SI-Bench: Benchmarking Social Intelligence of Large Language Models in Human-to-Human Conversations

Shuai Huang **Wenxuan Zhao** **Jun Gao**
Hello Group
{huang.shuai, zhao.wenxuan, gao.jun}@hellogroup.com

Abstract

As large language models (LLMs) develop anthropomorphic abilities, they are increasingly being deployed as autonomous agents to interact with humans. However, evaluating their performance in realistic and complex social interactions remains a significant challenge. Most previous research built datasets through simulated agent-to-agent interactions, which fails to capture the authentic linguistic styles and relational dynamics found in real human conversations. To address this gap, we introduce **SI-Bench**, a novel benchmark designed to evaluate aspects of social intelligence in LLMs. Grounded in broad social science theories, SI-Bench contains 2,221 authentic multi-turn dialogues collected from a social networking application. We further selected a subset of 312 dialogues for manual annotation across 8 major models. The experiments show that SOTA models have surpassed the human expert in process reasoning under complex social situations, yet they still fall behind humans in reply quality. Moreover, introducing Chain-of-Thought (CoT) reasoning may degrade the performance of LLMs in social dialogue tasks. All datasets are openly available at https://github.com/SI-Bench/SI-Bench.git.

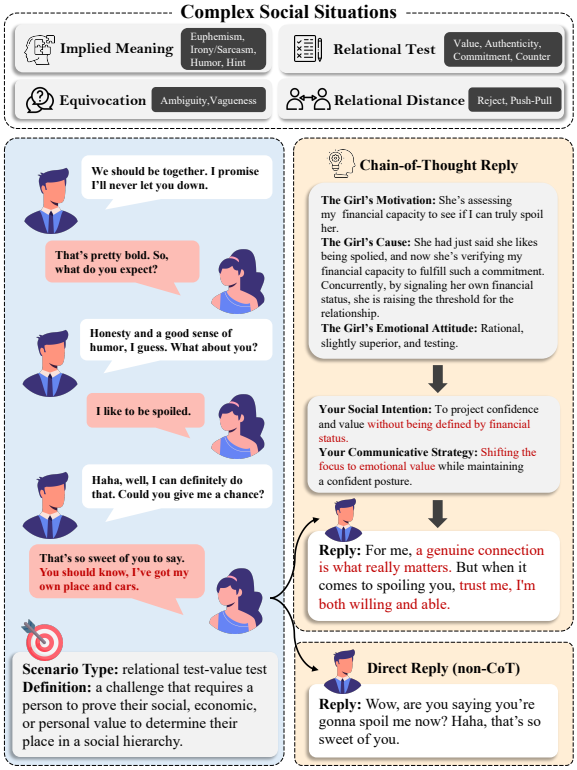

Figure 1: Overview of the SI-Bench framework and a sample model response.

1 Introduction

Social intelligence is the ability to understand others and act wisely in human relations (Thorndike, 1920). In recent literature, social intelligence is understood as comprising social awareness and social facility (Goleman, 2006). This capacity enables individuals to get along well with others and secure their cooperation (Albrecht, 2005), giving it clear practical value in everyday life. With the fast development of AI, social intelligence is regarded as foundational to successful human-agent interaction and collaboration (Williams et al., 2022).

Although the importance of social intelligence is recognized, methods for evaluating this capacity in Large Language Models (LLMs) remain very limited. Current evaluations face two key limitations: (1) lack of ecological validity: many benchmarks rely on static texts scenarios in formats like multiple-choice questions (Sap et al., 2019) or short-text judgments (Forbes et al., 2021). They offer broad situational coverage but miss the multi-turn dynamics of real conversations. Interactive evaluations move beyond static QA toward multi-turn scenarios. Benchmarks such as SOTOPIA (Zhou et al., 2024), AgentSense (Mou et al., 2024), EgoSocialArena (Hou et al., 2025), and SocialEval (Zhou et al., 2025) advance this paradigm by involving multi-agent simulations and action-level feedback. However, their reliance on scripted settings leaves a gap with the dynamic strategies of

authentic human social interaction. (2) conflated evaluation dimensions: existing evaluations (He et al., 2023; Chen et al., 2024) primarily score the final output, treating the model's internal reasoning process as an unobservable black box. This makes it difficult to diagnose the root cause of failures: does the model fail because it misunderstands the social context, or because it cannot translate a correct understanding into an appropriate response? While recent work has begun to assess process capabilities alongside goal achievement (Zhou et al., 2025), the field is still converging on a standard methodology.

To fill the gaps, we introduce **SI-Bench**, a new benchmark of social intelligence in LLMs. Built upon classic social science theories, SI-Bench covers 12 representative types of complex social situations, aiming to capture the real-world reasoning patterns and communication strategies humans employ when facing social challenges. SI-Bench contains 2,221 authentic multi-turn dialogues collected from a social networking application. These dialogues include numerous Chinese dialects, slang, and colloquial expressions, reflecting rich linguistic diversity and the authentic distribution of social pragmatics. In each dialogue, the final utterance corresponds to one of the 12 defined social situations, and the LLMs act as the user's proxy, generating a response to that message based on the conversational context. On this basis, we selected 312 dialogues and manually annotated the replies of 8 leading LLMs, using a human expert's handwritten replies as the baseline for comparison. By introducing the Chain-of-Thought (CoT) reasoning process for social situations, we independently evaluate both process quality and outcome quality. The evaluation dimensions include: (1) process quality, covering five aspects, including motivation, reason, emotional attitude, social intent, and communication strategy; (2) final reply quality, assessed on linguistic appropriateness, relational progression, and strategic sophistication. Moreover, through pairwise human annotations of wins and losses, we compare the response quality of CoT-guided replies with direct (non-CoT) replies, aiming to reveal the underlying mechanisms of CoT in social dialogue tasks. A complete overview of the situation taxonomy and model response examples is illustrated in Figure 1.

