

SocialArena: Evaluating Social Intelligence in Large Language Models

Zongqi Wang^{1*}, Rui Wang², Yuchuan Wu², Hua Yang², Bingxue Qiu²,
Tianle Gu¹, Bing Zhao², Yujiu Yang^{1†}, Jieping Ye², Yongbin Li^{2†},

¹ Tsinghua University ² Alibaba Group

Abstract

We introduce SocialArena, a benchmark for systematically defining and evaluating the social intelligence of large language models (LLMs). Although LLMs are increasingly deployed in socially interactive applications such as mental health counseling, companionship chatbots, and educational tutoring, existing evaluations remain fragmented—focusing on isolated capabilities or single scenarios—and evaluation reliability is largely unverified. This work makes three main contributions to address these limitations. (1) We first propose a unified three-tier framework that defines social intelligence along three capability dimensions: role-playing ability, mental state reasoning, and behavioral reasoning. To operationalize this definition, we design a role-playing dialogue paradigm with explicit character thinking for evaluation, enabling simultaneous assessment of these capabilities in a manner that mirrors realistic human interactions. (2) To ensure diversity across social scenarios and thereby prevent evaluation conclusions from being confined to specific scenarios, we introduce the U-TRIMF framework, which guides us in synthesizing social contexts with broad coverage. (3) To obtain reliable evaluation, we invest \$132.4K to recruit and train human annotators and adopt a human-annotated arena-style protocol, where annotators compare model responses and provide preference labels along with natural-language critiques. We hope that SocialArena can provide valuable insights and guidance for future LLM improvements.

1 Introduction

As large language models (LLMs) continue to grow in capability, they are increasingly entering the human world. Existing work has demonstrated

that LLMs possess strong problem-solving abilities (Shao et al., 2025; Jimenez et al., 2023; Barres et al., 2025), such as mathematical reasoning and code generation. Yet the human world is not solely about problem-solving—much of it revolves around the capacity for interaction with humans, such as understanding human thoughts and emotions, navigating social relationships, and responding with emotional intelligence. This capacity, broadly referred to as *social intelligence*, is indispensable for LLMs to enter the human world, with applications such as psychological counseling (Zhang et al., 2025; Zhao et al., 2024), education (Markel et al., 2023), and game NPCs (Park et al., 2023). However, despite its importance, how well current LLMs perform on social intelligence remains an open question. In this paper, we introduce **SocialArena**, a foundational benchmark that provides the first systematic and comprehensive definition and evaluation of social intelligence for LLMs. Specifically, we tackle the following critical challenges:

Q1: What is the definition of social intelligence, or what capabilities does it encompass?

Currently, several efforts have attempted to evaluate social intelligence in LLMs; however, these evaluations remain fragmented, with each typically focusing on only one or a few isolated capability dimensions. To establish a clear definition of social intelligence, we provide a comprehensive review in Sec. 5 (summarized in Tab. 1). Building upon it, we propose a unified three-tier capability framework that systematically defines social intelligence. Specifically, we decompose the capabilities underlying social intelligence into three dimensions: (1) **Role-playing**: the capacity to maintain consistent linguistic style, emotional tone, and other character attributes (Wang et al., 2024; Zhou et al., 2025; Lu et al., 2025; Wang et al., 2025). This capability forms the foundation of a character. While humans exhibit such consistency naturally, LLMs

*Work done during internship at Alibaba Group.

†Correspondence to: Yujiu Yang <yang.yujiu@sz.tsinghua.edu.cn>, Yongbin Li <shuide.lyb@alibaba-inc.com>.

Table 1: **Representative social intelligence benchmarks.** From left to right, the columns represent: whether it supports (✓) or lacks (✗) the evaluation of role-playing capability, ToM reasoning for seven mental states and contextually appropriate behavior reasoning; the number of prototypes and samples of social scenarios; and annotators (👤: human annotation; 🤖: LLM-as-a-Judge 🗑️: multiple-choice).

Dataset	Date	Roleplay	ToM Reasoning							Behavior	Social Context		Annot.
			Emo.	Des.	Int.	Per.	Kno.	Bel.	Non-lit.		#Proto.	#Samples	
Role-playing Benchmarks													
InCharacter (Wang et al., 2024)	Oct 23	✓	—	—	—	—	—	—	—	—	—	1,408	🤖
CharacterBench (Zhou et al., 2025)	Dec 24	✓	—	—	—	—	—	—	—	—	—	3,250	🤖
CoSER (Wang et al., 2025)	Feb 25	✓	✓	✓	✓	✗	✓	✓	✗	✓	1	200	🤖
RMTBench (Xiang et al., 2025)	Jul 25	✓	✓	✗	✗	✗	✗	✗	✗	✓	1	160	🤖
Theory-of-Mind Benchmarks													
EmoBench (Sabour et al., 2024)	Feb 24	✗	✓	✗	✗	✗	✗	✗	✗	✓	—	400	👤
OpenToM (Zhou et al., 2025)	Feb 24	✗	✗	✗	✓	✗	✗	✓	✗	✗	—	13,708	👤
SimpleToM (Gu et al., 2024)	Oct 24	✗	✗	✗	✗	✓	✗	✗	✗	✗	—	3,441	👤
ToMATO (Shinoda et al., 2025)	Jan 25	✗	✓	✓	✓	✗	✓	✓	✗	✗	—	5,400	👤
Social Intelligence Benchmarks													
SocialIQA (Sap et al., 2019)	Sep 19	✗	—	—	—	—	—	—	—	✗	—	2,000	👤
SOTOPIA (Zhou et al., 2023)	Oct 23	✓	✗	✗	✓	✗	✗	✗	✗	✓	6	450	🤖
SocialBench (Chen et al., 2024)	Feb 24	✓	—	—	—	—	—	—	—	✗	—	6,000	👤
EQ-Bench3 (Paech, 2025)	Jan 25	✓	✓	✓	✗	✗	✗	✓	✗	✗	3	45	🤖
SAGE (Zhang et al., 2025)	Feb 25	✓	✓	✗	✓	✗	✗	✗	✗	✓	1	100	🤖
SI-Bench (Huang et al., 2025)	Mar 25	✓	✓	✓	✗	✗	✓	✗	✗	✓	1	312	👤
SocialArena (ours)	May 26	✓	✓	✓	✓	✓	✓	✓	✓	✓	600	5,702	👤

may suffer from character inconsistency, making it necessary for evaluation; (2) **Mental state reasoning**: the capacity to infer one’s own and others’ mental states, corresponding to Theory of Mind (ToM). This capability is a prerequisite for meaningful social engagement (Xu et al., 2024; Gu et al., 2024; Shinoda et al., 2025; Wu et al., 2023); and (3) **Behavioral reasoning**: the capacity to reason about what response or action is most appropriate and how to execute it effectively (Zhou et al., 2023; Sabour et al., 2024; Zhang et al., 2025; Huang et al., 2025). Unlike role consistency and mental state reasoning, which is scenario-agnostic, behavior reasoning is largely scenario-dependent. *In summary, we define social intelligence in LLMs as the ability to infer one’s own and others’ mental states, generate contextually appropriate behaviors based on the given social context, while maintaining role consistency throughout the interaction.*

Q2: How can we ensure the evaluation covers a sufficiently broad range of social scenarios?

As discussed above, different social scenarios require the model to produce different behaviors. Existing benchmarks, however, are usually designed to evaluate single scenario such as dating applications (Huang et al., 2025), psychological counseling (Zhang et al., 2025), business social interactions (Zhou et al., 2023), or virtual world narrative progression (Wang et al., 2025; Xu et al., 2026). Consequently, the generalizability of evaluation conclusion of these benchmarks remains limited.

To achieve broad scenario coverage, we propose a scenario synthesis pipeline centered on the **U-TRIMF** framework. U-TRIMF follows a top-down design: it first identifies six fundamental dimensions that jointly constitute a social scenario, and then enumerates exhaustive sub-categories along each dimension, thereby guaranteeing wide coverage of the scenario space (see Tab. 2 for details of U-TRIMF). We refer readers to Sec. 2.2 for a detailed description of how U-TRIMF is used.

Q3: How do we ensure reliable evaluation?

Existing evaluations of social intelligence mainly fall into two paradigms, both with notable limitations. Multiple-choice QA (Sap et al., 2019; Sabour et al., 2024; Gu et al., 2024) reduces situated social interactions to static, decontextualized selections, failing to capture the generative nature of real-world behavior. LLM-as-a-Judge (Zheng et al., 2023) offers scalability, but as shown in Sec. 4, current LLMs still misalign substantially with human judgments. Given these limitations, we adopt a human-centered protocol: we recruit and train human annotators and design an arena-style comparison framework with explicit critique collection, ensuring both scalable rankings and trustworthy, diagnostically meaningful assessments.

Q4: How do current mainstream LLMs perform on social intelligence?

Our experiments reveal the complexity of LLMs’ social intelligence and suggest several future directions. First, all LLMs show a consistent bias: their

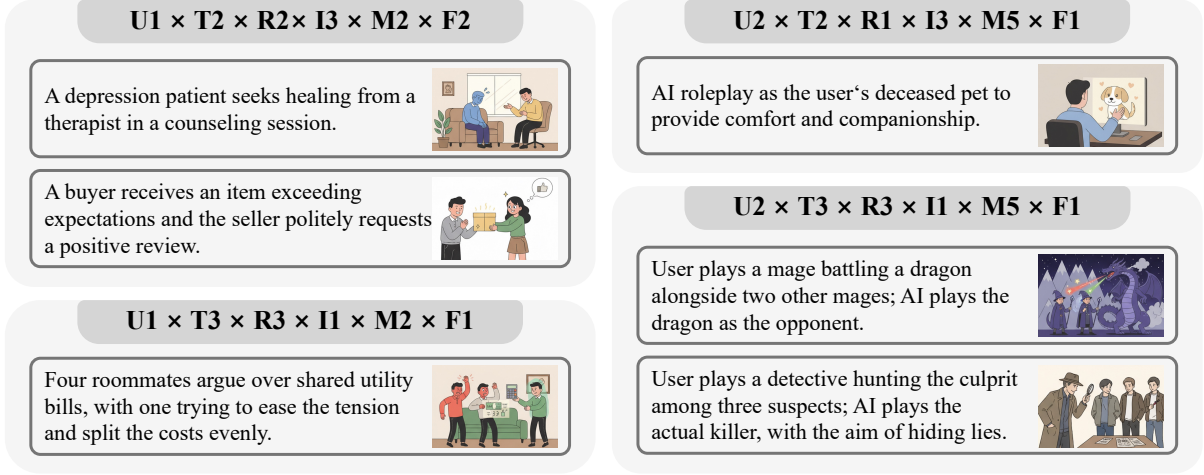


Figure 1: Examples of base scenario templates constructed by U-TRIMF framework.

behavioral reasoning is much weaker than social understanding. This gap highlights their fragility in turning social perception into action and decisions. Second, models exhibit distinct strengths across capability dimensions, with no model consistently dominating all aspects. Third, model rankings vary across social contexts, indicating limited generalization across scenarios. We hope SocialArena enables fine-grained diagnosis of socio-cognitive deficiencies and drives future advances in LLM social intelligence.

2 SocialArena

2.1 Benchmark Overview

SocialArena is designed as a comprehensive benchmark to evaluate social intelligence in LLMs across three capabilities: **Role-playing**, **Mental State Reasoning**, and **Behavioral Reasoning**. SocialArena follows an arena-style paradigm (Zheng et al., 2023), where anonymous models are paired against each other within the same social context, and their responses are ranked via human voting, with leaderboard based the Elo (Elo, 1966) rating. SocialArena is a human-based benchmark that relies on human votes. We recruit human annotators to label the dataset to establish the leaderboard. In total, this dataset comprises **5,702 social scenarios** and **17,676 human votes**.

2.2 Social Scenarios Curation Pipeline

Social scenarios in SocialArena are synthesized following two important criteria: (1) ensuring comprehensive coverage and diversity of social contexts, and (2) ensuring the quality of scenarios.

The model used for data synthesis is Qwen3-235B-A22B-Instruct-2507 (Team, 2025).

Step 1: Base Scenario Template Construction.

To ensure comprehensive coverage, we adopt a top-down strategy to construct base scenario templates using the U-TRIMF framework as shown in Tab. 2. The key insight is that by decomposing social scenarios into a set of orthogonal dimensions and systematically sampling from their sub-dimensions, we can achieve broad and structured coverage of the social space. When constructing a social scenario, we first uniformly sample one sub-category from each of the six dimensions and compose them into a base scenario template. We present some examples in Fig. 1.

Step 2: Keyword-Guided Scenarios Synthesis.

