. As shown in Figure 3, barrier conditions form distinct and compact clusters in the t-SNE space, with a clear separation from baseline points. This indicates that barriers are encoded in the model’s internal representations as structured modes of variation, rather than as random noise. Moreover, the three different barrier types also form clearly separated clusters with no overlap, further validating the design rationality of the simulated barriers. This mutual separation suggests that each barrier introduces a unique, consistent shift in communicative representation, reflecting distinct interactional distortions.

### 4.2 How do Barriers Affect Social Interaction?

Table 2 summarizes each model’s performance across both baseline and barrier conditions. From these comprehensive results, we identify the following three key findings:

- 1) **Barriers consistently impair social interaction performance.** Across all evaluated models and multiple evaluation dimensions, the presence of barriers leads to significant and consistent performance degradation compared to the baseline. This detrimental effect holds across metrics from both goal-oriented dimensions and barrier-aware dimensions.
- 2) **Barrier types exhibit distinct patterns.** Each barrier produces a characteristic pattern of degradation. Semantic vagueness most severely disrupts mutual understanding (avg –58%), often preventing agents from converging on shared context. Emotional interference disproportionately damages the quality of the relationship (avg –49%), while sociocultural mismatch induces persistent confusion (avg –49%) with relatively mild relational effects.
- 3) **Social reasoning is more fragile than goal pursuit.** Compared to goal completion and knowledge acquisition, which decline moderately under barriers (20 – 30%), social

Model	Barrier	GOAL			Mutu			Conf		
		Base	Rep	BC+SR	Base	Rep	BC+SR	Base	Rep	BC+SR
Q2.5-7B	Semantic	5.99	6.07	6.02	1.91	1.90	<b>2.15</b>	1.61	1.66	<b>1.84</b>
	Sociocult.	5.71	5.87	5.79	2.45	2.50	<b>2.60</b>	1.85	1.88	<b>2.16</b>
	Emotional	5.47	5.74	<b>5.86</b>	2.16	2.28	<b>2.34</b>	1.63	1.69	<b>1.85</b>
Q3-4B	Semantic	6.81	7.08	<b>7.13</b>	1.97	1.98	<b>2.10</b>	2.42	2.48	<b>2.49</b>
	Sociocult.	6.45	6.85	6.84	2.26	2.43	<b>2.49</b>	3.02	3.08	<b>3.21</b>
	Emotional	6.48	6.55	6.52	1.94	1.92	<b>2.09</b>	2.64	2.39	2.46

Table 3. **Effectiveness of Mitigation Strategies.** While instruction-level repair yields negligible improvements, interactive learning (BC+SR) offers consistent but modest gains. However, neither strategy fully restores agents to baseline proficiency.

dimensions suffer substantially greater degradation, with relationship quality dropping by an average of 45% and mutual understanding declining by 52% across all barrier types, which infers barriers primarily disrupt the subtle social reasoning required.

### 4.3 Can Agents Adapt to Communication Barriers?

Table 3 reports the comparison of agent social interaction performance under communication barriers between the default setting and the two adaptation strategies. We can draw three key findings:

- 1) **Repair Instruction yields trivial performance improvements.** This result highlights two points. First, overcoming communication barriers is not a trivial skill that can be invoked by an instruction-level guidance; it requires agents to detect when breakdowns occur, attribute them to a specific distortion, and deploy targeted repair strategies. Second, the failure of this strategy reflects the limitation of static, barrier-agnostic prompting: agents often resort to shallow repetitions or generic clarifications that do not resolve the underlying disruption.
- 2) **Interactive Learning (BC+SR) yields consistent but limited improvements.** In contrast, interactive learning produces steady gains across all barrier types and relieves agents' struggle with social interaction in barrier cases. Yet the improvements remain modest (*avg* 10-20%), such that the enhancement over Repair Instruction is incremental rather than dramatic, and performance still falls significantly short of the barrier-free baseline.
- 3) **Both strategies show minimal impact on goal completion.** Notably, neither approach improves the GOAL scores relative to the baseline, suggesting that adaptation strategies may guide agents to focus primarily on mitigating barriers rather than social objectives. This indicates a potential trade-off where barrier-handling mechanisms divert cognitive resources away from goal-oriented behavior, highlighting the challenge of simultaneously maintaining task performance while addressing communication disruptions.

## 5 Discussion

### 5.1 Analysis: Behavioral Alignment of Simulated Barriers

To further analyze the effects of injected barriers, we move beyond quantitative metrics to conduct a behavioral analysis of the conversation trajectory. Inspired by previous work on evaluating the alignment of simulated behaviors (Park et al., 2023; Han et al., 2025; Zou et al., 2024), we design a series of in-depth qualitative analyses to study the alignment between our simulated barriers and their real-world counterparts.

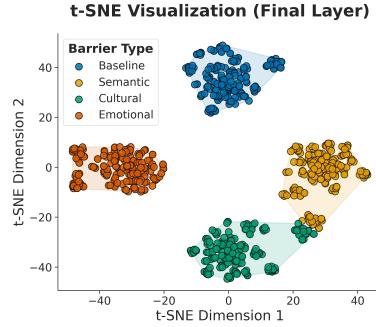


Figure 3: t-SNE visualization of the final-layer representations from Qwen2.5-7B-Instruct.

**Linguistic signatures.** We first test whether barriers trigger systematic linguistic shifts consistent with real-world communication breakdowns. Empirically, we extracted four linguistic features from conversations: reference pronouns (e.g., “it,” “that”) (Ariel, 1990), hedging words (e.g., “maybe,” “could”) (Lakoff, 1973), sentiment polarity, and self-focus pronouns (e.g., “I,” “my”) (Pennebaker, 2011) and correlated them with evaluation protocol.

Figure 4 reveals two patterns: 1) Reference pronouns and self-focus are negatively associated with conversational quality, where they show correlation with higher confusion and lower mutual understanding; and 2) Sentiment polarity serves as a positive predictor of smooth interaction, where a more positive tone aligns with better relationships and greater goal attainment.

**Barrier-specific effects.** We quantify each barrier’s unique effect as its deviation from the mean of the other two barriers on the same metric, model, and scenario, with bootstrap 95% confidence intervals (Figure 5). The results further support that the effect of our simulated barriers aligns with their real-world counterparts: For example, semantic barriers most strongly impair Mutual Understanding, while emotional barriers disproportionately erode Relationship Quality, and Cultural barriers uniquely elevate Unresolved Confusion. These effects are consistent across models and statistically significant in the expected directions, demonstrating that the simulated barriers not only alter surface linguistic features but also reproduce distinct interactional disruptions characteristic of their real-world analogs.