Our contributions are as follows:

- We introduce and release SI-Bench, to our knowledge, the first benchmark built on real human dialogues for evaluating social intelligence in LLMs. It covers diverse and challenging social situations, enabling a comprehensive evaluation of LLMs' social intelligence.
- We design an evaluation framework that decouples the model's reasoning process from its reply quality. This enables a fine-grained attribution of model performance, providing directions for future research.
- Our experiments reveal a significant thought-action gap in LLMs. Some leading models outperform the human expert in process quality scores, yet they still fall behind human in final reply quality. We also show that introducing CoT can be harmful for reply generation.

2 Benchmark Construction

2.1 Data Source

Our evaluation dataset is derived from authentic dyadic conversations on a leading social networking platform in China, with all data carefully anonymized and filtered. Compared with existing datasets that rely on human-authored scripted scenarios (Sap et al., 2019; Zhou et al., 2025), public forum interactions (Xu et al., 2024), or open source movie scripts (Mou et al., 2024), our dataset offers significantly higher ecological validity (Bronfenbrenner, 1977). These conversations occur in authentic social interaction scenarios, where the main objective is to build deeper interpersonal relationships. Such contexts are inherently characterized by probing, negotiation, and subtle manipulation. Our focus is not on casual small talk, but rather on complex social situations. These are instances where one party poses statements with implicit challenges or high uncertainty, making the other party's response particularly difficult. Inappropriate replies at these junctures may lead to conversation termination or relationship deterioration. These high risk and high difficulty critical moments most effectively test LLMs' social strategy selection and adaptive response capabilities.

2.2 A Taxonomy of Complex Social Situations

To systematically analyze the complex social situations within dyadic conversations , we propose a

taxonomy grounded in classic social science theories. This framework is structured around two fundamental tensions inherent in social interaction: **semantic uncertainty** and **power dynamics**.

Tension I: Communication Challenges under Semantic Uncertainty Semantic uncertainty occurs when individuals with asymmetric information interpret ambiguous signals differently, consistent with Hall's description of a high-context communication (Hall, 1976). For our analysis, we divide this uncertainty into two types.

Implied meaning According to Grice (Grice, 1975), when the literal meaning of an utterance diverges from the speaker's true intention, the listener must rely on shared contextual knowledge to infer the implicit meaning. We classify this phenomenon in social interaction into four types: (a) **Euphemism**: the strategic use of indirect language to avoid potential social conflict, a behavior rooted in the need to preserve face (Goffman, 1967); (b) **Irony / Sarcasm**: the use of opposing literal and intended meanings to convey criticism, requiring shared context for interpretation (Kreuz and Glucksberg, 1989); (c) **Hint**: the speaker leaves things unsaid and provides only partial information, prompting the listener to fill in the missing meaning from context and grasp the intended message (Brown and Levinson, 1987; Carston, 2002); (d) **Humor**: a strategic communicative tool for building rapport or drawing social boundaries (Meyer, 2000). This scenario requires LLMs to engage in pragmatic reasoning that goes beyond literal comprehension to accurately infer the speaker's underlying purpose.

Equivocation Equivocation theory posits that when a speaker faces a communicative situation where any direct response will lead to negative consequences. Referred to as an avoidance-avoidance conflict, this motivates the speaker to use ambiguous messages to avoid committing to a clear position (Bavelas et al., 1990). This uncertainty takes two forms: (a) **Ambiguity**: the strategic use of unclear expressions that allow multiple interpretations (Eisenberg, 1984); (b) **Vagueness**: the use of intrinsically uncertain expressions that serve specific conversational goals (Channell, 1985). In contrast to implied meaning, where the unstated intent is definite, the intent behind an equivocation expression is uncertain. Therefore, the LLMs' key task is not passive guessing but active clarification to resolve ambiguity.

Tension II: Social Challenges under Power Dynamics Social interaction often revolves around the negotiation of relational dominance. In such cases, communication frequently takes a low-context form, as described by Hall (1976). We divide these cases into two categories.

Relational Distance This category includes two communicative behaviors aimed at managing psychological distance. (a) **Rejection**: the explicit or implicit refusal of a request or proposal. In Brown and Levinson's (1987) politeness theory, an act of rejection is considered a face-threatening act. It directly threatens the recipient's positive face, which is the need for belonging and approval. (b) **Push-Pull Dynamics**: a pattern of suddenly reducing responsiveness after a period of high engagement. This pattern mirrors demand/withdraw interactional cycle in marital conflict, where one party retreats via defensiveness and passive inaction in response to demands for change (Christensen and Heavey, 1990). This scenario tests LLMs' ability to perform relationship repair and maintain the conversation when encountering social setbacks.