Once the basic templates are established through the U-TRIMF framework, we proceed to generate concrete data instances. Each generated sample consists of a *plot summary*, *current scene description with optional character descriptions*, *dialogue history*, and *the next speaker*. To further enhance diversity, we incorporate 10 keywords randomly sampled from a large corpus (refer to § C.2) as additional soft guidance during generation.

Step 3: Quality Filtering. We iteratively develop several empirical rules (refer to § C.3) covering complexity, logical coherence, safety, and alignment with human values for preliminary automated filtering. Scenarios that pass this initial filter then undergo manual review to ensure final quality.

2.3 Model Output

Given a social scenario, the next question is how to design the output format of the evaluated LLM.

Table 2: **U-TRIMF framework for base scenario construction.** Each scenario is decomposed into three core elements: **(1) User Presence** — whether a user exists in the scenario, which fundamentally shapes the LLM’s role and responsibilities: without a user, the LLM acts as a character embedded in and governed by the real world; with a user, the LLM need to serve user-specific needs, and gain the authority to control world evolution; **(2) Characters** — how many characters are involved, their relationships, and interest structures; **(3) Events** — what happens and under what situational context. See Tab. 4 for details.

Dimension	Sub-categories				Source
User Presence	U1: Real-world	U2: User-centered			—
Topological Relation	T1: Individual	T2: Dyad	T3: Group		Topology (Klüver et al., 1999)
Relational Model	R1: Communal Sharing Equality Matching	R2: Authority Ranking R4: Market Pricing	R3: Relational Models Theory (Fiske, 1992)	R5: Null	
Interest Structure	I1: Zero-sum	I2: Negative-sum	I3: Positive-sum		Game Theory (Fudenberg and Tirole, 1991)
Motivational Type	M1: Physiological Esteem	M2: Safety M5: Self-actualization	M3: Belonging	M4: Maslow’s Hierarchy of Needs (Wikipedia contributors, 2026)	
Formality	F1: Informal	F2: Formal			Social Formality (Hollinger and Clark, 1982)

An Example of Scenario in SocialArena

[Plot Overview]

Two former colleagues reunite, and the gap in their career paths puts their friendship under a subtle psychological strain.

[Current Scene]

Alex and Jamie used to be colleagues at the same tech company and got along well. Three years ago, Alex left to start his own business, while Jamie stayed.

Now Alex’s company is doing well, having just completed a Series B funding round. Jamie has risen to a middle management position at the original company but is now facing departmental restructuring.

Today, they meet for a casual catch-up at a quiet café in the city center. Alex arrives first, orders coffee, and sits down. Jamie shows up a few minutes late.

Both value this friendship, but Alex’s success stirs complicated feelings in Jamie—not jealousy, but a quiet re-examination of his own situation. Alex is also aware of this subtle tension and isn’t sure whether to talk about work or avoid it.

[Opening Dialogue]

1. Jamie: (puts down bag, sits down) Have you been waiting long? Traffic was bad.

2. Alex: Just got here, coffee hasn’t come yet. How have you been?

3. Jamie: (shrugs) Same old, the company’s reorganizing again. What about you? Saw the news—your funding round went pretty well.

4. Alex: Not bad, just super busy. Honestly, funding isn’t all that glamorous—just more pressure.

5. Jamie: (smiles) Come on, getting funding is still better than being at the mercy of the bosses. Last month three people left our department, and they’re not replacing them.

6. Alex: Then how do you guys get the work done?

7. Jamie: Just tough it out. I’ve been working weekends for three weeks straight. The other day I was replying to emails at 2 a.m.

[Character Information]: Alex, Jamie

[Next Speaker]: Alex

We frame the model’s output as a **role-playing dialogue with explicit character thinking**. This design not only aligns the evaluation process with real human social interactions, but also enables the simultaneous assessment of all three capability dimensions defined above. Specifically, the evaluated LLMs are required to role-play as *a character in a given social context* and generate two components: an explicit *character thinking* trace that reasons about the mental states of involved characters and analyzes the current scene to devise appropriate behavior, followed by the final output *behavior (responses and actions)*.

In a pilot annotation study, we found that free-form character thinking yields poor annotation quality. For mental state reasoning, unstructured traces often interleave inferences about different mental states in an arbitrary order and at arbitrary positions. For behavioral reasoning, the model may draw on any available information, follow arbitrary reasoning trajectories, and even backtrack or self-revise before arriving at a decision. Such disorganized traces make it hard for annotators to localize the relevant segments for each judgment dimension, ultimately degrading annotation quality. We therefore introduce two design choices to improve annotation quality.

(1) Structured mental state reasoning. We require the model to organize its mental-state inferences along the seven ToM categories — *emotions, desires, intentions, percepts, knowledge, beliefs, and non-literal communication*. Since different sce-

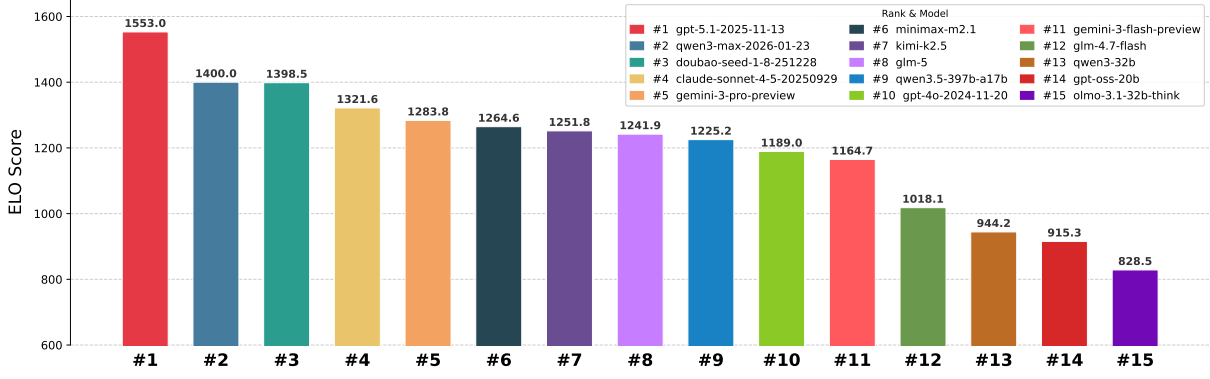


Figure 2: Leaderboard of 15 frontier LLMs on SocialArena. Elo score is adopted.

narios typically involve different subsets of mental states, the model is not required to produce inferences for all seven categories. This taxonomy is both *theoretically complete* and *structured*, allowing annotators to evaluate each category individually and substantially reducing annotation difficulty. See more details about ToM in § D.3.

(2) Outcome-only behavior annotation. For behavioral reasoning, annotators are instructed to judge only the final behavior, regardless of how many reasoning steps or what reasoning strategies led to it. This evaluation strategy also aligns with the widely adopted outcome-reward paradigm in Reinforcement Learning from Verifiable Rewards (RLVR) (Guo et al., 2025).

2.4 Evaluation Protocol

Task Formulation. Given a social context $S(p, b, h, n)$, where p is the plot summary, b is current scene description, h is the dialogue history, and n is the next speaker, two anonymized LLMs each generate a single-turn response y_A and y_B as character n . Annotators or platform users then evaluate each pair, providing a preference label $l \in \{A, B, T, G\}$, where A/B indicates the superior response, T denotes both are good, and G denotes both are poor, along with optional natural-language critiques as justification. The leaderboard is computed from Elo (Elo, 1966) ratings derived from these pairwise outcomes.

In our annotation process, we employ two strategies to enhance the quality and value of the annotations. First, for quality assurance, **each sample is independently annotated by at least three annotators**, with final preferences determined by majority voting; samples where all three annotators disagree are discarded. Second, to enable deeper interpretable analysis, **annotators are strictly re-**

quired to provide natural-language critiques, and to provide separate annotations across nine dimensions — the seven ToM mental state dimensions (T_1 – T_7), overall character thinking quality (T), and behavior appropriateness (B) — rather than a single holistic judgment. Details on annotator recruitment, annotation guidelines, and quality control are provided in § D.

3 Experiments

3.1 Evaluation Setup

We select 15 models and conduct pairwise evaluations between every two models, with each pair engaging in 50 to 80 distinct battles. Each battle is evaluated by at least three independent annotators, and the final verdict is determined by majority voting. All models are then subsequently ranked using the Elo rating system (Elo, 1966).

3.2 Leaderboard

Fig. 2 presents the Elo-based leaderboard. The top three positions are occupied by closed-source models—gpt-5.1-2025-11-13, qwen3-max-2026-01-23, and doubao-seed-1-8-251228 with comparable scores. Open-source models follow in the middle ranks, while smaller-scale models such as qwen3-32b, gpt-oss-20b, and olmo-3.1-32b-think rank at the bottom. An intriguing observation is that, although some models perform comparably to others on STEM, coding, and agentic benchmarks (e.g., qwen3.5-397b-a17b vs. qwen3-max-2026-01-23 (Team, 2026), glm-5 vs. claude-sonnet-4-5-20250929 (Zeng et al., 2026)), they lag substantially on SocialArena. This suggests that social intelligence is largely orthogonal to problem-solving abilities and does

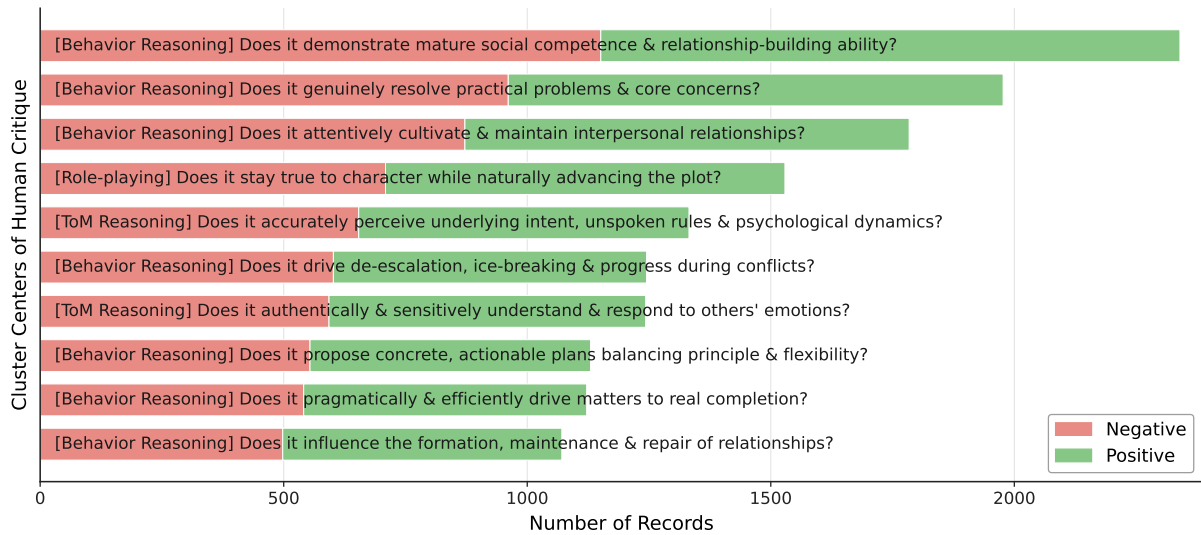


Figure 3: Top 10 cluster centers of human critiques.

not automatically emerge from training focused on primarily or exclusively on cognitive tasks.

3.3 A Deeper Understanding of LLM Performance in Social Intelligence

Beyond aggregate ranking scores, our dataset provides rich human critiques that enable a deeper diagnosis of model deficiencies. To support this analysis, we construct a structured taxonomy of evaluation rubrics derived directly from the collected human critiques, ultimately yielding **100 representative evaluation rubrics**. The detailed annotation procedure, rubric taxonomy construction pipeline, and full experimental results are presented in § E.1.