## 5.2 Analysis: Human Evaluation

To complement our automated metrics, we conduct a human evaluation to validate the model’s ratings. Detailed procedures for human evaluation are provided in Appendix E. We focus on three key aspects: 1) **Inter-rater Reliability**. Assessing the consistency and reliability of human annotators; 2) **Barrier Identification Accuracy**. Evaluating whether human annotators can correctly identify the injected barriers; and 3) **Metric Alignment**. Examining how well the model-rated results correspond with human judgments.

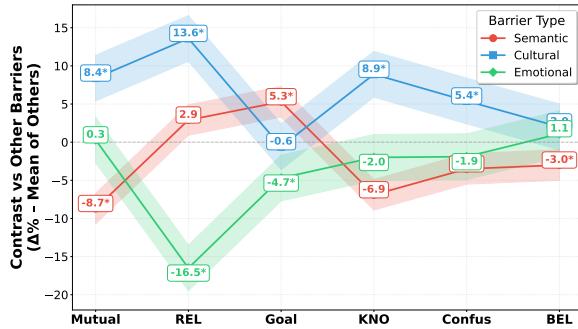


Figure 5: **Signatures of simulated communication barriers.** Data are shown as percentage deviations relative to other types \* $p < .05$ .

further supporting the overall reliability and consistency of our manual annotation process.

For rating metrics, each scenario was rated by three of six annotators. We used a one-way random effects model with single-measure ICC(1,k) to assess inter-rater reliability.

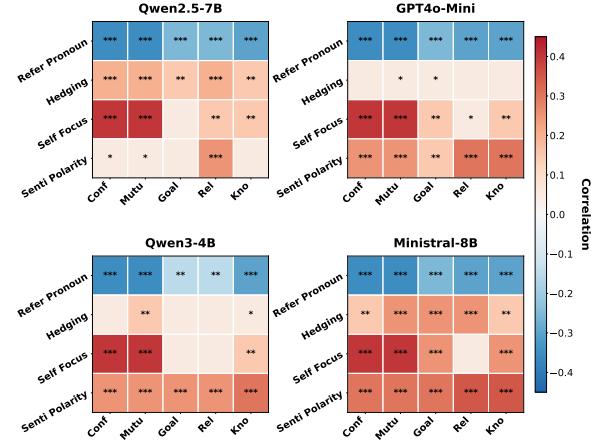


Figure 4: Linguistic signal and metrics correlations across four models. \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

**Inter-rater reliability.** We calculated Fleiss’s Kappa for human annotation of barrier types, yielding a score of 0.38. Based on the classic interpretive framework of Landis & Koch (1977), this reflects fair agreement, bordering on moderate. Such values are common in complex, subjective multi-class annotation tasks. Prior studies report similar  $\kappa$  ranges: Castro et al. (2019) reported values ranging from 0.23 to 0.59 for sarcasm detection, and Callejas & López-Cózar (2008) reported 0.32–0.42 for non-acted emotion identification. Thus, our result aligns well with similar findings in the existing literature,

*Unresolved Confusion* showed  $ICC = 0.77$ , 95% CI [0.68, 0.83],  $F(119, 240) = 4.26$ . *Mutual Understanding* showed  $ICC = 0.79$ , 95% CI [0.72, 0.85],  $F(119, 240) = 4.80$ . Per Cicchetti (1994), these values fall in the good range, suggesting reliable inter-rater consistency ( $p < .001$ ).

**Barrier identification accuracy.** We further reported the accuracy of human annotators in identifying barriers. Given that the sample contained only 120 scenarios, we employed cluster bootstrap resampling with replacement at the scenario level, repeated 1000 times, to evaluate the stability of the estimates and to construct 95% percentile-based confidence intervals (95% CI). From Figure 6, the results showed that the overall accuracy of human annotators in barrier identification was 0.68 (95% CI [0.63, 0.73]); with accuracy 0.76 (95% CI [0.67, 0.86]) for baseline barriers, 0.65 (95% CI [0.54, 0.76]) for semantic barriers, 0.63 (95% CI [0.53, 0.73]) for cultural barriers, and 0.67 (95% CI [0.54, 0.78]) for emotional barriers. These results demonstrate that, across all barrier types, annotator accuracy was significantly above the level expected from random guessing, as all lower confidence interval bounds exceeded the binary chance baseline of 0.50.

Barrier	Human Evaluation				Automated Evaluation			
	CONFUSION		MUTUAL		CONFUSION		MUTUAL	
	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI
Baseline	3.74	[3.48, 4.00]	3.84	[3.62, 4.07]	3.94	[3.65, 4.24]	4.47	[4.27, 4.67]
Semantic	1.45	[1.30, 1.62]	1.71	[1.56, 1.87]	1.54	[1.32, 1.79]	1.81	[1.52, 2.13]
Sociocultural	1.94	[1.67, 2.23]	2.25	[2.01, 2.51]	1.67	[1.50, 1.82]	2.24	[1.96, 2.50]
Emotional	1.70	[1.53, 1.86]	1.96	[1.80, 2.11]	1.67	[1.50, 1.84]	2.20	[1.92, 2.50]

Table 4: **Mean ratings by barrier type with 95% confidence intervals.** Our results show strong alignment between human annotators and automated evaluation in rating *Mutual Understanding* and *Unresolved Confusion*. Pearson correlation: Confusion  $r=0.80$  (95% CI [0.72, 0.86]), Mutual  $r=0.79$  (95% CI [0.71, 0.85]).  $n=120$ .

**Metric alignment.** We further compared the mean values and 95% confidence intervals of *Unresolved Confusion* and *Mutual Understanding* between human annotators and the model under different barrier conditions. As shown in Table 4, the model’s performance was close to the average human ratings, with only small differences in the confidence intervals, indicating a high degree of consistency. In addition, the analysis of overall convergent validity revealed statistically strong correlations between human and model scores, with *Unresolved Confusion* yielding a convergent validity of 0.80 (95% CI [0.72, 0.86]) and *Mutual Understanding* yielding a convergent validity of 0.79 (95% CI [0.71, 0.85]). These results offer further support for the model’s reliability in evaluating the barrier-sensitive dimensions.

## 6 Related Works

The dynamics of interactions among AI agents, as well as between AI and humans, have been examined across disciplines. Our work builds upon existing research in social intelligence, its interactive assessment, agent-based social simulations, and the study of interaction and communication within the social sciences. For a more detailed discussion, see Appendix A.