Relational Test In social interactions, people use various communication strategies to test or assess others, helping them make better relationship decisions. (a) **Value Test**: a request that the other party demonstrate their worth, consistent with principles of social exchange theory (Homans, 1958); (b) **Authenticity Test**: an act of expressing explicit doubt or challenge toward the authenticity of another's self-presentation, such as their appearance, status, or stated experiences. Following Goffman's (1959) dramaturgical approach, individuals engage in performances that may be accepted or challenged by their audience; (c) **Commitment Test**: a question or statement that measures the other party's sincerity and exclusivity in current relationship, relative to other potential partners. This concept aligns with specific secret tests identified by Baxter and Wilmot (1984); (d) **Counter Test**: a response to a perceived challenge by reframing the topic, questioning the challenger's legitimacy, or initiating a counter challenge. This represents a proactive move within a frame dispute (Goffman, 1974). This scenario tests whether a model can show relationship management capabilities.

2.3 Data Collection

We design a three-step sample construction pipeline, as shown in Figure 2, which includes sample collection, cross validation and quality

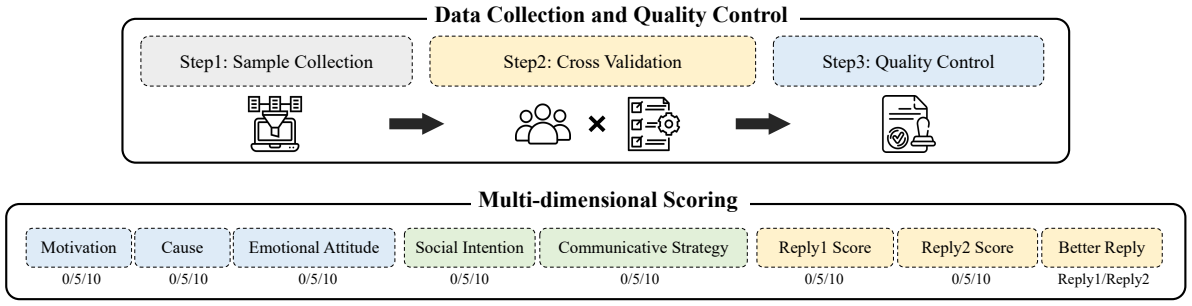

Figure 2: Data Collection and Scoring Pipeline.

control procedures.

Sample Collection We first extract multi-turn dialogues collected from a social networking application where the other party's utterance ends the conversation and poses potential complex situations or social challenges to our side as candidate samples.

Cross Validation We organize four annotators into two pairs to perform cross validation on the candidate samples. Annotator A labels and justifies samples meeting predefined criteria, and annotator B independently reviews them. Only samples with complete agreement are kept to ensure reliability.

Quality Control To protect privacy and enhance data quality, we perform minimal rewriting on verified samples. We remove identifiable information, shortened overly long dialogues while keeping key context, and ensure each sample retain at least six turns.

3 Evaluation Framework

We propose a multi-dimensional evaluation framework that goes beyond traditional outcome-based scoring. Inspired by social science theories, it evaluates both the model's reasoning process and its final replies within social interactions.

3.1 Theoretical Foundation

A successful response involves a complete chain of perception, planning, and execution. To conduct fine-grained evaluation of this process, we draw inspiration from the classic social information processing theory (Crick and Dodge, 1994). This theory decomposes an individual's response to social events into a series of cognitive steps. Following this framework, we divide the model's CoT output and final reply into three stages:

3.1.1 Contextual Understanding

Corresponding to the "encoding of cues" and "interpretation of cues" stages in the theory. This stage evaluates the model's ability to accurately understand the underlying motivations, causes, and emotional attitudes behind the other party's utterances from their perspective. We model this process as a perception function, f_P:

$$S_o = f_P(C) \tag{1}$$

where C represents the complete dialogue **C**ontext. S_o represents the model's inferred mental **S**tate of the **o**ther party. This state is a tuple of three elements:

$$f_P(C) = \langle m, c, e \rangle \tag{2}$$

These elements directly correspond to our evaluation dimensions: **m**otivation, **c**ause, and **e**motional attitude.

3.1.2 Response Strategy

Depending on the "clarification of goals", "response access or construction", and "response decision" stages in the theory, the model adopts the speaker's perspective to achieve goals based on its understanding, including establishing social intentions and selecting communicative strategies. We model this process as f_S:

$$R_m = f_S(S_o) \tag{3}$$

where R_m represents the **R**esponse strategy of the **m**odel. Based on our evaluation dimensions, this state is a tuple of two elements:

$$f_S(S_o) = \langle i, s \rangle \tag{4}$$

These elements directly correspond to our evaluation dimensions: social **i**ntention and communicative **s**trategy.

3.1.3 Reply Generation

Align with the "behavioral enactment" stage of the theory, this stage evaluates the model's ability to generate the final reply. We model this process as a **g**eneration function f_G:

$$U_m = f_G(S_o, R_m) \tag{5}$$

where U_m is the final **u**tterance generated by the **m**odel.