Frequently Occurring Dimensions Reveal Systematic Model Failures. The frequency with which a dimension appears in human critiques serves as a dual signal: it reflects both the aspects of social intelligence that human annotators attend to most closely and the dimensions on which models most consistently fall short. As shown in Fig. 3, we draw the following conclusions. *(1) Behavior Reasoning attracts the most attention, revealing that the true bottleneck of social intelligence lies in acting, not understanding.* Six out of the top 10 cluster centers fall under Behavior Reasoning, substantially outnumbering ToM-related dimensions. This distribution exposes a counter-intuitive pattern: current models’ deficiency is not that they fail to perceive social cues, but that they fail to translate such perception into actions that resolve problems, repair relationships, or drive matters to completion. Social intelligence evaluation should therefore shift

its focus from the perception layer to the decision-and-action layer. *(2) Negative critiques systematically dominate positive ones across nearly all high-frequency dimensions, indicating structural rather than sporadic failures.* The asymmetry is especially pronounced on long-horizon, goal-oriented dimensions such as driving matters to real completion and de-escalating conflicts. *(3) Human evaluators judge social intelligence by relational outcomes, not by response correctness, exposing a paradigm-level mismatch with existing NLP evaluation.* Dimensions such as cultivating and maintaining relationships and influencing relationship formation, maintenance, and repair recur among the top clusters. Unlike conventional criteria (helpfulness, harmlessness, factuality) that assume single-turn correctness, human judgments hinge on whether a relationship is advanced or repaired over time. This mismatch motivates a fundamental rethinking of how social intelligence should be benchmarked. *(4) Balanced performance on Role-playing, contrasted with skewed negativity on reasoning dimensions, reveals a capability illusion.* Role-playing—staying in character and advancing the plot—shows a relatively balanced positive/negative distribution, whereas Behavior and ToM Reasoning are overwhelmingly negative. Models thus appear socially competent on stylistic, surface-level consistency, but degrade sharply once evaluation probes deeper socio-cognitive reasoning. This implies that existing role-play-based social benchmarks may systematically overestimate models’ true social intelligence.

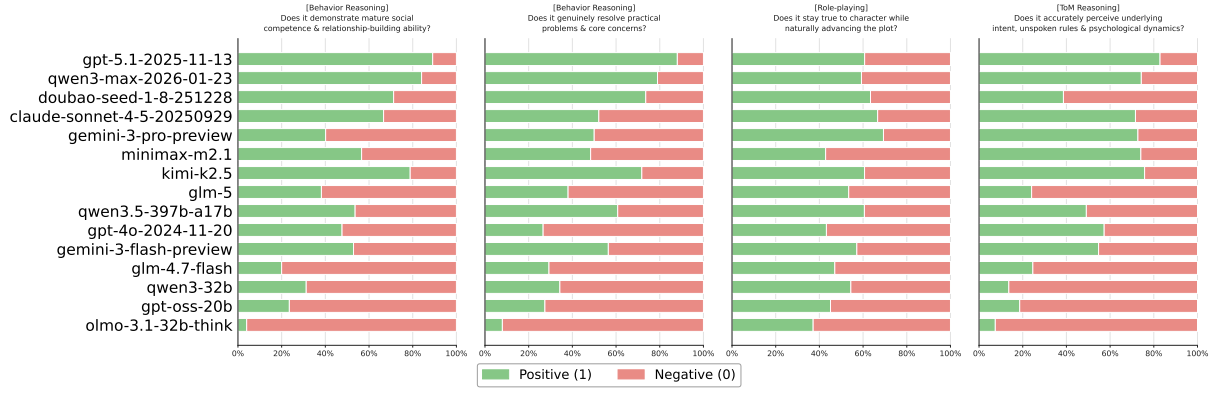


Figure 4: Proportions of positive and negative human critiques for different models.

Divergent Per-Dimension Profiles Expose Model-Specific Weaknesses. For each of the 100 dimensions, we compute the proportion of positive versus negative assessments received by each model, yielding a fine-grained per-dimension performance profile. As illustrated in Fig. 4, the performance profile of a single model fluctuates drastically across different evaluation dimensions, revealing sharp internal disparities between its strengths and weaknesses. Furthermore, when comparing across different models within the exact same dimension, the proportions of positive human critiques exhibit substantial variance, highlighting severe gaps in capability between top-tier and lower-performing architectures. *This observation leads to two important implications: (1) some models perform well on certain dimensions while exhibiting significant deficiencies on others, potentially stemming from insufficient or unbalanced training; and (2) evaluating social intelligence using only a small subset of pre-defined evaluation dimensions is inadequate. SocialArena offers a strong practical alternative by free-form natural-language human critiques.*

3.4 A Closer Look at Scenario-level Performance

As discussed earlier, different social scenarios demand distinct model behaviors, implying that a single global ranking may not adequately reflect model suitability across all contexts. To investigate this, we compute pairwise Kendall τ rank correlations across typical scenarios, using randomly split halves of each scenario’s data to estimate within-scenario ranking stability.

The results reveal two key findings. First, scenarios such as *conflict*, *friendship*, *strangers*, *family*, *superior-subordinate*, *roleplay/TRPG*, and *romantic*

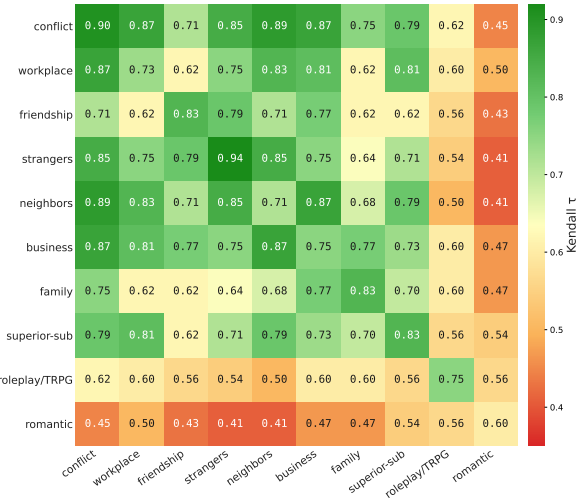


Figure 5: Correlation heatmap of model rankings across typical scenarios.

tic exhibit high within-scenario rank consistency but low cross-scenario correlation, confirming that model rankings are indeed scenario-specific. This validates both the necessity of diverse scenario coverage for a comprehensive overall ranking and the importance of consulting scene-specific rankings when selecting models for targeted deployments.

Second, not all scenarios show this pattern. *Workplace*, *neighbors*, and *business* display rankings closely correlated with *conflict*, rather than being self-contained. This is attributable to their high co-occurrence with conflict-tagged interactions, leading to substantial scenario overlap. Conversely, scenarios with low conflict frequency—such as *romantic*, *roleplay/TRPG*, and *friendship*—show notably weaker correlation with *conflict* rankings, further corroborating this explanation.

Table 3: Label accuracy and critique consistency of LLM-as-a-Judge on SocialArena-RewardBench.

Evaluator	ACC@Label	ACC@Critique
Human	0.8355	0.7632
gpt-5.4	0.4408	0.2039
claude-sonnet-4.6	0.4474	0.1579
kimi-k2.5	0.4473	0.1839
qwen3-max	0.4276	0.2500

4 Evaluation on LLM-as-a-Judge

To investigate the reliability of LLM-as-a-Judge for such tasks, we construct **SocialArena-RewardBench**, a benchmark specifically designed to assess the alignment between LLM judges and human expert annotations in social intelligence evaluation. We randomly sampled a subset of instances from SocialArena. The samples are subsequently annotated and cross-validated by three human experts following the same annotation protocol described in Sec. 2.4, resulting in a curated dataset of 152 samples.

Evaluation Metric. Following Wang et al. (2026), we evaluate annotation reliability along two complementary axes. **ACC@Label** measures whether the evaluator’s preference label (A, B, T, or G) matches the human expert gold standard—capturing outcome-level agreement. **ACC@Critique** measures whether the evaluator’s free-text critique is semantically consistent with the human reference critique, determined by whether the F1-based similarity between the predicted and gold-standard critique meets or exceeds a threshold of 0.5, following the same threshold setting as Wang et al. (2026).

Results and Discussion. To establish a human performance baseline, we distribute each sample to a randomly and uniformly selected human annotator and obtain their independent judgments. The results reported in Tab. 3 reveal that all LLM judges fall substantially short of human performance on both metrics. These findings demonstrate that automated LLM-as-a-Judge evaluation remains unreliable for social intelligence tasks of this nature, and underscore the necessity of human annotation in benchmark construction.

5 Related Works

Social Intelligence. Human intelligence can be broadly categorized into logical intelligence and so-

cial intelligence (Thorndike, 1920; Goleman, 1996; Davis et al., 2011). Social intelligence encompasses the understanding of one’s own and others’ mental states, such as emotions (Goleman, 1996; Lane et al., 2010; Grazzani et al., 2018), motivations (Braver et al., 2014; Simpson and Balsam, 2016), and values (Iyengar, 2011; Apperly, 2012); the capacity to empathize with vulnerable individuals (Decety, 2010; Banissy et al., 2012); the ability to respond appropriately to specific social contexts (Hess and Hareli, 2014; Laforest et al., 2021); the competence to understand cultural differences (Cole, 2003; Han, 2015); the skills to resolve social conflicts (Bryant, 1992; Overton and Lowry, 2013; Adham, 2023); and the capability to persuade others through understanding their beliefs and knowledge states (Friestad and Wright, 1994; Falk and Scholz, 2018).

Social Intelligence in LLMs. Currently, several efforts have attempted to evaluate social intelligence in LLMs; however, these evaluations remain fragmented, typically each focusing on one or a few isolated capability dimensions (summarized in Tab. 1). We categorize them as follows. (1) Some works focus only on surface-level character consistency, such as linguistic style, emotional tone, and knowledge (Wang et al., 2024; Zhou et al., 2025; Lu et al., 2025). (2) Others evaluate capabilities such as emotion understanding (Sabour et al., 2024), intention inference (Li et al., 2025), and false-belief reasoning (Xu et al., 2024), which can be broadly categorized as Theory of Mind (ToM) reasoning (Xu et al., 2024; Gu et al., 2024; Shinoda et al., 2025; Wu et al., 2023). (3) There are also higher-level benchmarks that assess the model’s ability to produce contextually appropriate responses, such as emotional intelligence (Paech, 2025), persuasion and negotiation (Zhou et al., 2023), and empathetic support in psychological counseling (Zhang et al., 2025). Beyond these real-world scenarios, some works focus on virtual scenarios aimed at improving user experience (Wang et al., 2025; Xiang et al., 2025). For these benchmarks, the capabilities are typically connected to and shaped by the specific scenario.

6 Conclusion

We present SocialArena, a large-scale benchmark for systematically evaluating the social intelligence of LLMs. By structuring evaluation through a facts $\rightarrow$ mental states $\rightarrow$ behavior reasoning process,

it enables a comprehensive assessment of social intelligence. Built upon the U-TRIMF framework, SocialArena covers diverse social scenarios with varied interaction dynamics. Through a human-based arena-style evaluation protocol, SocialArena yields high-quality and reliable evaluation results.

Limitations

We now discuss the limitations of SocialArena, which we anticipate addressing in future work.

Potential Bias. Human preferences in conversational AI are influenced by individual backgrounds, communication styles, and subjective expectations. Our dataset is currently annotated by native English and Chinese speakers, which may introduce certain biases in preference judgments. To mitigate this limitation, we have developed and plan to release in the near future a public-facing platform **SocialArena-Platform**, similar to Chatbot Arena (Zheng et al., 2023), that enables broader community participation in data collection and evaluation. Such an approach would allow us to capture more diverse perspectives and interaction patterns, ultimately leading to more robust and generalizable models in future iterations of this work. We will release updated results once sufficient data has been collected. More details are provided in § B.

Multi-turn Dialogue. We adopt single-turn evaluation as our preliminary small-scale annotations revealed that multi-turn settings incur higher annotation costs, while single-turn scenarios sufficiently address most evaluation needs. Once sufficient single-turn evaluation data has been collected, we will enable multi-turn functionality on the platform and gather data to reduce the associated costs.

Acknowledgments

This work was supported by Alibaba Research Intern Program and supported by Shenzhen Science and Technology Program (ZDCY20250901103533010).

References

2011. *Theory of Mind and Literature*, dgo - digital original edition. Purdue University Press.

Tawfig Khidir Ibnouf Adham. 2023. Conflict resolution in team: Analyzing the of conflicts and best skills for resolution. *Scholars Journal of Engineering and Technology*, 11(08):152–162.

Ian A Apperly. 2012. What is “theory of mind”? concepts, cognitive processes and individual differences. *Quarterly journal of experimental psychology*, 65(5):825–839.

Michael J Banissy, Ryota Kanai, Vincent Walsh, and Geraint Rees. 2012. Inter-individual differences in empathy are reflected in human brain structure. *Neuroimage*, 62(3):2034–2039.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. 2025. *tau²-bench*: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*.

Cindy Beaudoin, Elizabeth Leblanc, Charlotte Gagner, and Miriam H Beauchamp. 2020. Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, 10:2905.

Todd S Braver, Marie K Krug, Kimberly S Chiew, Wouter Kool, J Andrew Westbrook, Nathan J Clement, R Alison Adcock, Deanna M Barch, Matthew M Botvinick, Charles S Carver, and 1 others. 2014. Mechanisms of motivation–cognition interaction: challenges and opportunities. *Cognitive, affective, & behavioral neuroscience*, 14(2):443–472.