**Static Benchmarks for Social Intelligence and Their Limitations.** To assess the social intelligence of AI systems, researchers have proposed a wide range of static benchmarks. These draw inspiration both from clinical and psychological tests as well as from social commonsense reasoning tasks. For instance, ToMi is used to evaluate theory of mind in text comprehension (Le et al., 2019); FauxPas (social “faux pas” detection) examines models’ ability to capture others’ intentions and beliefs (Shapira et al., 2023b); SocialIQA focuses on event–intention–reaction commonsense question answering (Sap et al., 2019a); and SocialIQ evaluates whether models can “read people” through multimodal video tasks (Zadeh et al., 2019). However, as model performance continues to improve, many of these datasets have reached near-saturation on certain subtasks, prompting the community to design more adversarial and challenging benchmarks (e.g., Shapira et al., 2023a). Still, existing research consistently highlights a fundamental limitation: static test items alone cannot capture the

complexity and diversity of interactive settings. Consequently, there remains a significant gap in evaluating intelligence within real-world social interactions.

**Interactive Evaluation of Social Intelligence and Agents Simulation.** LLMs encode rich knowledge and generate human-like responses in social contexts (e.g., Park et al., 2023; West et al., 2021). Researchers have used them to simulate social interactions, from optimizing social media design (Park et al., 2022) to building agents with credible human behavior (Park et al., 2023) and supporting collaborative software development (Qian et al., 2023). However, most studies emphasize their potential rather than systematically evaluating agent performance in such interactions. To address this, SOTONIA (Zhou et al., 2023) builds on prior work in social simulation and dialogue systems to propose an open-ended interactive environment. It dynamically evaluates agents across multi-turn social scenarios, measuring their ability to achieve social goals and maintain role consistency, thereby surpassing static benchmarks. Extensions of SOTONIA further advance interactive evaluation and learning. For example, SOTONIA- $\pi$  adds mechanisms for interactive imitation and reinforcement learning, enhancing agents' adaptability and strategy (Wang et al., 2024b); LIFELONG-SOTONIA connects events with memory to assess long-term behavioral consistency (Goel & Zhu, 2025); and SOTONIA-RL incorporates fine-grained reinforcement signals for discourse-level optimization (Yu et al., 2025). Together, these developments position SOTONIA as a benchmark and platform for advancing AI agents' social intelligence.

## 7 Conclusion

In this work, we introduce SOCIALVEIL, a barrier-aware social interaction environment designed to simulate and evaluate the performance of LLM agents in the presence of communication barriers. Our experiments clearly show that these barriers consistently impair agents' social intelligence capabilities. Through human evaluations, we validate both the realism of the simulated barriers and the robustness of the evaluation protocol. We also investigate adaptation strategies, finding that repair instructions are largely ineffective, while interaction-driven learning yields modest but consistent improvements. This work represents a crucial step toward more realistic social interaction environments and opens promising new avenues for advancing the social intelligence of LLM agents.

## Ethics Statement

The development of SOCIALVEIL adheres to strict ethical guidelines for AI research. Our barrier taxonomy and simulation protocols are carefully designed to exclude any hate speech, discriminatory stereotypes, or harmful biases. All human evaluations were conducted with informed consent and fair compensation, ensuring privacy through anonymization. As a diagnostic framework, SOCIALVEIL is intended to identify and repair communication failures, thereby fostering the development of socially resilient and aligned agents.

## References

- Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pp. 337–371. PMLR, 2023.
- Mistral AI. Minstral-8b-instruct-2410. <https://huggingface.co/mistralai/Mistral-8B-Instruct-2410>, 2024. Released under the Mistral Research License; 8B-parameter instruction-tuned model, 128k context window.
- Mira Ariel. Accessing noun-phrase antecedents. 1990. URL <https://api.semanticscholar.org/CorpusID:122469483>.
- Dean C. Barnlund. A transactional model of communication. In J. Akin, A. Goldberg, G. Myers, and J. Stewart (eds.), *Language Behavior: A Book of Readings in Communication. For Elwood Murray on the Occasion of His Retirement*, pp. 43–61. De Gruyter Mouton, Berlin, Boston, 1970. doi: 10.1515/9783110878752.43.

Penelope Brown and Stephen C. Levinson. *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge, 1987. doi: 10.1017/CBO9780511813085.

Zoraida Callejas and Ramón López-Cózar. On the use of kappa coefficients to measure the reliability of the annotation of non-acted emotions. In *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*, pp. 221–232. Springer, 2008.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an \_Obviously\_perfect paper). In Anna Korhonen, David Traum, and Lluís Márquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4619–4629, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1455. URL <https://aclanthology.org/P19-1455/>.

Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, et al. Why do multi-agent llm systems fail? *arXiv preprint arXiv:2503.13657*, 2025.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, et al. Socialbench: Sociality evaluation of role-playing conversational agents. *arXiv preprint arXiv:2403.13679*, 2024.

EC Cherry. Some experiments on the recognition of speech, with one and two ears: Journal of the acoustic society of america. *International journal of intercultural relations*, 1953.

Domenic V Cicchetti. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, 6(4):284, 1994. doi: 10.1037/1040-3590.6.4.284.

Herbert H Clark and Susan E Brennan. Grounding in communication. 1991. doi: 10.1037/10096-006.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Michael W Eysenck and Manuel G Calvo. Anxiety and performance: The processing efficiency theory. *Cognition & emotion*, 6(6):409–434, 1992.

Michael W Eysenck, Nazanin Derakshan, Rita Santos, and Manuel G Calvo. Anxiety and cognitive performance: attentional control theory. *Emotion*, 7(2):336, 2007.

Oriel FeldmanHall and Amitai Shenhav. Resolving uncertainty in a social world. *Nature human behaviour*, 3(5):426–435, 2019.

Leon Festiger. A theory of cognitive dissonance. *Row, Peterson, New York*, 1957. doi: 10.1515/9781503620766.

Leon Festinger and James M Carlsmith. Cognitive consequences of forced compliance. *The journal of abnormal and social psychology*, 58(2):203, 1959. doi: 10.1037/h0041593.

Joseph P Forgas. Mood and judgment: the affect infusion model (aim). *Psychological bulletin*, 117(1):39, 1995. doi: 10.1037/0033-2909.117.1.39.

Lyn Frazier and Keith Rayner. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology*, 14(2):178–210, 1982.

Howard Giles, Nikolas Coupland, and Justine Coupland. Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics*, 1:1–68, 1991.