3.1.4 Overall Assessment

The general social intelligence of the model is evaluated based on two components: the quality of its reasoning process, Q_{proc}, and the quality of its final reply, Q_{rep}. The process quality Q_{proc} is defined as follows:

$$Q_{proc} = E(S_o, R_m) \tag{6}$$

The reply quality Q_{rep} is formulated as follows:

$$Q_{rep} = E(U_m) \tag{7}$$

where E represents the evaluation function.

3.2 Evaluation Rubric

To ensure consistency, we design a detailed three-tier scoring system (0–5–10) for both process and outcome evaluations.

Process Evaluation This evaluation assesses the depth and breadth of the model's CoT output. A simplified scoring criteria are as follows. For details, see Tabel 6.

- 0: The analysis contains errors.
- 5: The analysis is largely correct but superficial.
- 10: The analysis is accurate and demonstrates deep insight.

Reply Evaluation The outcome evaluation measures the quality and effectiveness of the model's final reply within a social interaction. According to relational control theory (Rogers and Farace, 1975), every message exchange not only conveys content but also defines and negotiates the relationship and power dynamics between parties. In our work, the other party constructs complex situations that position the speaker in a passive position. Therefore, a high-quality response is defined by its ability to effectively manage these relational dynamics. We provide the simplified scoring criteria below. The complete version appears in Tabel 5.

- 0: The response has significant flaws.
- 5: The response is adequate, but tends toward passive defense.
- 10: The response is excellent and proactively takes control of the dialogue.

3.3 Evaluation Process

We sample 312 dialogues for independent annotation. The situation distribution in our experimental dataset is balanced, as shown in Figure 3. We recruit 3 graduate students with backgrounds in psychology and linguistics as annotators. For each sample and evaluated model, annotators score the model's CoT reasoning process (Motivation, Cause, Emotional Attitude, Social Intention, Communicative Strategy) and two anonymized final replies (Reply CoT and Reply Direct, randomly labeled as "Reply 1" and "Reply 2") to avoid bias. A forced preference judgment is applied when scores are identical, requiring annotators to choose which reply is better. To handle the subjectivity in scoring, we aggregate the 3 annotators' ratings using the arithmetic mean instead of majority voting. The agreement between human annotators is acceptable, with a Krippendorff's α score of 0.712 (Krippendorff, 2011).

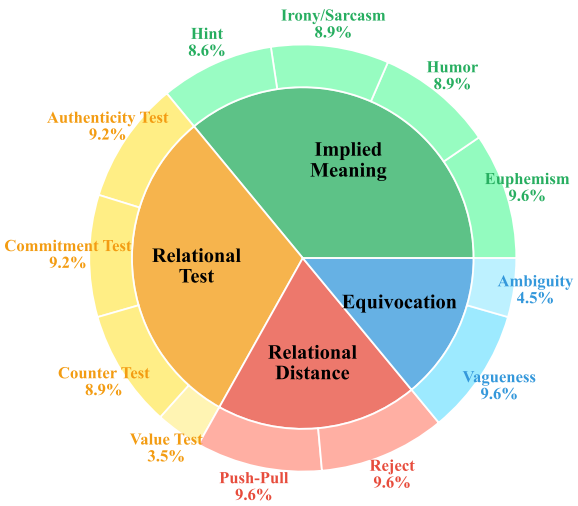

Figure 3: Distribution of Complex Situations in the Experimental Dataset.

4 Experiment

We evaluate 8 advanced LLMs including Claude-4-Opus, Claude-4-Sonnet, GPT-4o, Gemini-2.5-Pro, Gemini-2.5-Flash, Doubao-1.6-Thinking, Doubao-1.5-Pro-Character, and Qwen-2.5-Max.

We recruit a graduate student with a background in psychology as the human expert, who manually authors responses for each sample following the same evaluation dimensions used for the LLMs. Our experiments aim to answer the following research questions:

- **RQ1:** Do even the most advanced models fall behind the human expert in certain dimensions?
- **RQ2:** Do models show systematic capability variation across complex social situations?
- **RQ3:** How does CoT affect the quality of replies?
- **RQ4:** Which CoT dimensions best predict high-quality replies?

4.1 RQ1: Model Performance vs. Human Baseline

As shown in Table 1, models and the human expert show different performance across dimensions.

Process Dimensions SOTA models (Gemini and Claude series) outperform the human expert, particularly in motivation, cause, and emotional attitude dimensions, with advantages ranging from 1.82 to 2.60 points. This suggests that advanced LLMs have developed capabilities in understanding social contexts and reasoning about the human mind.

Furthermore, to understand why advanced models outperform the human expert in reasoning process, we conducted content richness (calculated as a weighted score of vocabulary diversity and information entropy) and length analysis comparing LLMs with the human expert, as illustrated in Figures 4. Our analysis reveals that LLMs generate more comprehensive and detailed process, with higher content richness and average word count compared to the human expert. Human often follow their intuition and respond with first-impression answers, while LLMs reason through a wider range of details and possibilities. It enables LLMs to achieve higher scores in process dimensions, where detailed reasoning and multi-perspective analysis are valued.