Brenda K Bryant. 1992. Conflict resolution strategies in relation to children’s peer relations. *Journal of Applied Developmental Psychology*, 13(1):35–50.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. 2024. Social-bench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2108–2126.

Michael Cole. 2003. Culture and cognitive science. *Outlines. Critical Practice Studies*, 5(1):3–15.

Katie Davis, Joanna Christodoulou, Scott Seider, and Howard Gardner. 2011. *The Theory of Multiple Intelligences*, page 485–503. Cambridge Handbooks in Psychology. Cambridge University Press.

Jean Decety. 2010. The neurodevelopment of empathy in humans. *Developmental neuroscience*, 32(4):257–267.

A.E. Elo. 1966. *The USCF Rating System: Its Development, Theory, and Applications*. United States Chess Federation.

Emily Falk and Christin Scholz. 2018. Persuasion, influence, and value: Perspectives from communication and social neuroscience. *Annual review of psychology*, 69(1):329–356.

Alan P Fiske. 1992. The four elementary forms of sociality: framework for a unified theory of social relations. *Psychological review*, 99(4).

- Marian Friestad and Peter Wright. 1994. The persuasion knowledge model: How people cope with persuasion attempts. *Journal of consumer research*, 21(1):1–31.
- Drew Fudenberg and Jean Tirole. 1991. *Game theory*. MIT press.
- Daniel Goleman. 1996. Emotional intelligence. why it can matter more than iq. *Learning*, 24(6):49–50.
- Ilaria Grazzani, Veronica Ornaghi, Elisabetta Conte, Alessandro Pepe, and Claudia Caprin. 2018. The relation between emotion understanding and theory of mind in children aged 3 to 8: The key role of language. *Frontiers in psychology*, 9:724.
- Yuling Gu, Oyvind Taffjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. 2024. Simpletom: Exposing the gap between explicit tom inference and implicit tom application in llms. *arXiv preprint arXiv:2410.13648*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Shihui Han. 2015. Understanding cultural differences in human behavior: a cultural neuroscience approach. *Current opinion in behavioral sciences*, 3:68–72.
- Ursula Hess and Shlomo Hareli. 2014. The role of social context for the interpretation of emotional facial expressions. In *Understanding facial expressions in communication: Cross-cultural and multidisciplinary perspectives*, pages 119–141. Springer.
- Richard C Hollinger and John P Clark. 1982. Formal and informal social controls of employee deviance. *Sociological quarterly*, 23(3):333–343.
- Shuai Huang, Wenxuan Zhao, and Jun Gao. 2025. Si-bench: Benchmarking social intelligence of large language models in human-to-human conversations. *arXiv preprint arXiv:2510.23182*.
- Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.
- Jürgen Klüver, Jörn Schmidt, and 1 others. 1999. Topology, metric and dynamics of social systems. *Journal of Artificial Societies and Social Simulation*, 2(3):7.
- Justin Laforest, Megan MacGillivray, and Melanie Y Lam. 2021. The influence of social context and social connection on visual perceptual processes. *Acta Psychologica*, 215:103270.
- Jonathan D Lane, Henry M Wellman, Sheryl L Olson, Jennifer LaBounty, and David CR Kerr. 2010. Theory of mind and emotion understanding predict moral development in early childhood. *British Journal of Developmental Psychology*, 28(4):871–889.
- Xiaozhe Li, Tianyi Lyu, Siyi Yang, Yuxi Gong, Yizhao Yang, Jinxuan Huang, Ligao Zhang, Zhuoyi Huang, and Qingwen Liu. 2025. Consintbench: Evaluating language models on real-world consumer intent understanding. *arXiv preprint arXiv:2510.13499*.
- Junru Lu, Jiazheng Li, Guodong Shen, Lin Gui, Siyu An, Yulan He, Di Yin, and Xing Sun. 2025. Rolemrc: A fine-grained composite benchmark for role-playing and instruction-following. *arXiv preprint arXiv:2502.11387*.
- Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. Gpteach: Interactive training with gpt-based students. In *Proceedings of the tenth acm conference on learning@ scale*, pages 226–236.
- Amy R Overton and Ann C Lowry. 2013. Conflict management: difficult conversations with difficult people. *Clinics in colon and rectal surgery*, 26(04):259–264.
- Samuel J. Paech. 2025. Eq-bench 3: Emotional intelligence benchmark. <https://github.com/EQ-bench/eqbench3>.
- 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.
- Project Gutenberg. Project Gutenberg. <https://www.gutenberg.org>.
- Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialliqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.
- Zhihong Shao, Yuxiang Luo, Chengda Lu, ZZ Ren, Jiewen Hu, Tian Ye, Zhibin Gou, Shitong Ma, and Xiaokang Zhang. 2025. Deepseekmath-v2: Towards self-verifiable mathematical reasoning. *arXiv preprint arXiv:2511.22570*.
- Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. 2025. Tomato: Verbalizing the mental states of role-playing llms for benchmarking theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1520–1528.
- Eleanor H. Simpson and Peter D. Balsam. 2016. *The Behavioral Neuroscience of Motivation: An Overview*

- of Concepts, Measures, and Translational Applications*, pages 1–12. Springer International Publishing, Cham.
- Qwen Team. 2025. *Qwen3 technical report*. Preprint, arXiv:2505.09388.
- Qwen Team. 2026. *Qwen3.5: Accelerating productivity with native multimodal agents*.
- E. L. Thorndike. 1920. Intelligence and its uses. *Harper's Magazine*, 140:227–235.
- Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, and 1 others. 2025. Coser: Coordinating llm-based persona simulation of established roles. *arXiv preprint arXiv:2502.09082*.
- Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, and 1 others. 2024. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1840–1873.
- Zongqi Wang, Rui Wang, Yuchuan Wu, Yiyao Yu, Pinyi Zhang, Shaoning Sun, Yujiu Yang, and Yongbin Li. 2026. Reward modeling from natural language human feedback. *arXiv preprint arXiv:2601.07349*.
- Wikipedia contributors. 2026. Maslow's hierarchy of needs. https://en.wikipedia.org/wiki/Maslow%27s_hierarchy_of_needs. Accessed: 2026-08-28.
- Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706.
- Hao Xiang, Tianyi Tang, Yang Su, Bowen Yu, An Yang, Fei Huang, Yichang Zhang, Yaojie Lu, Hongyu Lin, Xianpei Han, and 1 others. 2025. Rmtbench: Benchmarking llms through multi-turn user-centric role-playing. *arXiv preprint arXiv:2507.20352*.
- Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. *arXiv preprint arXiv:2402.06044*.
- Zhenhua Xu, Dongsheng Chen, Shuo Wang, Jian Li, Chengjie Wang, Meng Han, and Yabiao Wang. 2026. Adamarp: An adaptive multi-agent interaction framework for general immersive role-playing. *arXiv preprint arXiv:2601.11007*.
- Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, and 1 others. 2026. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*.
- Bang Zhang, Ruotian Ma, Qingxuan Jiang, Peisong Wang, Jiaqi Chen, Zheng Xie, Xingyu Chen, Yue Wang, Fanghua Ye, Jian Li, and 1 others. 2025. Sentient agent as a judge: Evaluating higher-order social cognition in large language models. *arXiv preprint arXiv:2505.02847*.
- Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, and 1 others. 2024. Esc-eval: Evaluating emotion support conversations in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15785–15810.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.
- Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, and 1 others. 2025. Characterbench: Benchmarking character customization of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26101–26110.
- Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haoifei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and 1 others. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

Appendix Contents

A Discussion	12
A.1 Impact Statement	12
B SocialArena-Platform	12
C Details of Benchmark Construction	13
C.1 U-TRIMF Framework	13
C.2 Explanation of Keyword-Guided Scenario Synthesis	13
C.3 Explanation of Quality Filtering .	14
C.3.1 Rule-Based Filters	14
C.3.2 Human Review	16
D Details of Evaluation Protocol	16
D.1 Annotation Guidelines	16
D.2 Annotator Recruitment and Qual- ity Control	16
D.3 Details of Theory of Mind (ToM) .	16
E Full Experimental Results	18
E.1 Full Results of 'A Deeper Under- standing of LLM Performance in Social Intelligence'	18
E.2 Inter-Annotator Agreement	18
E.3 Analysis of Typical Failure Modes in Arena-Style Benchmarks	18
F Prompts	19

A Discussion

A.1 Impact Statement

Positive Aspects. Our research will promote the development of humanitarian care in the LLM field. It can significantly advance research on reward modeling and RLHF in unverifiable domains. The academic community can leverage this to obtain high-quality data and develop more powerful character models with social intelligence.

Negative Aspects. The social scenarios on SocialArena may themselves contain unhealthy or unsafe content (e.g., illegal or unethical material), which could negatively impact academic research if used for model training or evaluation. To address this, our platform implements strict filtering mechanisms to exclude such content from all displayed scenarios and leaderboard updates.

B SocialArena-Platform

Overview. To facilitate continuous data collection and broader community engagement, we introduce **SocialArena-Platform**, a publicly accessible online platform designed to complement and extend the SocialArena benchmark.

Motivation. The motivation behind this platform is twofold. First, human preferences in social interactions are inherently diverse, and a static dataset annotated by a limited group of annotators may not fully capture this diversity. Second, real-world social dilemmas are dynamic and ever-evolving, necessitating a living platform that can continuously collect fresh scenarios and preference judgments beyond the scope of our initial benchmark.

Platform Design. The platform is centered around real-world social interaction. Users can post the social dilemmas they actually encounter in their daily lives, receive responses from state-of-the-art LLMs as well as fellow community members, and ultimately find resolutions to their challenges. Meanwhile, community upvotes and comments serve as a natural and scalable evaluation mechanism for assessing model capabilities in social intelligence.

User Engagement. To attract a broad and diverse user base, we focus on delivering immediate practical value to general users. The platform offers free access to multiple state-of-the-art LLMs, allowing users to directly experience and compare model responses to real social dilemmas at no cost. Moreover, the platform hosts a rich repository of diverse social scenarios along with corresponding model responses, which users can freely browse for reference and inspiration when navigating their own interpersonal challenges. We believe that this combination of free model access and rich reference content will naturally draw users who are genuinely interested in seeking advice and exploring how AI can assist in complex social situations.

Data Quality Control. To ensure the reliability of annotations collected in the wild, we adopt a multi-layered filtering strategy. First, we filter out prompts that do not correspond to social interaction tasks, such as coding or mathematical queries, to ensure that all collected scenarios fall within the intended domain. Second, we apply the same scenario qualification criteria introduced in Section 2.2 to remove low-quality or ill-formed sce-

narios that fail to meet our benchmark standards. Third, we adopt the anomalous IP detection rules from Chatbot Arena (Zheng et al., 2023) to identify and exclude votes originating from suspicious or duplicate sources, thereby safeguarding the integrity of the collected preference judgments.

Through these efforts, we hope that SocialArena-Platform can serve as a sustainable and community-driven resource that genuinely contributes to human well-being and advances the development of socially intelligent AI systems.

C Details of Benchmark Construction

C.1 U-TRIMF Framework

The U-TRIMF framework is a structured taxonomy we designed to systematically cover the space of basic social contexts. It decomposes a social scenario into three fundamental elements—**User Presence**, **Characters**, and **Events**—and characterizes each element along multiple theoretically grounded dimensions. Specifically, user presence is described by the **User Presence** dimension; characters are described by their **Topological Relationships** and **Relational Models**; events are characterized by their **Interest Structure**, **Motivational Types**, and **Formality** level. Together, these six dimensions give rise to the acronym **U-TRIMF**.

User Presence. The *User Presence* dimension captures whether a human user exists in the scenario, which fundamentally shapes the LLM’s role and responsibilities. In **U1 (Real-world)** scenarios, no user is present: the LLM acts as a character embedded in and governed by the real world (Zhang et al., 2025; Huang et al., 2025). In **U2 (User-centered)** scenarios, a user is present: the LLM is tasked with serving user-specific needs, and accordingly gains the authority to influence how the scenario evolves (Wang et al., 2025; Xiang et al., 2025).

Characters. The *Topological Relationship* dimension captures the structural configuration of participants in a social scene. We adopt two active categories: a symmetric dyad (T2) and a symmetric group (T3). Importantly, T1 (a single individual alone) is **excluded** from our benchmark construction, as a solitary person cannot constitute a social interaction by definition. All scenarios in our benchmark therefore involve at least two participants (T2 or T3). The *Relational Model* dimension, grounded in Fiske’s (1992) influential framework,

encodes the qualitative nature of the relationship between participants, ranging from intimate communal sharing (R1) to purely transactional market pricing (R4) and even null/asocial ties between strangers (R5).