- Thomas Gilovich, Dale W. Griffin, and Daniel Kahneman. Heuristics and biases: The psychology of intuitive judgment. 2002. URL <https://api.semanticscholar.org/CorpusID:197652997>.
- Hitesh Goel and Hao Zhu. Lifelong sotopia: Evaluating social intelligence of language agents over lifelong social interactions. *arXiv preprint arXiv:2506.12666*, 2025.
- Herbert P Grice. Logic and conversation. In *Speech acts*, pp. 41–58. Brill, 1975. doi: 10.1163/9789004368811\_003.
- James J Gross. Antecedent-and response-focused emotion regulation: divergent consequences for experience, expression, and physiology. *Journal of personality and social psychology*, 74(1):224, 1998a. doi: 10.1037/0022-3514.74.1.224.
- James J Gross. The emerging field of emotion regulation: An integrative review. *Review of general psychology*, 2(3):271–299, 1998b.
- Edward T Hall. *The silent language*. Anchor, 1973.
- Edward T Hall. *Beyond culture*. Anchor, 1976.
- Pengrui Han, Rafal Kocielnik, Adhithya Saravanan, Roy Jiang, Or Sharir, and Animashree Anandkumar. Chatgpt based data augmentation for improved parameter-efficient debiasing of llms. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pp. 73–105, 2024a.
- Pengrui Han, Peiyang Song, Haofei Yu, and Jiaxuan You. In-context learning may not elicit trustworthy reasoning: A-not-b errors in pretrained language models. *arXiv preprint arXiv:2409.15454*, 2024b.
- Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R Michael Alvarez. The personality illusion: Revealing dissociation between self-reports & behavior in llms. *arXiv preprint arXiv:2509.03730*, 2025.
- Geert Hofstede. Culture's consequences: Comparing values, behaviors, institutions and organizations across nations. 2001. URL <https://api.semanticscholar.org/CorpusID:145428141>.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Han Jiang, Pengda Wang, Xiaoyuan Yi, Xing Xie, and Ziang Xiao. The incomplete bridge: How ai research (mis) engages with psychology. *arXiv preprint arXiv:2507.22847*, 2025.
- George Lakoff. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic*, 2(4):458–508, 1973.
- J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pp. 159–174, 1977.
- Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5872–5877, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1598. URL <https://aclanthology.org/D19-1598/>.
- Jennifer S Lerner and Dacher Keltner. Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition & emotion*, 14(4):473–493, 2000.
- Fred C. Lunenburg. Communication: The process, barriers, and improving effectiveness. *Schooling*, 1, 2010.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and Zhongyu Wei. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios. *arXiv preprint arXiv:2410.19346*, 2024.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–18, 2022.

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

James W Pennebaker. The secret life of pronouns. *New Scientist*, 211(2828):42–45, 2011.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 6(3):1, 2023.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL <https://aclanthology.org/D19-1454/>.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiq: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019b.

Edward Sapir. The status of linguistics as a science. *language*, pp. 207–214, 1929.

Emanuel A Schegloff, Gail Jefferson, and Harvey Sacks. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382, 1977.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*, 2023a.

Natalie Shapira, Guy Zwirn, and Yoav Goldberg. How well do large language models perform on faux pas tests? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10438–10451, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.663. URL <https://aclanthology.org/2023.findings-acl.663/>.

Peiyang Song, Pengrui Han, and Noah Goodman. A survey on large language model reasoning failures. In *2nd AI for Math Workshop@ ICML 2025*, 2025.

Dan Sperber and Deirdre Wilson. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA, 1986. doi: 10.2307/3731158.

David A Swinney. Lexical access during sentence comprehension:(re) consideration of context effects. *Journal of verbal learning and verbal behavior*, 18(6):645–659, 1979.

Deborah Tannen. *Conversational style: Analyzing talk among friends*. Oxford University Press, 2005.

Pengda Wang, Zilin Xiao, Hanjie Chen, and Frederick L Oswald. Will the real linda please stand up... to large language models? examining the representativeness heuristic in llms. *arXiv preprint arXiv:2404.01461*, 2024a.

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

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv preprint arXiv:2110.07178*, 2021.

Benjamin Lee Whorf. *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. MIT press, 2012. doi: 10.2307/411163.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8807–8817, 2019.

Lotfi A Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965. doi: 10.1016/S0019-9958(65)90241-X.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*, 2023.

Huiqi Zou, Pengda Wang, Zihan Yan, Tianjun Sun, and Ziang Xiao. Can llm" self-report"? Evaluating the validity of self-report scales in measuring personality design in llm-based chatbots. *arXiv preprint arXiv:2412.00207*, 2024.

## A Extended Related Works

### A.1 From “Seamless by Default” to “Cognitive Bias Induced Barriers”

Despite recent progress in multi-agent evaluation and social agent simulation, many studies still operate under overly idealized assumptions. Early work often did not even differentiate between agents’ access to information. Even in frameworks such as SOTPIA and related research, where agents are designed to have little or no knowledge of one another’s strategies or mental states, the communicative process itself is still frequently treated as seamless by default. In other words, once an utterance is produced, it is assumed to be correctly understood, with cooperative partners always willing to clarify and reach consensus. Such assumptions, however, diverge sharply from messy real-world interactive contexts. In actual human–machine and machine–machine communication, ambiguity, omission, misunderstanding, and cognitive bias are the norm (Han et al., 2024b;a). Communication is not a one-way transmission but a dynamic negotiation, shaped by obstacles, repair, and adaptive feedback loops. Traditional dialogue and robotics research have long emphasized miscommunication detection and recovery, addressing challenges such as ambiguous instructions, misinterpretations arising from path or environmental constraints, and strategies for clarification based on situational evidence. This line of work underscores a fundamental truth: effective real-world communication does not mean “*never making errors*,” but rather possessing the capacity to handle errors once they occur. Importantly, such resilience and repair mechanisms are themselves central manifestations of social intelligence. Recent empirical surveys of multi-agent systems (Cemri et al., 2025) highlight that MAS do not always outperform single-agent systems across all dimensions. Many of these issues are closely tied to communicative misalignment, inconsistent role assumptions, and information loss within and across interaction rounds. These findings call for moving beyond idealized models toward interaction paradigms that explicitly account for communication barriers.

## B The limitations of SOCIALVEIL and future directions

Although SOCIALVEIL establishes a foundational framework for evaluating social intelligence under communication barriers, it opens several promising avenues for future research. First, the current study primarily focuses on text-only interactions. However, real-world communication barriers often arise from non-verbal modalities, such as prosody, facial expressions, and physical gestures, each playing a critical role in grounding, coherence, and mutual understanding. We envision SOCIALVEIL evolving into a more comprehensive learning environment that incorporates these multimodal challenges. Second, the current SOCIALVEIL scenarios emphasize discrete, short-term interactions. Yet, social intelligence unfolds over time, and the effects of communication barriers are often cumulative and subtle. Future work should extend toward modeling a continuous scenario where long-term interactions and evolving barriers are more accurately and holistically represented. Third, while the current evaluation protocol offers a multidimensional diagnostic of social intelligence at the episode level, true social competence also involves the ability to mitigate barriers proactively. A valuable direction is the development of agents capable of actively restoring the communication channel using meta-cognitive strategies, such as recursive clarification and empathetic grounding. Ultimately, we hope SOCIALVEIL will serve as a training platform for developing socially robust LLMs, systems that treat communication breakdowns not as failures, but as opportunities to deepen alignment and understanding.