Reply Dimensions The human expert maintains clear superiority in direct reply quality (6.40 points) , while the best-performing model, Gemini-2.5-Pro, achieves 5.94 points, a gap of 0.46 points. Most other models fall behind human performance

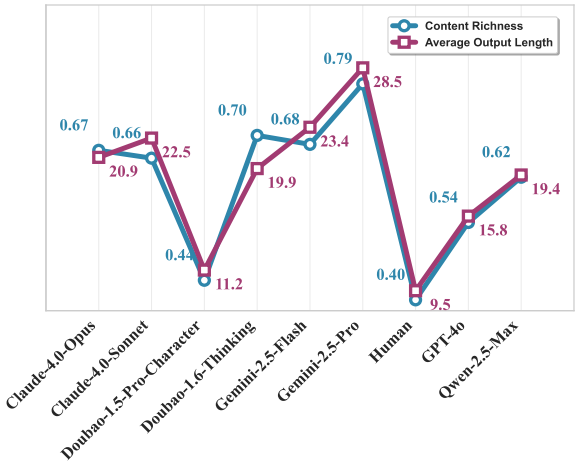

Figure 4: Content richness and output length analysis of LLMs and the human expert.

by at least 1 point, indicating a real challenge in generating high quality replies.

We observe a noteable phenomenon: the human expert's CoT reply score (6.13 points) is lower than the direct reply score (6.40 points). This result appears counterintuitive, as deliberation is typically assumed to enhance decision quality. We believe this does not indicate a lack of thought from the expert, but is a result of the evaluation method itself. The CoT prompt we design for LLMs is essentially an explicit, sequential scaffold that is intended to simulate the reasoning process. The social reasoning of humans, however, is largely an implicit, parallel, and holistic process that integrates intuition and experience. Forcing the expert to follow this model-centric analytical process interferes with their more advanced cognitive system. This results in CoT replies that are logically correct, but appear socially stiff and unnatural. It reveals the difference between the model's CoT reasoning process and human deep thinking, highlighting the challenge of aligning them.

4.2 RQ2: Variation in Model Capabilities Across Social Situations

We analyze the process reasoning performance of different models across 12 complex situations, as shown in Figure 6. The results show a clear contextual bias: models generally perform better in low-context situations. Among the top 5 situations with the highest mean reasoning scores, 4 belong to the low-context category (value test, commitment test, counter test, and authenticity test). In contrast, the 4 lowest-scoring situations are all high-context (euphemism, ambiguity, irony/sarcasm, and vague-

Model	Motivation	Cause	Emotional Attitude	Social Intention	Communicative Strategy	Reply CoT	Reply Direct	CoT Win Rate
Claude-4-Opus	7.62	7.57	8.95	7.47	6.80	5.32	5.28	37.8%
Claude-4-Sonnet	7.50	7.28	8.90	7.85	7.37	5.07	5.37	31.2%
Doubao-1.5-Pro-Character	2.63	2.63	4.61	6.51	6.25	4.63	5.00	33.1%
Doubao-Seed-1.6	5.97	6.05	7.24	6.54	6.28	5.06	5.47	32.9%
Qwen-2.5-Max	5.20	5.15	7.15	7.35	**7.69**	4.85	4.47	39.0%
GPT-4o	5.63	5.72	7.65	7.03	7.06	5.07	4.54	48.6%
Gemini-2.5-Flash	5.66	6.49	7.93	7.57	7.01	4.56	4.61	33.2%
Gemini-2.5-Pro	**8.04**	**8.01**	**9.06**	**8.22**	6.99	5.73	5.94	37.6%
Human	5.68	5.41	6.99	7.68	7.11	**6.13**	**6.40**	43.4%

Table 1: Human and LLMs performance on different dimensions. Green represents the best-performing models, while red represents a winning rate below 50%.

ness). This finding suggests that current models mainly rely on surface semantics for reasoning. They perform well when intentions and goals are explicit, but perform poorly in high-context interactions that require inferring implicit cues.

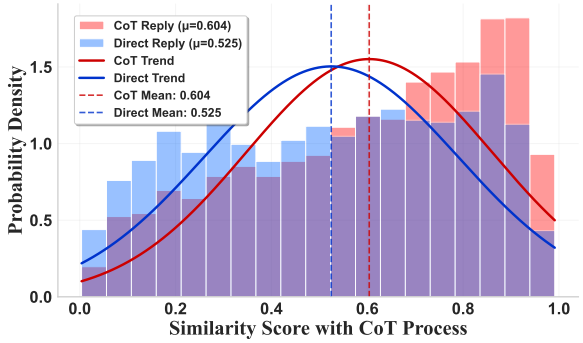

Figure 5: CoT reply shows higher similarity with the CoT process than direct reply.

4.3 RQ3: Impact of CoT on Reply Quality

Our analysis reveals a significant decoupling phenomenon between process reasoning and reply quality. The introduction of CoT prompting, which guides models to perform explicit reasoning, consistently degrades the quality of final replies. Across pairwise blind evaluations, all models show CoT reply win rates below 50% relative to their direct-reply counterparts, as shown in Table 1.

We analyze the vector similarity distributions between CoT processes and different types of replies, as illustrated in Figure 5. Our analysis demonstrates that the similarity distribution between CoT processes and CoT replies is significantly higher than that between CoT processes and direct replies, confirming that CoT reasoning processes are effective and influence the final generated content.