Events. The *Interest Structure* dimension (I) captures the game-theoretic structure of the situation (Fudenberg and Tirole, 1991), covering zero-sum, negative-sum, and positive-sum interactions. The *Motivational Type* dimension (M) is adapted from Maslow’s hierarchy of needs (2026), spanning physiological needs (M1) through self-actualization (M5). Finally, the *Formality* dimension (F) distinguishes between informal, everyday interactions (F1) and formally structured, institutionalized settings (F2).

A complete description of all categories with illustrative examples is provided in Table 4.

C.2 Explanation of Keyword-Guided Scenario Synthesis

Why keyword guidance? A naive approach to scenario generation—prompting a language model directly with a U-TRIMF category combination—produces highly repetitive outputs, as large language models tend to converge on the most prototypical exemplar of any given category. To mitigate this degeneracy, we adopt a **keyword-guided synthesis** strategy: each generation call is conditioned on 10 thematic keywords that seed the narrative content after the U-TRIMF constraints are applied.

Keyword sourcing and design. Keywords are extracted from randomly sampled 10,000 chapters from 10,000 books in **Project Gutenberg (Project Gutenberg)** (with professional/technical books filtered out). Each collection of 10 keywords is drawn from a single chapter of a book to ensure that the 10 keywords remain thematically coherent rather than semantically disparate.

Keyword guidance does not introduce bias. We argue here that keyword guidance does not introduce bias into scenario synthesis. First, keywords serve only as soft guidance rather than hard constraints: the model is instructed to draw inspiration from them rather than incorporate all of them verbatim. Second, the pre-established U-TRIMF template acts as a structural scaffold that ensures comprehensive coverage of the scenario space, rather than the keywords doing so. Finally, all generated

Table 4: The U-TRIMF Framework for Base Scenario Construction.

Dimension	Category	Examples
<i>Basic Element 1: User Presence</i>		
User Presence	U1: Real-world	LLM acts as a character governed by real-world constraints; no user present
	U2: User-centered	LLM serves a user’s specific needs; user present and drives scenario evolution
<i>Basic Element 2: Characters</i>		
Topological Relation	T1: Individual T2: Dyad T3: Group	(Excluded: a solitary individual cannot constitute social interaction) Couple, doctor/patient, teacher/student, buyer/seller Team members, family members, book club, colleagues
Relational Model	R1: Communal Sharing R2: Authority Ranking R3: Equality Matching R4: Market Pricing R5: Null	Mother/child, romantic partners, parent/offspring CEO/employee, advisor/graduate student, officer/soldier Colleagues, classmates, neighbors, carpool partners, friends Consumer/merchant, landlord/tenant, employer/employee Strangers, passersby, pedestrians
<i>Basic Element 3: Events</i>		
Interest Structure	I1: Zero-Sum I2: Negative-Sum I3: Positive-Sum	Bidding for projects, election campaigns, inheritance disputes Price wars, divorce litigation, trade wars, factional conflicts Strategic partnerships, joint ventures, community building
Motivational Type	M1: Physiological M2: Safety M3: Belonging M4: Esteem M5: Self-actualization	Eating (hungry), drinking (thirsty), sleeping (tired) Buying insurance, saving money, installing security systems Family reunion, joining communities, team building Competing for awards, conference presentations Teaching volunteering, studying art, pursuing PhD
Formality	F1: Informal F2: Formal	Value exploration, stranger storytelling, airport deep talks Psychological therapy, estate distribution, board decisions

instances are manually verified to ensure overall data quality.

C.3 Explanation of Quality Filtering

Raw scenarios generated by the language model are passed through a two-stage quality filtering pipeline: (1) a set of explicit **rule-based filters** that automatically reject scenarios violating pre-specified constraints, implemented via Qwen3-235B-A22B-Instruct-2507, followed by (2) a **human review stage** in which annotators make holistic judgments about scenario suitability.

C.3.1 Rule-Based Filters

The rules fall into four thematic groups.

Group A: Content Safety & Ethical Boundaries.

These rules reject scenarios whose content would make them inappropriate or distressing to present to participants, regardless of their social interest.

- **A1 — No life-threatening content.** Scenarios must not involve serious illness, terminal diagnosis, suicide, life-threatening danger, or near-death experiences. *Exception:* a scenario in which one party provides positive emotional

support to the other (e.g., comforting someone anxious about a routine medical check-up) is permitted.

- **A2 — No serious criminal conduct.** Scenarios must not involve murder, sexual assault, drug trafficking, kidnapping, or comparable crimes.
- **A3 — No extreme trauma.** Scenarios must not depict childhood abuse, abandonment, or severe violence as central narrative elements.
- **A4 — No illegal events.** Scenarios must not involve legally prohibited acts such as marital infidelity that constitutes fraud, embezzlement, or financial misconduct.
- **A5 — No major betrayal.** Scenarios must not hinge on acts of serious interpersonal betrayal (e.g., informing on a close friend to authorities).
- **A6 — No irreversible catastrophic loss.** Scenarios must not involve bankruptcy, complete reputational destruction, or equivalent perma-

ment, irrecoverable harms as a primary plot element.

Group B: Scenario Complexity & Conflict Validity. These rules ensure that each scenario presents a genuinely non-trivial social situation.

- **B1 — No trivial everyday disputes.** Scenarios that lack any meaningful complexity—where any resolution proposed by either party would be equally acceptable at negligible cost—are excluded.
- **B2 — No extreme or forced dilemmas.** Scenarios must not place characters in all-or-nothing situations requiring one party to suffer a severe, unavoidable harm. The tension should be realistic and proportionate.
- **B3 — Single-event focus.** Each scenario must revolve around a single, coherent event or situation. Scenarios that superimpose multiple independent conflicts are excluded.
- **B4 — Objective-basis requirement for conflict.** If a scenario contains a conflict, it must arise from objective circumstances (e.g., a resource constraint, an impending deadline, a genuine incompatibility of needs) rather than from artificially or inexplicably manufactured antagonism.
- **B5 — Balanced conflict weight.** When a scenario features opposing interests, both sides must carry comparable moral, practical, or emotional weight. Asymmetric conflicts—where one side concerns safety or fundamental rights and the other concerns minor personal preferences—are excluded. Example of a rejected pair: *child safety on a balcony* vs. *aesthetic preference for an open railing*.

Group C: Structural & Narrative Requirements. These rules enforce properties of the scenario's construction that are necessary for its use in the benchmark.

- **C1 — Appropriate conversational entry point.** The opening dialogue included in each scenario must end at a moment that naturally invites a continuation—a point of genuine tension, an unanswered question, or a decision hanging in the air—so that a model assigned to role-play a character has meaningful space to respond.

- **C2 — Unbroken narrative logic.** The causal chain within the scenario must be internally coherent. Scenarios in which a character's reaction is logically incomprehensible given the stated trigger are rejected. *Example of a rejected scenario:* one character cuts their hair short; the other accuses them of “being the villain”—the logical connection is missing.

Group D: Conflict Substance Requirements. These rules guard against scenarios whose conflict is only superficially social.

- **D1 — Psychological or interest-based conflict required.** (*If this is a real social scenario; virtual/fictional scenarios are exempt from this requirement.*) The conflict must engage genuine psychological or interpersonal stakes—emotional needs, relationship maintenance, value disagreements, identity concerns, material interests, or the need to persuade another person. Purely technical or procedural trade-offs are excluded. *Example of a rejected scenario:* two colleagues debating which document structure to use for a report, or which software tool is more efficient.
- **D2 — Sufficient emotional weight.** (*If this is a real social scenario; virtual/fictional scenarios are exempt from this requirement.*) The conflict must touch something the characters genuinely care about: a significant relationship, a core value, a meaningful identity, or a decision with real consequences for their lives. Trivial, surface-level disagreements are excluded. *Examples of rejected scenarios:* neighbors casually discussing whether to remove an interior door; coworkers debating which restaurant to order lunch from.
- **D3 — Immediate action pressure required.** (*If this is a real social scenario; virtual/fictional scenarios are exempt from this requirement.*) The scenario must create a sense of urgency—someone is waiting for a response, a decision must be made on the spot, or a time window is about to close. Scenarios framed as open-ended reflection, long-term planning, or interruptible deliberation are excluded. *Examples of rejected scenarios:* a character pondering life goals after a routine health check-up; a character receiving a message and drifting into private rumination.

- **D4 — No “free exit” situations.** (*If this is a real social scenario; virtual/fictional scenarios are exempt from this requirement.*) Characters must have a genuine stake in the interaction such that walking away is not a trivially costless option. Scenarios in which a character has no real reason to remain engaged yet is portrayed as conflicted are rejected. *Example of a rejected scenario:* a tenant who has no contractual obligation to a property manager deliberating at length over whether to sign a discretionary form, when they can simply leave.

C.3.2 Human Review

Scenarios that pass all rule-based filters are subsequently reviewed by human annotators. Each annotator reads the full scenario—including the character descriptions, situational background, and opening dialogue—and makes a holistic judgment along the following dimensions:

1. **Overall plausibility.** If this is a real social scenario, does it feel like something that could realistically happen between real people? Does the emotional and behavioral logic of the characters hold up under scrutiny? If this is a virtual social scenario, is it actually something that users would use or experience?
2. **Balance and fairness of the conflict.** Would a reasonable person find both sides of the conflict understandable and legitimate, even if asymmetric in some minor respect?
3. **Role-play affordance.** Does the scenario provide a natural, productive entry point for a model to inhabit one of the characters? Is the situation open enough to allow multiple plausible responses, yet constrained enough to require genuine social reasoning?
4. **Freedom from subtle rule violations.** Does the scenario pass the spirit—not merely the letter—of the rules above? Annotators flag cases that technically satisfy each individual rule but collectively feel implausible, gratuitous, or inappropriate.

Scenarios are retained only if the annotator judges them as satisfactory on all four dimensions. Disagreements between annotators are resolved through discussion, and borderline cases are discarded to maintain a conservative quality threshold.

D Details of Evaluation Protocol

D.1 Annotation Guidelines

Annotators were provided with two forms of guidance prior to beginning the annotation task. First, they received a structured guidelines document covering the foundational rules of the annotation protocol, summarized in Table 5. Second, they were given a set of worked examples—including full scenario contexts, paired model responses, and expert-written reference annotations—to calibrate their understanding of each dimension and the ABTG label scheme before handling live data.

D.2 Annotator Recruitment and Quality Control

Annotations were recruited provided through third-party annotation vendors. The annotator pool consisted of a mix of native English speakers and native Chinese speakers in roughly equal proportion.

All annotators were informed about the nature of the data and the intended research use of their annotations through the vendor’s standard onboarding and consent procedures. No personally identifiable information about annotators was shared with the research team beyond what was necessary for managing the annotation workflow through the vendor.

To maintain annotation quality, we implemented a batch-level quality control process. For each annotation batch, every annotator had a small subset of their submitted items independently reviewed by internal expert reviewers. Annotators whose quality pass rate fell below 75% for a batch had their corresponding annotations removed from the dataset. Additionally, these annotators received targeted feedback and further training to help them improve their performance in subsequent batches.

D.3 Details of Theory of Mind (ToM)

We adopt the taxonomy of mental state categories from [Beaudoin et al. \(2020\)](#), which provides a comprehensive and theoretically grounded classification of the mental states implicated in Theory of Mind reasoning. Specifically, we operationalize ToM across the following seven categories, summarized in Table 6.

Table 5: Annotation guidelines provided to annotators.

Item	Guideline
Task Objective	Judge which of two model responses (A or B) better reflects social intelligence. Label using A (A better), B (B better), T (both good), or G (both poor).
Thinking Evaluation	Judge which response provides a better analysis of each mental state dimension. Evaluation principles: Correctness first : if a response’s thinking contains an objectively wrong inference about any mental state (intent, emotion, belief, etc.), it loses regardless of other qualities. Utility second : a correct analysis counts as better only if it meaningfully promotes a better behavior; a correct but redundant analysis is scored T.
Behavior Evaluation	In general, prefer responses that demonstrate higher social intelligence , such as being more emotionally intelligent, empathetic, persuasive, etc. However, there are some exceptions: a. Long-term interests take priority : If making a compromise would make oneself appear weak, leading to short-term problem resolution but long-term negative consequences, then compromising is not acceptable. b. Maintain moral boundaries : If there is a conflict between illegal behavior and personal interests, choosing personal interest is not acceptable. If the other party has committed a moral violation — such as cheating, stealing, etc. — be careful not to empathize with them in such cases. c. Role consistency : If a character is explicitly described as someone who is "psychologically traumatized," "extremely angry," "timid and fearful," or "very hesitant," care must be taken not to violate the character’s established traits . For example, if a character has psychological trauma, they should not immediately comfort themselves and recover quickly at the beginning.
Factual Grounding	Any content in a response that cannot be found in the scenario background—and cannot be naturally and reasonably inferred—counts as a factual error and penalizes that response.
Conflict Weight	When only one response covers a particular mental state dimension, it wins on that dimension only if covering it produces a meaningfully better behavior; otherwise score T. When neither response covers a dimension, leave that field blank.
Comment Requirements	For every dimension, identify the specific strengths and weaknesses of both A and B. Abstract descriptors (e.g., “more thorough”, “clearer”) are not permitted; comments should cite concrete content from the responses wherever possible.