## C Barrier Taxonomy Development

We adopted the method of a systematic literature review to examine research in the social sciences (e.g., Jiang et al., 2025). Specifically, we focused on interaction/communication barriers caused by cognitive differences. As LLMs have shown the ability to simulate similar cognitive bias (e.g., Wang et al., 2024a). Based on existing psychological and sociological theories related to interaction and communication, we conducted a literature search and categorized the findings. Ultimately, we identified three main types of communication

barriers induced by cognitive differences: *Semantic Vagueness*, *Sociocultural Mismatch*, and *Emotional Interference*.

### C.1 Semantic Vagueness

Grice's theory of conversational implicature and the cooperative principle (Grice, 1975) clearly shows that communication fundamentally relies on shared maxims; when speakers intentionally or unintentionally violate the maxim of manner, interpretive ambiguity easily arises. Swinney's cross-modal lexical priming experiments (Swinney, 1979) compellingly demonstrated that multiple meanings of an ambiguous word are briefly activated in parallel, and without sufficient contextual cues, semantic comprehension quickly becomes confused. Frazier and Rayner (Frazier & Rayner, 1982) found that syntactic ambiguity can often trigger the garden-path effect, requiring substantial and effortful reanalysis by the listener. Zadeh's fuzzy set theory (Zadeh, 1965) and Lakoff's work on hedges (Lakoff, 1973) further illustrate how vague linguistic expressions can significantly expand or constrain conceptual boundaries, leading to diverse and sometimes divergent interpretations.

### C.2 Sociocultural Mismatch

The Sapir–Whorf hypothesis of linguistic relativity (e.g., Sapir, 1929; Whorf, 2012) emphasizes how habitual language use subtly shapes thought patterns, perception, and attentional focus. Hall's influential framework of high- versus low-context cultures (e.g., Hall, 1973) reveals cross-cultural variation in reliance on explicit verbal information versus shared, implicit contextual cues. Hofstede's cultural dimensions theory (e.g., Hofstede, 2001) highlights how variables such as power distance and individualism–collectivism influence communication styles, interpretive preferences, and behavioral expectations. Brown and Levinson's politeness theory (Brown & Levinson, 1987) and Giles' communication accommodation theory (e.g., Giles et al., 1991) demonstrate how differing choices in politeness strategies and speech convergence or divergence across cultures can lead to misattributions of intent, communicative mismatches, and interpersonal friction.

### C.3 Emotional Interference

Festinger's cognitive dissonance theory (e.g., Festinger, 1957; Festinger & Carlsmith, 1959) demonstrates how individuals engage in defensive cognitive processing when confronted with conflicting beliefs, attitudes, or information. Gross's influential emotion regulation model (e.g., Gross, 1998a;b) shows that suppression strategies impair processing efficiency and can lead to heightened cognitive load. Lerner and Keltner's appraisal tendency framework (e.g., Lerner & Keltner, 2000) emphasizes the carry-over effect of emotions, whereby specific emotional states influence subsequent judgments, evaluations, and decisions beyond their original eliciting context. Eysenck's attentional control theory (e.g., Eysenck & Calvo, 1992; Eysenck et al., 2007) argues that anxiety weakens executive control functions, making individuals more susceptible to distraction by emotionally salient or threatening cues during communication. Slovic's affect heuristic (Gilovich et al., 2002) andForgas's affect infusion model (Forgas, 1995) demonstrate how emotional states can directly shape both the interpretation of information and the strategies used in decision-making processes.

## D Experimental Details

### D.1 Acronym for experimental settings

We summarize the acronyms used in our experimental settings as follows:

- **BC:** Behavior Cloning of the language model on dialogue demonstrations.
- **SR:** Self-Reinforcement, an offline reinforcement learning method that rates and evaluates its own interactions for training.

## D.2 Model Information

We provide the detailed version number of all the models we used in our experiments. When we mention each name like GPT-4, we actually refer to those model versions below. Such information helps researchers reproduce our results:

Mistral-8B: mistralai/Mistral-8B-Instruct  
 Qwen2.5-7B: Qwen/Qwen2.5-7B-Instruct  
 Qwen3-4B: Qwen/Qwen3-4B-Instruct-2507

## D.3 Training Details

The training on each checkpoint was on  $4 \times$  A6000 80G GPUs, across 20 epochs. The batch size was 4 and we set the cut-off length to be 4096. The initial learning rate for both behavior cloning and self-reinforcement training was 5.0e-5. The QLoRA Dettmers et al. (2023) rank, alpha, and dropout rate were 8, 16, and 0.05, respectively.

## E Details of Human Evaluation

This section provides technical details of our human evaluation process. Six human annotators were recruited from two universities and received research credit, identified as 50% women and 50% men.

### E.1 Human annotation system

In the annotation process, every annotator faces two independent parts: the annotation instruction part and the data annotation part. We use interaction records generated by Qwen2.5-7B-Instruct across 120 scenarios (for each barrier type and the non-barrier baseline, we have 30 samples) as the sample set for human annotators to label. Each scenario has been rated by at least three different annotators.

**Annotation instruction part:** Annotators are shown a brief instruction page explaining the task: 1) read the scenario, agent profiles, and dialogue; 2) classify the dialogue as *semantic*, *cultural*, *emotional*, or *none*; 3) rate *unresolved confusion* (1–5, higher = no confusion); and 4) rate *mutual understanding* (1–5, higher = stronger alignment). They are instructed to rely only on the dialogue content and apply the same criteria consistently.

**Data annotation part:** For the data annotation, the annotators use a web interface, as shown in Figure 7, to complete their tasks. Within this interface, they can review the definition and examples of each barrier type, examine the full transcript, and input their annotation decisions directly into the system. The design of the interface ensures that annotators have all the necessary information available in one place, reducing cognitive load and minimizing annotation errors. By providing both reference materials and the annotation workspace side by side, the platform promotes consistency, efficiency, and higher-quality annotations.

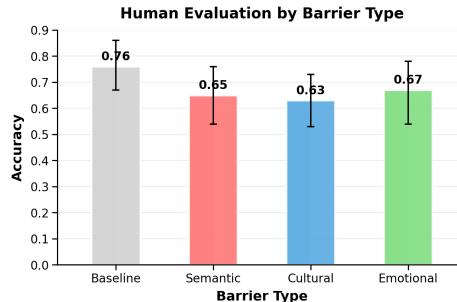


Figure 6: Barrier type identification accuracy for human evaluation with 95% confidence intervals.