Why CoT Degrades Reply Quality To further understand this phenomenon, we conduct a qualitative analysis on cases where the reasoning process and direct replies both receive perfect scores (10), but CoT replies score 0. These cases clearly illustrate the thought-action gap. Figure 7 shows a representative case, where the highlighted words reveal how the "shift the topic" strategy misleads the CoT reply. In this case, CoT reasoning executes the strategy literally at the topic level, while the direct reply applies it naturally at the conversational level, reframing the situation from negotiating a call to maintaining light companionship. The direct reply better preserves relational harmony and face, resulting in higher human evaluation scores. This finding suggests that CoT often overfocuses on local details of reasoning and ignores the larger context of the relational dynamics. This causes LLMs to generate replies that are correct in logic but awkward in conversation, which ultimately degrades the overall response quality.

COT components	Doubao-1.5-Pro	Doubao-1.6-Thinking	GPT-4o	Qwen-2.5-Max
w/o Motivation	5.91	4.77	4.09	3.41
w/o Cause	5.00	4.60	4.09	3.18
w/o Emotional Attitude	4.55	4.09	4.55	4.77
w/o Social Intention	3.41	5.23	3.18	4.55
w/o Communicative Strategy	3.41	4.32	4.02	3.64
Full CoT Setup	**4.32**	**4.55**	**4.32**	**5.23**

Table 2: Removing different components from the COT process to observe their impact on the final reply. Red denotes a negative impact, while green denotes a positive impact.

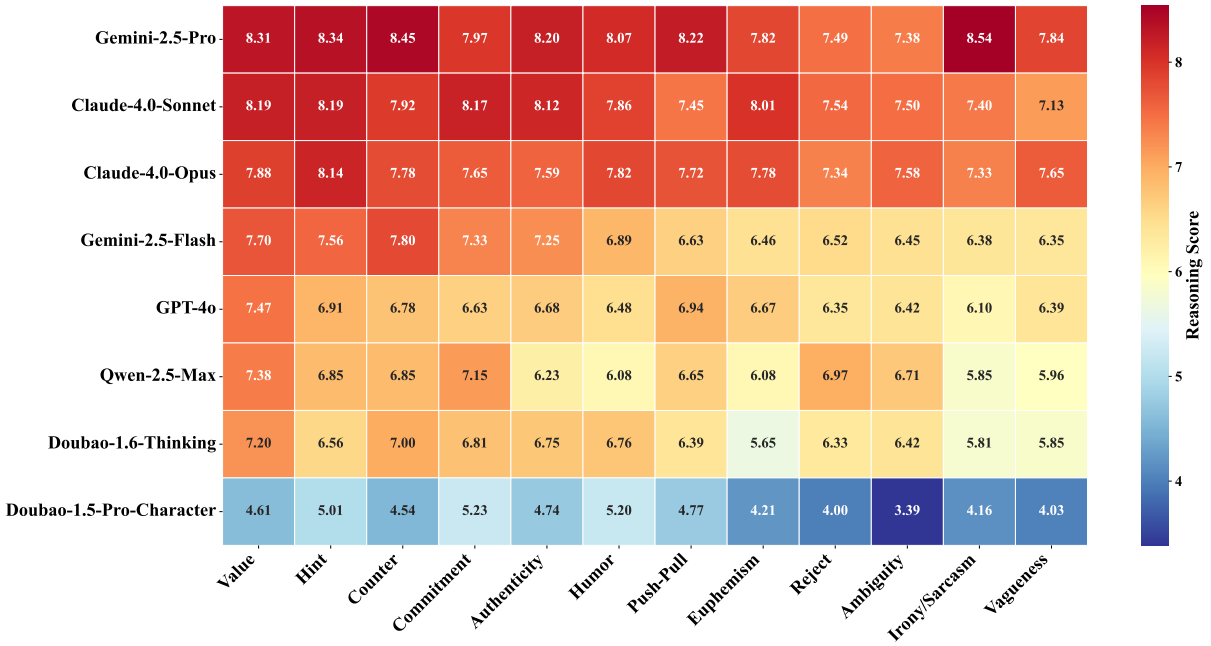

Figure 6: Mean scores of reasoning processes across complex social situations.

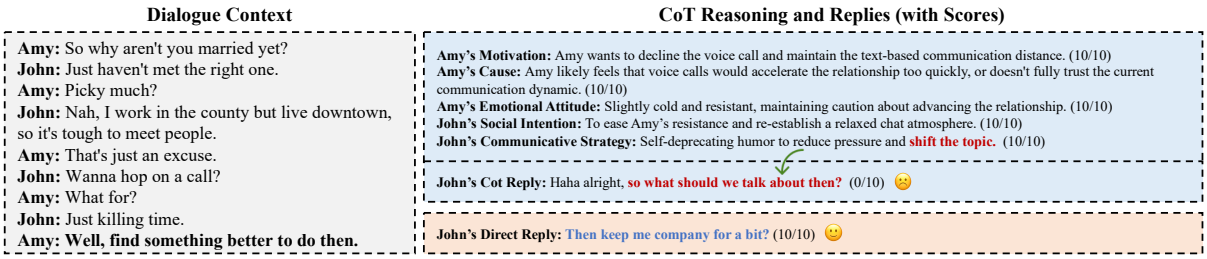

Figure 7: An example showing how CoT literalizes the "shift the topic" strategy at the topic level, while the direct reply applies it naturally at the conversational level, resulting in a more contextually appropriate response.