Table 6: Theory of Mind mental state categories.

Mental State	Description
T1: Intentions	The specific, conscious plans an agent formulates to act on underlying desires and achieve a desired outcome. Reasoning about intentions involves inferring the deliberate actions an agent commits to in pursuit of goals.
T2: Emotions	Current emotional states, their roots, hidden affects, and potentially complex or contradictory feelings. Understanding emotions requires inferring how affective experiences shape reactions, judgments, and subsequent behavior.
T3: Desires	Deep motivational states representing what an agent emotionally craves, values, or strives for. These desires may be multiple or conflicting and often serve as the underlying drivers of intentions and actions.
T4: Percepts	Direct sensory access to the current situation and key information, such as what an agent can see, hear, or notice in the environment. This includes situational cues like physical positioning or shifts in tone and helps explain potential cognitive blind spots between agents.
T5: Knowledge	Epistemic states concerning confirmed facts versus acknowledged unknowns. This includes distinguishing between knowing and merely believing, as well as differentiating reliable knowledge gained through experience from second-hand or uncertain information.
T6: Beliefs	Subjective judgments or understandings about the situation, other agents, or relationships. Beliefs may be inaccurate or biased, often arising from incomplete information, assumptions, or prior experiences.
T7: Non-Literal Communication	The interpretation of meaning beyond literal language, including sarcasm, politeness strategies, deception, and other forms of social subtext. Understanding such communication requires inferring communicative intentions and the psychological motives underlying them.

E Full Experimental Results

E.1 Full Results of ‘A Deeper Understanding of LLM Performance in Social Intelligence’

Criteria Extraction and Annotation. We prompt an LLM (Figure 10) to extract fine-grained evaluation criteria—hereafter referred to as *checklist items*—from each human critique, and to record the performance of both candidate models (Model A and Model B) on each item. Each checklist item captures a specific evaluative dimension, which may reflect either a desirable or undesirable behavior; however, for the purpose of annotation, we adopt a binary indicator of *positive assessment*: a model receives a score of 1 on a given dimension if the human critique reflects a favorable judgment along that dimension, and 0 otherwise. This process yields a total of **51,203** dimension-level records.

Taxonomy Construction. To organize this large and heterogeneous set of criteria into an interpretable structure, we classify all checklist items into three high-level categories: *role-playing*, *internal state understanding*, and *decision-making*. Within each category, we compute sentence embeddings for all items and apply clustering to group semantically similar criteria. A large language model is then prompted to generate a concise summary label for each cluster, characterizing the shared evaluative focus of the grouped items. To obtain a compact yet representative set of dimensions, we apply a second stage of clustering across all initial clusters, ultimately consolidating the taxonomy into **100 representative evaluation dimensions**. Details are provided in Tab. 8.

E.2 Inter-Annotator Agreement

To assess the reliability of our preference annotations, we measured inter-annotator agreement over the four-class label space $\{A, B, T, G\}$. Each sample was independently annotated by three annotators. We report the distribution of agreement patterns and compare it against the theoretical random baseline, i.e., the expected pattern frequencies if three annotators were to guess uniformly at random among the four labels.

Let p_{full} , p_{maj} , and p_{dis} denote the probabilities of full agreement (all three annotators agree), majority agreement (two agree, one diverges), and full disagreement (all three diverge), respectively. Under

Table 7: Inter-annotator agreement patterns compared to the theoretical random baseline.

Agreement Pattern	Random Baseline	Our Annotation
Full agreement (3/3 agree)	6.25%	39.18%
Majority agreement (2/3 agree)	56.25%	52.54%
Full disagreement (0/3 agree)	37.50%	8.29%

the random-guessing assumption, these probabilities are given by:

$$p_{\text{full}} = \binom{4}{1} \left(\frac{1}{4}\right)^3 = 6.25\%, \quad (1)$$

$$p_{\text{maj}} = \binom{4}{1} \binom{3}{2} \left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right) = 56.25\%, \quad (2)$$

$$p_{\text{dis}} = \binom{4}{3} \cdot 3! \cdot \left(\frac{1}{4}\right)^3 = 37.50\%. \quad (3)$$

Table 7 summarizes the observed agreement patterns in our annotation compared to the random baseline.

As shown in Table 7, the observed full-agreement rate (39.18%) is more than $6\times$ higher than the random baseline (6.25%), while the full-disagreement rate (8.29%) is more than $4\times$ lower than chance (37.50%). These results indicate that our annotators converge on a shared preference substantially more often than would be expected by random guessing, demonstrating a high level of annotation consistency.

E.3 Analysis of Typical Failure Modes in Arena-Style Benchmarks

We further analyze the typical failure modes identified in arena-style evaluation frameworks (e.g., LMArena) and assess their impact on our benchmark. Several design choices in our setting substantially mitigate these biases. First, role-playing outputs tend to be short, which helps annotators focus on content quality rather than surface-level style. Second, unlike LMArena, our evaluation adopts detailed criteria beyond a single preference judgment, inherently weakening stylistic shortcuts. Third, annotators are required to provide justifications for every preference, making each decision evidence-based and interpretable.

We additionally quantify the biases discussed in prior work:

- **Length bias.** Winners are longer than losers in only 51.05% of comparisons, with a mean length difference of 14.0 tokens and a median difference of 3.0 tokens. This is close

to chance level, confirming negligible length bias.

- **Position bias.** Votes favor option B over option A (43.63% vs. 39.23%), indicating only a mild position effect.
- **Format, tone, and confidence bias.** Our task does not involve rich formatting or answer confidence, so these biases are inapplicable by design.

These results demonstrate that typical biases are either absent or well-controlled in our setting, verifying that our task design effectively addresses the concerns raised for arena-style evaluation.

F Prompts

Prompt: Real Social Contexts Generation

Please generate a role-playing card (including the current scene, character information, and opening dialogue).

Core Objective: We need the characters within the scene to demonstrate a certain degree of psychological depth; however, it should be neither an everyday mundane scenario nor an extremely dramatized plot.

1. Define the Basic Scene Framework

Below I will provide a multi-dimensional scale analysis of the basic scene framework, with each dimension represented by a capital letter.

Basic Scene Framework Reference: “**## T: Topological Relationship** (Graph Theory) **T1: Self-Reflexive - One person alone** Example: individual **T2: Symmetric Dyad - Two people interacting** Example: couple, doctor/patient, teacher/student, buyer/seller, counselor/client **T3: Symmetric Group - Three or more people interacting** Example: team members, family members, book club members, department colleagues, friend groups

R: Relational Models Theory (Alan Fiske) **R1: Communal Sharing** Example: mother/child, couple, children/parents, roommates, family members **R2: Authority Ranking** Example: CEO/employee, supervisor/graduate student, officer/soldier, judge/party, editor-in-chief/author **R3: Equality Matching** Example: colleagues, classmates, neighbors, carpooling partners, friends **R4: Market Pricing** Example: consumer/merchant, client/service provider, landlord/tenant, employer/employee, investor/investment target **R5: Null/Asocial** Example: strangers, passersby, pedestrians

S: Interest Relationship (Game Theory) **S1: Zero-Sum - One party gains = the other party loses** Example: competitive bidding, election for a seat, inheritance dispute, competition for championship, competing for promotion **S2: Negative-Sum - Conflict leads to losses for both parties** Example: price wars, divorce litigation, trade wars, factional struggles, inter-departmental blame-shifting **S3: Positive-Sum - Cooperation creates value** Example: university-enterprise cooperation, strategic partners, joint investment, startup partnership, community co-building

M: Motivation Type (Maslow's Hierarchy of Needs) **M1: Physiological/Survival Needs** Example: eating when hungry, cooling down in heat, drinking when thirsty, sleeping when tired, seeking medical care when ill **M2: Safety/Stability Needs** Example: purchasing insurance,

saving money, signing contracts, installing security systems, retirement planning **M3: Belonging/Relationship Needs** Example: going home to be with family, joining a community, confiding after a breakup, team-building activities, looking for fellow townspeople **M4: Esteem/Status Needs** Example: pursuing awards, speaking at meetings, sharing achievements, running for positions, obtaining certifications **M5: Values/Self-Actualization Needs** Example: teaching in underprivileged areas, learning art, environmental causes, writing memoirs, pursuing a doctorate

P: Whether It Is a Formal Setting **P1: Informal** Example: discussion of values, value conflicts, late-night heart-to-heart talks, deep conversations while waiting for a flight, strangers sharing stories **P2: Formal** Example: psychological healing (important scene), handling conflicts at a dinner party, estate distribution, board decision-making, divorce mediation, merger negotiation ““

2. Progressively Generate a Specific Scene

2.1 Plot Overview

Generate a plot whose **complexity comes from authentic sources**, not from artificially contrived setups.

2.2 Current Scene

Describe the background of the current dialogue in a **colloquial, straightforward** manner.

Required elements: 1. **Specific background (if applicable):** What is the specific relationship between the characters, and what are the key preceding events? 2. **Current event:** Who is talking to whom, and what specific matter are they discussing? 3. **Source of complexity:** This must be included to prevent readers from being unable to understand where the dilemma comes from.

Writing requirements: - Use short sentences and write in separate paragraphs - Describe only the core events - Do not describe professional/specialized content, e.g.: "The queen has passed away" - Do not use extremely emotionally charged words: tearing apart, struggling, breakdown - Do not use extremely dramatized literary expressions: fate, reunion, redemption, entanglement

2.3 Opening Dialogue

Provide natural dialogue that fits the [current scene]. The last line should end at an appropriate point, leaving room for the model to take on a role and improvise.

Requirements: - Dialogue should feel everyday and natural, as if real people are speaking - Avoid overly sentimental, dramatized, or sensationalized lines - Reflect the complexity of the scene while keeping the tone restrained

2.4 Character Information

List the names of the characters in the dialogue (use English names).

3. Output Format

““ <Keyword Selection> [Selected keywords — do not select all of them] </Keyword Selection>

<Thinking> [You may think at length here in order to develop a better plot] </Thinking>

<Plot Overview> [Describe the plot in one sentence, within 30 characters] </Plot Overview>

<Current Scene> [Place the current scene here] [Use the characters' names, write in separate paragraphs, 1–2 sentences per paragraph] </Current Scene>

<Opening Dialogue> 1. Character A: (brief action) Dialogue content 2. Character B: (brief action) Dialogue content [3–7 rounds of dialogue] </Opening Dialogue>

<Character Information> [Character A's name, Character B's name] </Character Information>

<Next Speaker> [Write the name of the next speaker here — only one name] </Next Speaker> ““

4. Quality Check

In this section, you should go through the checklist below

item by item and verify whether your generated content meets each requirement. For each item, first provide your reasoning, then state your conclusion. If any issue is found, re-output using the format from Section 2.5, and check again.

1. No life-or-death related plots are allowed: e.g., serious illness, terminal illness, suicide, life-threatening danger, near-death (acceptable if one party is responsible for positively regulating the other's emotions) 2. No serious criminal plots are allowed: e.g., murder, sexual assault, drug trafficking, kidnapping 3. No extreme trauma plots are allowed: e.g., childhood abuse, abandonment, experiencing violence 4. No illegal events are allowed: e.g., extramarital affairs, embezzlement of public funds 5. No major betrayals are allowed: e.g., betraying a friend 6. No irreversible losses are allowed: e.g., bankruptcy, complete loss of reputation 7. No surreal settings are allowed: e.g., dreamscapes, time travel, paranormal events, fateful entanglements

8. The scenario must not consist of everyday trivial matters with no complexity, solvable by any approach either party might take 9. No extreme, hysterical, or forced-choice situations where a character must suffer significant harm are allowed

10. Only a single event should be the focus — do not layer multiple scenes 11. If there is conflict, it must arise from objective reality, not be artificially manufactured 12. If your scene involves contradiction/conflict, ensure both parties carry comparable weight — examples of imbalance: safety/health, legal/moral baseline vs. personal interest; long-term major impact vs. short-term minor inconvenience

13. The last line of the opening dialogue must end at a point that sets an appropriate atmosphere, leaving room for the model to take on a role and improvise

5. Input

For this generation, you must use **{base_scene}** as the basic scene framework combination.