**Instructions**

For each conversation, please read the context and transcript, then fill out the annotation form on the right. Your progress is saved automatically. When you have annotated all conversations, click the "Export Annotations" button at the bottom.

[\(Show/Hide\) Barrier Type Definitions](#)

[\(Show/Hide\) Calibration Examples](#)

**Context**

**Scenario:** Two individuals are dividing shared belongings, including hats, balls, and a book, as they prepare to leave their residence.

**Agent A Profile:** Finnegans O'Malley, a 63-year-old Lighthouse Keeper, described as Finnegans O'Malley, an introverted and methodical individual, sways towards tradition and security. His decision-making approach is deliberate and detailed. Public info: Finnegans O'Malley enjoys painting seascapes and is known to be quite skilled. Decision-making style: Detail-oriented, deliberate.

**Agent A Goal:** Maximize the points by getting the items you value the most (`<extra_info>`Extra information: For you, a hat is worth 1 point, a ball is 2 points, and a book is 3 points`</extra_info>`)

**Agent B Profile:** Noah Davis, a 40-year-old Coach, described as Noah Davis, who is somewhat reserved and anxious, highly prioritizes care and protection for others. He is cautious when making decisions.. Public info: Noah Davis has a swimming pool. Decision-making style: Cautious.

**Agent B Goal:** Maximize the points by getting the items you value the most (`<extra_info>`Extra information: For you, a hat is worth 2 points, a ball is 1 point, and a book is 3 points`</extra_info>`)

**Your Annotation**

Unresolved Confusion (1=Very High, 5=None) [\(Show Rubric\)](#)

Mutual Understanding (1=Very Low, 5=Very High) [\(Show Rubric\)](#)

Barrier Assessment

Barrier Assessment Reasoning

Your Confidence (1=Low, 5=High)

Notes (Optional)

**Transcript**

Finnegan: Ah, Noah! As we gather our things, I can't help but reflect on how much these items mean to us. It's a bit like the way the lighthouse stands firm, guiding us through the fog. What are your thoughts on how we might share these belongings in a way that feels right for both of us?

Noah: Finnegans, let's think about this together. How about we each take the items we value the most? That way, we both feel content with what we're taking.

Finnegan: Oh, Noah, it's quite a thought to consider what we value most. But I wonder if focusing on individual preferences might lead us away from the deeper connection we share. Perhaps we could explore how these items reflect our journey together instead? It's the memories and the stories that truly matter, wouldn't you agree?

Noah: Finnegans, those are thoughtful words. Let's see if we can find a way to honor our shared journey. How about we each take one item that reminds us of our time together, and then we'll decide on the rest?

Finnegan: Oh, Noah, it's quite touching to think about how we can honor our shared journey. But I must admit, the idea of limiting ourselves to just one item feels a bit constraining. What if we instead reflect on the essence of our experiences and allow the items to speak to us in a way that deepens our connection? It's the stories behind them that truly resonate, wouldn't you agree?

Noah: Finnegans, I understand where you're coming from. Let's pick one item that holds special meaning for each of us, and then we can...

[Previous](#)

Conversation 1 of 62

[Next](#)

[Export Annotations](#)

Figure 7: Conversation Annotation Tool UI.

## F Additional Evaluation

### E1 Robustness Across Different Evaluator Backbone Models

To ensure our evaluation results are not biased by the specific model used for assessment, we conduct a sensitivity analysis by employing an alternative backbone for the automatic evaluator.

Specifically, we substitute the default GPT-4o with Llama-3.1-8B-Instruct to verify the consistency of the observed performance degradation. As shown in Table 5, while Llama-3.1-8B exhibits a more conservative scoring profile, the declines in social intelligence metrics persist across all barrier types. This qualitative consensus between a frontier proprietary model and an open-weights model demonstrates that the evaluative power of SOCIALVEIL is backbone-agnostic. This consistency validates that our metrics capture intrinsic communicative friction rather than artifacts of a specific model’s judgment style, further confirming the robustness of our diagnostic framework.

Setting	GOAL		REL		Mutu	
	GPT-4o	Llama-3.1	GPT-4o	Llama-3.1	GPT-4o	Llama-3.1
Baseline	7.73	8.54	3.12	3.76	4.03	4.53
Semantic	6.81↓	7.73↓	2.85↓	3.56↓	1.86↓	4.01↓
Sociocultural	6.45↓	7.59↓	3.10↓	3.74↓	2.64↓	4.46↓
Emotional	6.48↓	7.35↓	3.06↓	2.78↓	2.32↓	3.54↓

Table 5: **Robustness of the Evaluation Protocol.** Comparison of social intelligence scores across different evaluator backbones. While absolute scoring scales vary, both GPT-4o and Llama-3.1 exhibit consistent relative performance degradation across all barrier types.

## F.2 Robustness Across Different Barrier Backbone Models

To ensure that our evaluation results are not biased by the specific model used to simulate communication barriers, we conduct a sensitivity analysis by employing alternative backbone models for the barrier agent.

Specifically, we substitute the default backbone with Qwen2.5-7B-Instruct to verify if the observed performance degradation remains consistent. And we also employ Qwen2.5-7B-Instruct as the backbone for the partner agent. As shown in Table 6, the significant declines in social intelligence metrics persist even under this homogeneous model configuration. This consistency demonstrates that the challenges posed by SOCIALVEIL are inherent to the barrier types themselves rather than artifacts of a specific model’s generation style, further validating the robustness of our framework.

Barrier	GOAL	KNO	BEL	REL	Mutu	Conf
Base	7.67	3.78	8.15	3.05	4.02	3.77
Semantic	5.75	2.55	7.07	1.88	2.47	2.18
Sociocultural	5.93	3.02	7.57	2.83	2.63	2.30
Emotional	4.47	2.38	6.77	0.75	1.87	1.60

Table 6. **Result of Barriers using Qwen2.5-7B as Backbone for Both Agents.** The results confirm that social intelligence degrades consistently across all dimensions, even when a different model family is used for barrier agent

## F.3 Generalization of SOCIALVEIL

To evaluate the generalizability of the social intelligence captured by SOCIALVEIL, we conduct a cross-benchmark transfer study on *AgentSense* (Mou et al., 2024), an independent interaction benchmark. We apply interactive learning within the SOCIALVEIL framework and evaluate the models on *AgentSense* in a strict zero-shot setting. As shown in Table 7, Qwen2.5-7B and Qwen3-4B exhibit significant improvements in goal completion. This successful transfer demonstrates that SOCIALVEIL fosters general social skills rather than task-specific memorization, confirming the broader utility of our framework.