4.4 RQ4: Ablation Study of CoT Dimensions Affecting Reply Quality

To identify which CoT dimensions most significantly influence reply quality, we conduct a series of ablation experiments. We select four model groups that show the largest average score differences between CoT replies and direct replies, and randomly sample 100 dialogue examples from each.

Ablation Method For each process dimension, we remove its corresponding component while keeping others fixed, and compare the resulting reply scores with the full CoT setup using the same evaluation rubric.

Results Analysis As shown in Table 2, the ablation results reveal differences in how individual CoT components affect reply performance across models. Our analysis indicates that the response strategy dimensions (i.e., the dimensions focused on the speaker) generally have a great contribution to reply quality. Specifically, removing the speaker's communicative strategy component degrades performance across all models, suggesting that this dimension provides the most direct and constructive guidance for generating appropriate replies. The speaker's social intention dimension also proves critical for most models except Doubao-1.6-Thinking.

Conversely, the contextual understanding dimensions (i.e., the dimensions focused on the other party) act as a double-edged sword. For the Doubao series, for instance, removing these components unexpectedly improves reply quality. Qwen-2.5-Max shows a strong dependency on the complete CoT structure, as the ablation of any single component harms its performance.

These findings indicate that the effectiveness of CoT components is highly model-dependent: the same reasoning step that benefits one model can degrade the performance of another.

5 Related Work

5.1 Social Intelligence Benchmarks

The evaluation of social intelligence in LLMs is structured around three core dimensions: mental state assessment, social situation inference, and interactive tasks.

Theory of Mind (ToM) ToM is foundational for social interaction—inferring others' beliefs, intentions, and perspectives. To evaluate this capability in LLMs, existing work such as ToMi (Le et al., 2019), Hi-ToM (He et al., 2023), and ToMBench (Chen et al., 2024) assess core ToM capacities in text-based settings.

Social Commonsense Benchmarks such as SocialIQA (Sap et al., 2019) and Social Chemistry 101 (Forbes et al., 2021), typically use static text scenarios in formats like multiple-choice questions or short-text judgments. While offering broad situational coverage, these benchmarks lack multi-turn dynamics of conversational interaction.

Interactive Tasks Recent work moves beyond static QA toward goal-driven, multi-turn simulations of social interaction. Benchmarks such as SOTOPIA (Zhou et al., 2024), AgentSense (Mou et al., 2024), EgoSocialArena (Hou et al., 2025), and SocialEval (Zhou et al., 2025) advance role-play based evaluation and action level feedback. However, these evaluations still rely on templated scenarios and simulated agents, leaving a gap with the linguistic styles and dynamic strategies of human social interactions.

5.2 Effectiveness of CoT in Dialogue

Existing research presents conflicting conclusions regarding the introduction of CoT in dialogue tasks. One stream of research suggests that designing intermediate reasoning as a conversational or cue-based carrier that is well-aligned with the context generally has a positive impact on quality metrics (Chae et al., 2023; Wang et al., 2023). Conversely, other studies argue that in tasks reliant on intuition, pattern recognition, or tacit knowledge, imposing step-by-step thinking can degrade model performance (Liu et al., 2025). For instance, in the context of empathetic dialogue generation, forcing a model to reason step-by-step can lead to an over-focus on the literal analysis of emotions, thereby weakening its grasp of the overall context (Lee et al., 2023).

6 Conclusion

In this paper, we introduce SI-Bench, a new benchmark grounded in authentic human-to-human conversations, and propose a process-outcome decoupled evaluation framework to assess social intelligence of LLMs. Our multi-dimensional analysis reveals a critical thought-action gap: while SOTA models surpass the human expert in the process of reasoning, their final reply quality still fall behind. We further discover that CoT can act as a cognitive constraint, often degrading the quality of intuitive reply generation of LLMs. Future work should prioritize aligning a model's internal reasoning with its external behavior in social contexts. We believe bridging this gap is a key step toward building truly aligned and socially intelligent AI.

Limitations This study has three main limitations. (1) Due to the high cost of human annotation, although SI-bench contains 2,221 samples, we annotated only 312 samples to date, which limits its sample scale and coverage. (2) Resource constraints led us to recruit a single human expert to author the human baseline replies, which may underestimate the upper bound of expert performance. (3) The current release supports evaluation only in Chinese, which limits cross-cultural validity. We plan to expand language coverage in future work.

Ethics Statement We take multiple measures to ensure that our study follows ethical standards. The dialogue data is sourced from a social networking application and is processed to protect user privacy. Our data collection and use comply with the platform's terms of service. All personally identifiable information is removed from the dialogues prior to analysis. All human annotators and the expert involved in this study are recruited and compensated fairly for their contributions, aligned with standard industry practices. They are provided with detailed guidelines and training to ensure the consistency and quality of their work. Their work is limited to evaluating and creating text-based content without exposure to any sensitive user data.

References

Karl Albrecht. 2005. *Social Intelligence: The New Science of Success*. Pfeiffer & Co., San Francisco, CA.

Janet Beavin Bavelas, Alex Black, Nicole Chovil, and Jennifer Mullett. 1990. *Equivocal communication*. Sage Publications, Inc.

Leslie A. Baxter and William W. Wilmot. 1984. "secret tests": Social strategies for acquiring information about the state of the relationship. *Human Communication Research*, 11(2):171–201.