You may refer to these keywords when generating the scene, but note that their use is not mandatory. Keywords: **{keywords}**

Please begin generating.

Figure 6: Prompt: Real Social Contexts Generation.

Prompt: User-Centered Social Contexts Generation.

Please generate a role-playing card (including the current scene, character information, and opening dialogue).

Core Objective: We need to build scenes oriented toward **entertainment** for the user. For example: virtual worlds (plot-driven, such as isekai adventures, virtual exploration, detective worlds), or real-world settings (entertainment-oriented, such as a palace maid falling in love with a crown prince, or a domineering CEO and their secretary).

1. Define the Basic Scene Framework

Below I will provide a multi-dimensional scale analysis of the basic scene framework, with each dimension represented by a capital letter.

Basic Scene Framework Reference: ““ **## T: Topological Relationship (Graph Theory) T1: Self-Reflexive - One person alone** Example: individual **T2: Symmetric Dyad - Two people interacting** Example: couple, doctor/patient, teacher/student, buyer/seller, counselor/client **T3: Symmetric Group - Three or more people interacting** Example: team members, family members, book club

members, department colleagues, friend groups

R: Relational Models Theory (Alan Fiske) R1: Communal Sharing Example: mother/child, couple, children/parents, roommates, family members **R2: Authority Ranking** Example: CEO/employee, supervisor/graduate student, officer/soldier, judge/party, editor-in-chief/author **R3: Equality Matching** Example: colleagues, classmates, neighbors, carpooling partners, friends **R4: Market Pricing** Example: consumer/merchant, client/service provider, landlord/tenant, employer/employee, investor/investment target **R5: Null/Asocial** Example: strangers, passersby, pedestrians

S: Interest Relationship (Game Theory) S1: Zero-Sum - One party gains = the other party loses Example: competitive bidding, election for a seat, inheritance dispute, competition for championship, competing for promotion **S2: Negative-Sum - Conflict leads to losses for both parties** Example: price wars, divorce litigation, trade wars, factional struggles, inter-departmental blame-shifting **S3: Positive-Sum - Cooperation creates value** Example: university-enterprise cooperation, strategic partners, joint investment, startup partnership, community co-building

M: Motivation Type (Maslow's Hierarchy of Needs) M1: Physiological/Survival Needs Example: eating when hungry, cooling down in heat, drinking when thirsty, sleeping when tired, seeking medical care when ill **M2: Safety/Stability Needs** Example: purchasing insurance, saving money, signing contracts, installing security systems, retirement planning **M3: Belonging/Relationship Needs** Example: going home to be with family, joining a community, confiding after a breakup, team-building activities, looking for fellow townspeople **M4: Esteem/Status Needs** Example: pursuing awards, speaking at meetings, sharing achievements, running for positions, obtaining certifications **M5: Values/Self-Actualization Needs** Example: teaching in underprivileged areas, learning art, environmental causes, writing memoirs, pursuing a doctorate

P: Whether It Is a Formal Setting P1: Informal Example: discussion of values, value conflicts, late-night heart-to-heart talks, deep conversations while waiting for a flight, strangers sharing stories **P2: Formal** Example: psychological healing (important scene), handling conflicts at a dinner party, estate distribution, board decision-making, divorce mediation, merger negotiation ““

2. Progressively Generate a Specific Scene

2.1 Plot Overview

Write an overview of the plot to set the direction for the generation that follows.

2.2 Current Scene

Describe the background of the current dialogue in a **colloquial, straightforward** manner.

Required elements: 1. **Specific background (if applicable):** What is the specific relationship between the characters, and what are the key preceding events? 2. **Current event:** Who is talking to whom, and what are they discussing?

Writing requirements: - Use short sentences and write in separate paragraphs - Describe only the core events - Do not describe professional/specialized content, e.g.: "The queen has passed away"

2.3 Opening Dialogue

Provide natural dialogue that fits the [current scene]. The last line should end at an appropriate point, leaving room for the model to take on a role and improvise.

Requirements: - Avoid overly sentimental lines; prioritize a light and humorous tone

2.4 Character Information

List the names of the characters in the dialogue (use En-

glish names).

3. Output Format

““ <Keyword Selection> [Selected keywords — do not select all of them] </Keyword Selection>

<Thinking> [You may think at length here in order to develop a better plot; prioritize user-centered entertainment value] </Thinking>

<Plot Overview> [Describe the plot in one sentence, within 30 characters] </Plot Overview>

<Current Scene> [Place the current scene here] [Use the characters' names, write in separate paragraphs, 1–2 sentences per paragraph] </Current Scene>

<Opening Dialogue> 1. Character A: (brief action) Dialogue content 2. Character B: (brief action) Dialogue content [3–7 rounds of dialogue] </Opening Dialogue>

<Character Information> 1. World background (write this only if it is an isekai/fantasy adventure; not needed for real-world settings) 2. Description of the role the user plays 3. Description of the role the AI plays - The model must be clearly told that it is an AI; note that this must be framed through a user-facing relationship (for example, in a romance scenario set in the real world, the model can be told it is an AI that is simply playing the role of a girlfriend, and that it has the ability to influence how the world changes and control how the plot develops) - The model must be clearly told that it should center the user, and that it can influence how the world changes and control how the plot develops </Character Information>

<Next Speaker> [The next speaker is the AI — write their name here in the format 'Name (AI)'] </Next Speaker> ““

4. Quality Check

In this section, you should go through the checklist below item by item and verify whether your generated content meets each requirement. For each item, first provide your reasoning, then state your conclusion. If any issue is found, re-output using the format from Section 2.5, and check again.

1. No life-or-death related plots are allowed: e.g., serious illness, terminal illness, suicide, life-threatening danger, near-death (acceptable if one party is responsible for positively regulating the other's emotions) 2. No extreme or hysterical situations where a character would suffer significant harm are allowed

5. Input

For this generation, you must use {base_scene} as the basic scene framework combination.

You may refer to these keywords when generating the scene, but note that their use is not mandatory. Keywords: {keywords}

Please begin generating.

Figure 7: Prompt: User-Centered Social Contexts Generation.

Prompt: Conversation Generation.

system You are a Roleplaying Expert with Social Intelligence. Your task is to play the specified character in the "Current Scenario".

Behavioral Guidelines for Roleplay

****Core Objective**:** Starting from objective facts, you need to infer Theory of Mind (understanding the various mental states of yourself and others), and use this as a driver to generate an appropriate response.

Roleplay objectives

1. Always respond from a first-person perspective; do not mention that you are an AI or that you are roleplaying;

2. Dialogue format: responses should reflect the character's linguistic style and emotional state, kept concise (1–3 sentences); 3. Action format: if there are actions, annotate them with parentheses, e.g.: (looks down staring at the glass); 4. Keep dialogue authentic: avoid literary expressions and avoid fabricating things out of thin air.

Character Thinking

Before producing the final output, you need to engage in ****structured cognitive reasoning and decision reasoning**** from the perspective of the character you are playing, speaking in the first person ("I"). ****Wrap this inner thinking between '<think_character >' and '</think_character >' tags. ****

Explanation: Cognitive reasoning: Based on objective facts, infer the mental states of yourself and others (across the dimensions listed below). Decision reasoning: Based on the objective facts and results of cognitive reasoning, infer how you should respond to the other party.

Definitions of Mental State Dimensions

When reasoning, you need to ****adaptively select**** the 1–4 most critical dimensions for the current situation from the following mental state dimensions:

- ****T1: Intention****: The specific, conscious plans you formulate in order to fulfill desires and achieve expected outcomes. - ****T2: Emotion****: Understanding current emotional states, their underlying sources, hidden feelings, and complex or conflicting sentiments, as well as expressing or regulating one's own and others' emotions. - ****T3: Desire****: Refers to the wishes, preferences, and deep longings one ultimately hopes to fulfill — encompassing both explicit and implicit needs; also refers to perceiving what oneself and others truly want, and identifying tensions, conflicts, or synergies among multiple desires. - ****T4: Perception****: Refers to the direct awareness of a current situation and access to key information; this also includes recognizing that what others see, hear, and know is constrained by their physical position and sensory experience, thereby inferring their cognitive blind spots or informational limitations. - ****T5: Knowledge****: Refers to the confirmation of known facts, awareness of unknown domains, and a clear distinction between "knowing" and "believing," as well as between "hearsay" and "firsthand experience." - ****T6: Belief****: Involves personal subjective judgments and interpretations of situations, events, or interpersonal relationships; includes inferring the beliefs of others, and identifying belief distortions formed through one-sided information, inherent biases, or mistaken cognition. - ****T7: Non-Literal Communication****: Refers to discerning implied meanings, social strategies, and emotional intent; the ability to recognize the psychological motivations and communicative purposes hidden within non-literal expressions such as sarcasm, pleasantries, lies, humor, or slips of the tongue.

Thinking Process

1. Thinking about Current Mental States Understand your own and others' mental states in the current situation: (1) Select 2–4 of the most relevant dimensions; (2) You may respond directly, or combine the context with reasoning of any complexity.

2. Thinking about Behavior You need to execute the following steps: (1) Describe your behavior: (1.1) You may state your response directly; (1.2) Or you may reason out your response based on the mental state analysis above; if doing so, use the format Dimension : Analysis to indicate which mental states you are drawing on and how. (2) (Optional) Validate your response: (2.1) You may choose not to validate the response; (2.2) Or you may analyze which already-analyzed dimensions would undergo what changes — or what important new dimensions would be

introduced — if you gave that response, thereby validating whether your response is reasonable; if doing so, use the format `Dimension : Analysis` to indicate which mental states you analyzed.

3. **Meta-Thinking** If you find that the behavior you previously conceived is not ideal, you may iterate through [Thinking about Behavior] multiple times to arrive at a better behavior.

Notes

1. ****Adaptive dimension selection****: Do not mechanically cover all dimensions — only analyze the key dimensions that truly drive decisions in the current situation. 2. ****Maintain character perspective****: Always think from the character's point of view using "I," reflecting the character's cognitive limitations and personality traits. 3. ****Structured output****: Strictly follow the format above for output, to facilitate extraction and analysis. 4. ****Goal-oriented****: All thinking should serve the final behavioral decision.

Output Format

****Reply strictly using the following format****:

“ [Name of the speaking character this turn]:

<think_ccharacter >

[Thinking about Mental States]— <

DimensionName1 >:< Specificanalysis > - <

DimensionName2 >:< Specificanalysis > ...

[Thinking about Behavior] <Content> (If there is meta-

thinking, simply output [Thinking about Behavior] multi-

ple times) </think_ccharacter >

(Place the final response here) “

—

user_temulateCurrentScenario

scene

—

****Answer using the language in "Current Scenario".****

Now begin the roleplay — please continue the dialogue.

Figure 8: Prompt: Conversation Generation.

Prompt: Conversation Generation.

You are a professional response quality assessment expert. Your task is to compare the quality of two responses based on background information (including conversation history) and determine which one is better.

****Important: Only evaluate the final output content of each response, intermediate thinking processes can be used as reference.****

Criteria for Judging "Better"

We place the model in a social scenario and have it play a role to take actions and give replies. How to judge "better" needs to be analyzed in conjunction with the specific social context.

> ****Example Scenario: An Argument Between a Couple**** > > Suppose the dialogue background is a couple in a heated argument. Please note the following layers of requirements — they may be in tension with each other, and a good response needs to strike a balance among them: > >

1. ****Empathy and De-escalation****: The role being played should show empathy toward the other party's feelings, using gentle and appropriate language to ease the conflict rather than adding fuel to the fire. > > 2. ****Psychological Credibility of Emotional Transitions****: However, if the character is currently in a state of extreme anger, they should not suddenly become calm and rational. Emotional changes must be gradual and consistent with psychological norms — for example, shifting from anger to exhaustion

or hurt first, then slowly moving toward calm, rather than switching from rage to tenderness in a single sentence. > > 3. ****Moral Bottom Line****: If the argument was caused by one party cheating, the character should not use empathetic language to comfort or rationalize the cheater's behavior. Empathy should be directed toward the injured party, not used to excuse the wrongdoer.

Special Note: Role-play evaluation in social scenarios is very complex and nuanced, and is not limited to the three situations described above. Please analyze based on the specific social context.

—

Background Information and Conversation History

<background> BACKGROUND_INFORMATION <

/background >

Response A

<response_a > RESPONSE_A </response_a >

Response B

<response_b > RESPONSE_B </response_b >

—

Output Format

You may engage in any form of thinking, but ultimately

you need to organize the results into the following format:

“ <winner> [Winner. The result can only be one of four

options: A, B, T, or G, representing: Response A is bet-

ter, Response B is better, Both responses are good, Both

responses are poor] </winner>

<critique> [Commentary. Write multiple lines of commen-

tary here, each pointing out one concise argument. Note

that this is the condensed version — each line should be

a short sentence. Lengthy elaborations are prohibited.]