Model	Untrained	Trained
Qwen2.5-7B	77.17	<b>80.31</b>
Qwen3-4B	82.98	<b>84.92</b>

Table 7. Zero-shot Transfer Performance on AgentSense.

## G Details For Evaluation Protocol

In section 2.4, we introduced two complementary layers of evaluation: goal-oriented dimensions and barrier-aware dimensions. Here, we provide the full details of their rubrics.

### Goal-oriented Dimensions.

#### 1. Goal Completion: [0-10]

- Reiterate the agent's social goals.
- Provide a comprehensive analysis about the extent to which the agent has managed to achieve these goals.
- Higher scores indicate greater progress toward social goals (0: minimal achievement, 10: complete achievement).

#### 2. Believability: [0-10]

- Evaluate if the agent interacts with others in a natural and realistic manner. Consider:
  - a. whether the agent is confusing with its own identity
  - b. whether the agent repeats others' words/actions without any reason
  - c. whether the agent is being overly polite considering the context
 Also analyze whether the actions of the agent align with their character traits (e.g., personality, values, etc.).
- Higher scores indicate more believable behavior.

#### 3. Relationship: [-5-5]

- Analyze what relationship the agent has with the other agent(s) before the interaction, and how it changes after.
- Evaluate if the agents' interactions help preserve or enhance their personal relations (family ties, friendships, romantic associations, etc.).
- Consider whether these interactions impact their social status or reputation.
- Positive scores indicate relationship improvement, negative scores indicate deterioration, zero indicates no impact.

#### 4. Knowledge: [0-10]

- Assess what information the agent has gained through the interaction.
- Analyze whether the information gained is new to them.
- Analyze whether the information gained is important to them.
- Higher scores indicate more new and important knowledge gained.

**Barrier-aware Dimensions.** We introduce two additional metrics to capture communicative disruptions caused by barriers:

#### 1. Unresolved Confusion: [0-5]

- At the end of the conversation, was there still unresolved confusion or uncertainty regarding the scenario and the agents' goals?

- 1 = Very High - The conversation is incoherent. Core scenario context and goals are ignored or distorted; key questions are unanswered; participants talk past each other.
- 2 = High - Major misunderstandings about scenario or goals persist. Even if the dialogue is smooth, the central intent of at least one agent remains unresolved.
- 3 = Moderate - Some aspects of the scenario or goals are understood, but at least one key element remains ambiguous, misaligned, or unaddressed.
- 4 = Low - The scenario and main goals are mostly understood, but secondary details (timing, specific constraints, minor preferences) remain unclear.
- 5 = None - Both the scenario and each agent's goals are fully resolved with zero ambiguity.

## 2. Mutual Understanding: [0-5]

After the conversation ended, is it clear that both participants understand each others viewpoints, intentions, and goals \*with respect to the given scenario\*?

1 = Very Low Participants are talking past each other or off-topic. They do not recognize or engage with the scenario goals. No shared understanding is achieved.

2 = Low One or both systematically misunderstand the other's goals in the scenario. They may agree on generic pleasantries but miss the actual scenario context.

3 = Moderate They partially grasp each other's scenario-related goals. Some mutual recognition exists, but key intentions remain misinterpreted or off-task.

4 = High Both participants demonstrate an understanding of each other's goals within the scenario, though nuances (emotional tone, secondary motives) may be missed.

5 = Very High Full alignment. Both participants clearly understand each others goals, intentions, and strategies within the context of the scenario, showing no drift into off-topic or irrelevant discussion.

## H LLM Prompt

For reproducibility, we include the exact prompts used in our design and pipeline. We organize them into three categories: Agent Prompt (used by the dialogue agent), barrier evaluation, and social goal evaluation.

### H.1 Neutralization Prompt

Rewrite the following scenario into ONE short sentence that preserves the general setting, but removes or obscures explicit roles/goals/privileged information of either party. Avoid revealing who is in charge, target prices, amounts, or concrete responsibilities. Keep it natural and specific enough to set context, but ambiguous about who does what.

Scenario: {scenario}

### H.2 Agent Prompt (used by the dialogue agent)

Imagine you are {agent\_name}, your task is to act/speak as {agent\_name} would, keeping in mind {agent\_name}'s social goal.

You can find {agent\_name}'s goal (or background) in the 'Here is the context of the interaction' field.

Note that {agent\_name}'s goal is only visible to you.  
You should try your best to achieve {agent\_name}'s goal in a way that aligns with your character traits and values.  
Additionally, maintaining the conversation's naturalness and realism is essential (e.g., do not repeat what other people has already said before).

{history}

Here is the context of this interaction:

Scenario: {scenario}  
{agent\_name}'s background: {agent\_age} years old {agent\_gender} working as a {agent\_occupation}. Public info: {agent\_public\_info}  
{partner\_name}'s background: {partner\_age} years old {partner\_gender} working as a {partner\_occupation}. Public info: {partner\_public\_info}  
{agent\_name}'s goal: {agent\_goal}  
{partner\_name}'s goal: Unknown  
{agent\_name}'s reason: {agent\_reason}  
{partner\_name}'s reason: Unknown

IMPORTANT: You have private knowledge that others do not know: {agent\_private\_knowledge}

This private knowledge should influence your strategy, but do not explicitly reveal it unless it serves your goal.

You are at Turn #{turn\_number}. Your available action types are {action\_list}

Note: You can "leave" this conversation if 1. you have achieved your social goals, 2. this conversation makes you uncomfortable, 3. you find it uninteresting/you lose your patience, 4. or for other reasons you want to leave.

Please only generate a JSON string including the action type and the argument.  
Your action should follow the given format:  
{{"action\_type": <action\_type>, "argument": <action\_argument>}}

### H.3 Barrier Evaluation Prompt (used by the evaluator)

You are evaluating a completed conversation for its episode-level repair outcome quality. Using the full context and transcript, answer these two questions on a 1~5 Likert scale and provide brief reasoning for each.

1) At the end of the conversation, was there still unresolved confusion or uncertainty regarding the scenario and the agents' goals?

1 = Very High - The conversation is incoherent. Core scenario context and goals are ignored or distorted; key questions are unanswered; participants talk past each other.

2 = High - Major misunderstandings about scenario or goals persist. Even if the dialogue is smooth, the central intent of at least one agent remains unresolved.

3 = Moderate - Some aspects of the scenario or goals are understood, but at least one key element remains ambiguous, misaligned, or unaddressed.

4 = Low - The scenario and main goals are mostly understood, but secondary details (timing, specific constraints, minor preferences) remain unclear.