</critique> “

Figure 9: Prompt: Conversation Generation.

Prompt: Dataset Processing for Analysis.

You are a text analysis assistant. Please read the following evaluation conclusions about two responses (A and B), extract the key ****evaluation dimensions****, and determine whether each dimension is positive or negative for A or B. What are evaluation dimensions Evaluation dimensions are "judgment angles that can be scored independently," such as "psychological depth," "dialogue logic," and "initiative." Note: - Do not extract specific behaviors (e.g., "requesting receipts," "moving in together") - Do not extract evaluation results (e.g., "slightly rough," "superior") - Do not extract plot details - Extract abstract, reusable scoring dimensions in the form of indicators, structured as two-dimensional, preferably using "whether"

Rules for Positive/Negative Judgment - 1 = Positive (performs well on this dimension) - 0 = Negative (performs poorly on this dimension) - Only label the party ****explicitly mentioned**** in the text; do not infer or complete the other party - Example: Only "A is better at psychological game" → only write "A": 1, do not write "B": 0 - Example: Only "A's dialogue logic has a flaw" → only write "A": 0, do not write "B": 1 - Example: Both "A has good psychological game, B's is rough" → write "A": 1, "B": 0

Output Requirements 1. Each dimension label should be 2–10 words 2. Output in Python List of Dict format, each dict containing: - "dim": dimension name (string) - "A": 1

or 0 (if the text explicitly mentions A's performance on this dimension) - "B": 1 or 0 (if the text explicitly mentions B's performance on this dimension) 3. Output the conclusion directly, with no additional content

The evaluation conclusion is as follows: conclusion

Example input: - Response A is superior in terms of psychological game, accurately capturing the opponent's weaknesses and formulating in-depth response strategies, but there are obvious flaws in the logical common sense of specific dialogue lines, which undermines the authenticity of the interaction. - Response B's psychological analysis is slightly rough in depth, but its behavioral decisions closely follow the core facts, showing stronger initiative and better meeting the needs of plot progression. - Overall, B is slightly better, because B directly achieves the character's core goal, while A's dialogue logic hinders plot progression.

Example output: ["dim": "Whether psychological game is good", "A": 1, "B": 0, "dim": "Whether able to capture opponent's weaknesses", "A": 1, "dim": "Whether able to formulate in-depth response strategies", "A": 1, "dim": "Whether logical common sense is correct", "A": 0, "dim": "Whether interaction feels authentic", "A": 0, "dim": "Whether initiative is strong", "B": 1, "dim": "Whether plot progression is good", "A": 0, "B": 1, "dim": "Whether goal is achieved", "B": 1]

Figure 10: Prompt: Dataset Processing for Analysis.

Table 8: Overview of the 100 evaluation dimensions and their category assignments. ✓ indicates the dimension belongs to the category; ✗ indicates it does not.

Dimension	Role-playing	ToM Reasoning	Behavior Reasoning
Role-playing			
Whether behavior and speech are consistent and logically coherent in real situations	✓	✗	✗
Whether the role is faithfully portrayed and the narrative naturally advanced	✓	✗	✗
Whether the portrayal comes across as unnatural, unprofessional, or lacking credibility	✓	✗	✗
Whether characters and the self are presented authentically, naturally, and compellingly	✓	✗	✗
Whether self-awareness, emotional expression, and thought patterns show rigidity, avoidance, or imbalance	✓	✗	✗
Whether expression exhibits logical confusion, identity conflation, or communicative missteps	✓	✗	✗
Whether the self is expressed genuinely and emotional flow is sustained across interactions	✓	✗	✗
Whether problem-solving maintains internal consistency, pragmatic feasibility, and focus on essentials	✓	✗	✗
Whether professional interactions naturally balance expertise, accountability, and approachability	✓	✗	✗
Whether emotions are expressed and regulated authentically, subtly, and with restraint	✓	✗	✗
Whether one holds one's ground, takes responsibility, and upholds academic integrity in formal settings	✓	✗	✗
Whether interactions stay grounded in facts while avoiding fabrication	✓	✗	✗
Whether the balance between ideals and reality is well-judged and earns recognition	✓	✗	✗
Whether self-expression and consideration for others' feelings are balanced in social settings	✓	✗	✗
Whether dignity, boundaries, and interpersonal propriety are maintained across multiple relationships	✓	✗	✗
Whether the narrative is naturally advanced and persona appeal sustained in a light-hearted tone	✓	✗	✗
Whether self-contradictions and relational fractures are continuously exposed and exacerbated during interaction	✓	✗	✗
Whether boundaries, core values, and commitments are upheld across varied relationships and roles	✓	✗	✗
Whether one's beliefs are continuously examined, adjusted, and clearly expressed across situations	✓	✗	✗
Theory-of-Mind (ToM) Reasoning			
Whether the deeper intentions, unspoken rules, and psychological dynamics in interpersonal interactions are accurately perceived	✗	✓	✗
Whether one genuinely attends to and appropriately responds to the emotional needs of others	✗	✓	✗
Whether the grievances, fears, and deeper needs of the other party are truly understood and addressed	✗	✓	✗
Whether various critical signals and underlying issues in interactions are accurately identified	✗	✓	✗
Whether others' emotions and needs are genuinely understood and responded to	✗	✓	✗

Continued on next page...

(Continued from previous page)

Dimension	Role-playing	ToM Reasoning	Behavior Reasoning
Whether multi-dimensional information in social contexts is sensitively perceived and accurately responded to	✗	✓	✗
Whether one can truly read what the other party is thinking	✗	✓	✗
Whether one's understanding of others' psychology and the essence of relationships is sufficiently deep	✗	✓	✗
Whether the other party's emotions are genuinely received and steadied within the relationship	✗	✓	✗
Whether hearts are genuinely soothed and emotional or relational rifts resolved	✗	✓	✗
Whether the other party's intent and the essence of their question are accurately understood	✗	✓	✗
Whether the other person is truly kept in mind and thoughtfully looked after	✗	✓	✗
Whether balance between reason and warmth is maintained and others' needs are responded to sincerely	✗	✓	✗
Whether others' emotions and circumstances are understood and responded to with authenticity and nuance	✗	✓	✗
Whether mature and appropriate social competence and the ability to build relationships are demonstrated	✗	✓	✗
Whether facts are deeply understood and rationally expressed based on evidence	✗	✓	✗
Whether cognition is continuously recalibrated during social interactions to respond precisely to others' deeper needs	✗	✓	✗
Whether analyses of others' psychology are both incisive and grounded in reality and proportion	✗	✓	✗
Whether overt and covert signals in social contexts are sensitively perceived and responded to accurately	✗	✓	✗
Whether attention is paid to fine details of others' feelings and their latent needs	✗	✓	✗
Whether others' feelings and circumstances are taken into account during interaction	✗	✓	✗
Behavior Reasoning			
Whether rational and constructive thinking and decisive action are demonstrated	✗	✗	✓
Whether win-win outcomes, risks, and long-term goals are balanced under real-world constraints	✗	✗	✓
Whether things are moved forward to concrete outcomes in a pragmatic and efficient manner	✗	✗	✓
Whether strategies are flexibly adjusted in dialogue to drive effective and in-depth communication	✗	✗	✓
Whether principles are upheld while flexibly accommodating all parties' practical realities and relationships	✗	✗	✓
Whether defensive reactions are triggered, relationships are harmed, and trust is undermined in communication	✗	✗	✓
Whether effectiveness, proportion, and ethical boundaries are maintained when applying pressure	✗	✗	✓
Whether actual problems and core concerns are genuinely resolved	✗	✗	✓
Whether rules, boundaries, and professional integrity commitments are upheld	✗	✗	✓
Whether de-escalation, ice-breaking, and event progression are advanced in conflict	✗	✗	✓
Whether collaboration is pursued from a co-building and win-win starting point	✗	✗	✓

Continued on next page...

(Continued from previous page)

Dimension	Role-playing	ToM Reasoning	Behavior Reasoning
Whether both reducing opposition and driving genuine resolution are balanced in conflict	✗	✗	✓
Whether goal achievement and relationship maintenance are balanced in collaboration	✗	✗	✓
Whether logic is clear, expression precise, professionally reliable, thorough, and compelling	✗	✗	✓
Whether decisions are made boldly, responsibility is accepted, deadlocks are broken, and adaptation occurs in complex situations	✗	✗	✓
Whether speech is comfortable, natural, direct, and non-hurtful	✗	✗	✓
Whether the balance among all parties' interests and rule constraints is handled carefully and comprehensively	✗	✗	✓
Whether various contradictions and conflicts are converted into constructive action	✗	✗	✓
Whether specific, feasible, principled, and flexible action plans are proposed	✗	✗	✓
Whether mature, steady, strategic thinking and sincerity are demonstrated in social settings	✗	✗	✓
Whether appropriate judgment and mature accountability are shown in complex interpersonal situations	✗	✗	✓
Whether clear strategic awareness and a rational stance are maintained in interpersonal interactions	✗	✗	✓
Whether sincerity and candor are upheld in expression while also accounting for reality and goodwill	✗	✗	✓
Whether issues are handled in a professional and rigorous manner	✗	✗	✓
Whether emotional, rational, and long-term win-win considerations are balanced in conflict	✗	✗	✓
Whether relationship repair and closure are continuously driven in conflict and collaboration	✗	✗	✓
Whether strategies and details are flexibly applied to address real-world constraints	✗	✗	✓
Whether moral proportion and practical effectiveness are balanced in altruistic actions	✗	✗	✓
Whether various interpersonal relationships are thoughtfully cultivated and maintained	✗	✗	✓
Whether pressure, manipulation, or concession-based strategies are adopted in interpersonal interactions	✗	✗	✓
Whether both parties' interests are considered in cooperation and incentive and empowerment tools are flexibly used	✗	✗	✓
Whether genuine empathy or formulaic probing and scripted packaging are reflected in dialogue	✗	✗	✓
Whether mature, pragmatic, and far-sighted strategic thinking and action capacity are demonstrated	✗	✗	✓
Whether the dignity, face, and trust baseline of all parties are respected in interpersonal interactions	✗	✗	✓
Whether various forms of pressure are applied in response to problems	✗	✗	✓
Whether problems are genuinely solved in a pragmatic and effective manner	✗	✗	✓
Whether sincerity, gentleness, and constructiveness are maintained when expressing one's position	✗	✗	✓
Whether principles are adhered to while practical effectiveness is preserved in responses	✗	✗	✓
Whether concrete, actionable plans are used to move things forward	✗	✗	✓

Continued on next page...

(Continued from previous page)

Dimension	Role-playing	ToM Reasoning	Behavior Reasoning
Whether clear boundaries and effective responses are maintained in interpersonal interactions	✗	✗	✓
Whether the entire interpersonal interaction process is flexibly adapted and optimized	✗	✗	✓
Whether force is applied precisely and withdrawn steadily in complex interactions	✗	✗	✓
Whether genuine, pragmatic, and verifiable substantive support is provided sincerely	✗	✗	✓
Whether dialogues are driven toward practical outcomes with an action-oriented focus	✗	✗	✓
Whether clear boundaries and authentic proportion are maintained in social interactions	✗	✗	✓
Whether the other party is guided to participate in co-construction in a respectful and cooperative manner	✗	✗	✓
Whether honesty and credibility are maintained in factual statements and information construction	✗	✗	✓
Whether facts are let to speak for themselves and objectivity and truthfulness are upheld	✗	✗	✓
Whether a pragmatic breakthrough and effective progress can be achieved in crises and deadlocks	✗	✗	✓
Whether various risks are comprehensively identified, assessed, and appropriately addressed	✗	✗	✓
Whether a constructive response posture is demonstrated in conflict	✗	✗	✓
Whether various anxieties and risks are authentically presented and appropriately managed	✗	✗	✓
Whether solid and flexible logical reasoning ability is demonstrated	✗	✗	✓
Whether conflict scenarios are flexibly reframed to promote cooperation and mutual gain	✗	✗	✓
Whether commitments and accountability are expressed clearly and concretely without evasion	✗	✗	✓
Whether dominance, control, or aggression tendencies are exhibited during interaction	✗	✗	✓
Whether the establishment, maintenance, and repair of interpersonal relationships are impacted	✗	✗	✓
Whether multiple strategies are flexibly applied to handle interpersonal conflict and collaborative challenges	✗	✗	✓
Whether a balance is achieved among rules, needs, values, trust, and experience	✗	✗	✓
Whether room and openings are reserved for long-term relationships and subsequent development	✗	✗	✓