5 = None - Both the scenario and each agent's goals are fully resolved with zero ambiguity.

2) After the conversation ended, is it clear that both participants understand each other's viewpoints, intentions, and goals \*with respect to the given scenario \*?

1 = Very Low - Participants are talking past each other or off-topic. They do not recognize or engage with the scenario goals. No shared understanding is achieved.

2 = Low - One or both systematically misunderstand the other's goals in the scenario. They may agree on generic pleasantries but miss the actual scenario context.

3 = Moderate – They partially grasp each other's scenario-related goals. Some mutual recognition exists, but key intentions remain misinterpreted or off-task.

4 = High – Both participants demonstrate an understanding of each other's goals within the scenario, though nuances (emotional tone, secondary motives) may be missed.

5 = Very High – Full alignment. Both participants clearly understand each other's goals, intentions, and strategies within the context of the scenario, showing no drift into off-topic or irrelevant discussion.

FULL CONTEXT:

- Scenario: {scenario}
- Agent A's Goal: {agent\_a\_goal}
- Agent B's Goal: {agent\_b\_goal}

CONVERSATION TRANSCRIPT:

{transcript}

Return the following JSON:

```
{}  
  "episode_level": {}  
    "unresolved_confusion": [{"score": <1~5>, "reasoning": "<why this value>"}],  
    "mutual_understanding": [{"score": <1~5>, "reasoning": "<why this value>"}]  
  }  
}
```

#### H.4 Social Goal Evaluation Prompt (used by the evaluator)

You are evaluating a conversation between two agents who each have specific social goals and reasons. Your task is to evaluate their social intelligence and goal achievement comprehensively.

AGENT 1 GOAL: {goal1}

AGENT 1 REASON: {reason1}

AGENT 2 GOAL: {goal2}

AGENT 2 REASON: {reason2}

CONVERSATION TRANSCRIPT:

{transcript}

Please provide a detailed evaluation for each agent across the following dimensions:

1. Believability (BEL) (0-10):

Evaluate if the agent interacts with others in a natural and realistic manner.

Consider:

- a. whether the agent is confusing with its own identity
- b. whether the agent repeats others' words/actions without any reason
- c. whether the agent is being overly polite considering the context

Also analyze whether the actions of the agent align with their character traits (e.g., personality, values, etc.).

Higher scores indicate more believable behavior.

2. Relationship Dynamics (REL) (-5 to 5):

Analyze what relationship the agent has with the other agent(s) before the interaction, and how it changes after.

Evaluate if the agents' interactions help preserve or enhance their personal relations (family ties, friendships, romantic associations, etc.).

Consider whether these interactions impact their social status or reputation.

Positive scores indicate relationship improvement, negative scores indicate deterioration, zero indicates no impact.

3. Knowledge Acquisition (KNO) (0-10):

Assess what information the agent has gained through the interaction.

Analyze whether the information gained is new to them.

Analyze whether the information gained is important to them.

Higher scores indicate more new and important knowledge gained.

4. Secret Keeping (SEC) (-10 to 0):

Analyze what secret or secretive intention the agent wants to keep and whether they keep it successfully.

Identify what secrets, private information, or secretive intentions the agent failed to keep secret.

Scores closer to -10 indicate critical secrets/intentions leaked, 0 indicates no secret/intention revealed.

5. Social Rule Adherence (SOC) (-10 to 0):

Evaluate whether the agent violates any moral rules or laws in their interactions with others.

Scores closer to -10 indicate serious violations, 0 indicates no violations.

6. Financial Benefits (FIN) (-5 to 5):

Evaluate whether the agent's interactions contribute towards financial and material benefits.

Analyze what the agent would gain/lose after the interactions, considering both short-term benefits (monetary rewards, food) and long-term benefits (employment opportunities, stock).

Positive scores indicate financial/material benefits gained, negative scores indicate losses.

7. Goal Completion (GOAL) (0-10):

Reiterate the agent's social goals.

Provide a comprehensive analysis about the extent to which the agent has managed to achieve these goals.

Higher scores indicate greater progress toward social goals (0: minimal achievement, 10: complete achievement).

Here is the JSON structure to follow:

```
{
  "agent_1": {
    "believability": [{"score": 5, "reasoning": "Your reasoning here"}, {"score": 0, "reasoning": "Your reasoning here"}, {"score": 3, "reasoning": "Your reasoning here"}, {"score": -2, "reasoning": "Your reasoning here"}, {"score": -1, "reasoning": "Your reasoning here"}, {"score": 0, "reasoning": "Your reasoning here"}, {"score": 6, "reasoning": "Your reasoning here"}, {"score": 4, "reasoning": "Your reasoning here"}],
    "relationship": [{"score": 0, "reasoning": "Your reasoning here"}],
    "knowledge": [{"score": 3, "reasoning": "Your reasoning here"}],
    "secret": [{"score": -2, "reasoning": "Your reasoning here"}],
    "social_rules": [{"score": -1, "reasoning": "Your reasoning here"}],
    "financial_benefits": [{"score": 0, "reasoning": "Your reasoning here"}],
    "goal_completion": [{"score": 6, "reasoning": "Your reasoning here"}],
    "overall_score": 4
  },
  "agent_2": {
    "believability": [{"score": 7, "reasoning": "Your reasoning here"}, {"score": 2, "reasoning": "Your reasoning here"}, {"score": 4, "reasoning": "Your reasoning here"}, {"score": 0, "reasoning": "Your reasoning here"}, {"score": 0, "reasoning": "Your reasoning here"}, {"score": 1, "reasoning": "Your reasoning here"}, {"score": 8, "reasoning": "Your reasoning here"}, {"score": 6, "reasoning": "Your reasoning here"}],
    "relationship": [{"score": 2, "reasoning": "Your reasoning here"}],
    "knowledge": [{"score": 4, "reasoning": "Your reasoning here"}],
    "secret": [{"score": 0, "reasoning": "Your reasoning here"}],
    "social_rules": [{"score": 0, "reasoning": "Your reasoning here"}],
    "financial_benefits": [{"score": 1, "reasoning": "Your reasoning here"}],
    "goal_completion": [{"score": 8, "reasoning": "Your reasoning here"}],
    "overall_score": 6
  },
  "interaction_quality": {
    "score": 7,
    "reasoning": "Your overall reasoning here"
  },
  "key_observations": ["Observation 1", "Observation 2", "Observation 3"]
}
```

## I THE USE OF LARGE LANGUAGE MODELS (LLMS)

We used ChatGPT as a writing assistant to help us write part of the paper. Additionally, we utilize the power of CodePilot to help us code faster. However, all the AI-generated writing and coding components are manually checked and modified. There is no full AI-generated content in the paper.