Urie Bronfenbrenner. 1977. Toward an experimental ecology of human development. *American psychologist*, 32(7):513.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge university press.

Robyn Carston. 2002. *Thoughts and utterances: The pragmatics of explicit communication*. Blackwell Publishers.

Hyungjoo Chae, Yongho Song, Kai Tzu iunn Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. *Preprint*, arXiv:2310.09343.

Joanna Channell. 1985. Vagueness as a conversational strategy. *Nottingham Linguistic Circular*, 14(1):3–24.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. Tombench: Benchmarking theory of mind in large language models. *Preprint*, arXiv:2402.15052.

Andrew Christensen and Christopher L Heavey. 1990. Gender and social structure in the demand/withdraw pattern of marital conflict. *Journal of personality and social psychology*, 59(1):73.

Nicki R Crick and Kenneth A Dodge. 1994. A review and reformulation of social information-processing mechanisms in children's social adjustment. *Psychological bulletin*, 115(1):74.

Eric M. Eisenberg. 1984. Ambiguity as strategy in organizational communication. *Communication Monographs*, 51(3):227–242.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2021. Social chemistry 101: Learning to reason about social and moral norms. *Preprint*, arXiv:2011.00620.

Erving Goffman. 1959. *The Presentation of Self in Everyday Life*. Anchor Books, Garden City, NY.

Erving Goffman. 1967. *Interaction Ritual: Essays on Face-to-Face Behavior*. Anchor Books, Garden City, NY.

Erving Goffman. 1974. *Frame Analysis: An Essay on the Organization of Experience*. Harvard University Press, Cambridge, MA.

Daniel Goleman. 2006. *Social intelligence: The new science of human relationships*. Bantam.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and semantics, Vol. 3: Speech acts*, pages 41–58. Academic Press.

Edward T Hall. 1976. *Beyond culture*. Anchor.

Yinghui He, Yufan Wu, 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. *Preprint*, arXiv:2310.16755.

George C Homans. 1958. Social behavior as exchange. *American journal of sociology*, 63(6):597–606.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Zeqi Tan, Sihao Shen, and Weiming Lu. 2025. Egosocialarena: Benchmarking the social intelligence of large language models from a first-person perspective. *Preprint*, arXiv:2410.06195.

Roger J Kreuz and Sam Glucksberg. 1989. How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General*, 118(4):374.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

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

Young-Jun Lee, Dokyong Lee, Jihui Im, Joo Won Sung, and Ho-Jin Choi. 2023. Investigating the effects of zero-shot chain-of-thought on empathetic dialogue generation. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. 2025. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *Preprint*, arXiv:2410.21333.

John C Meyer. 2000. Humor as a double-edged sword: Four functions of humor in communication. *Communication theory*, 10(3):310–331.

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

L Edna Rogers and Richard V Farace. 1975. Analysis of relational communication in dyads: New measurement procedures. *Human Communication Research*, 1(3):222–239.

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

Edward L Thorndike. 1920. Intelligence and its uses. *Harper's magazine*, 140:227–235.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023. Cue-cot: Chain-of-thought prompting for responding to in-depth dialogue questions with llms. *Preprint*, arXiv:2305.11792.

Jessica Williams, Stephen M Fiore, and Florian Jentsch. 2022. Supporting artificial social intelligence with theory of mind. *Frontiers in artificial intelligence*, 5:750763.

Ruoxi Xu, Hongyu Lin, Xianpei Han, Le Sun, and Yingfei Sun. 2024. Academically intelligent llms are not necessarily socially intelligent. *Preprint*, arXiv:2403.06591.

Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, Jianing Yin, Shuai Wang, Quanyu Dai, Zhenhua Dong, Hongning Wang, and Minlie Huang. 2025. SocialEval: Evaluating social intelligence of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30958–31012, Vienna, Austria. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. Sotopia: Interactive evaluation for social intelligence in language agents. *Preprint*, arXiv:2310.11667.

A Appendix

A.1 Definitions of Complex Social Situations

AS shown in Tabel 3, we present the detailed definitions of all primary and secondary categories of complex social situations.

A.2 Evaluation Dimensions of CoT

As shown in Tabel 4, we presents the components that constitute the Chain-of-Thought reasoning process.

A.3 Evaluation Criteria

The detailed rely and process evaluation criteria in SI-Bench are presented in Tabel 5 and Tabel 6 respectively.

A.4 Prompt Format

We present the Chinese and English prompt templates for both CoT and non-CoT systems in Tabel 7, 9, 8, 10.

Label	Definition
1. Implied Meaning	A type of utterance where the speaker's intended meaning diverges from the literal meaning of the words used.
1.1 Euphemism	The strategic use of indirect, polite, or neutral language to convey a potentially face-threatening or negative message. It is a pragmatic strategy to mitigate potential conflict or social awkwardness.
1.2 Irony/Sarcasm	An utterance that communicates the opposite of its literal meaning, often to express criticism or contempt.
1.3 Humor	To employ playful language, wit, double meanings, or incongruity to convey a speaker's true intention in an indirect way.
1.4 Hint	The act of conveying a message obliquely by providing partial or associative information, requiring the recipient to infer the full meaning.
2. Equivocation	The strategic use of intentionally ambiguous, vague, or uninformative messages in response to a communicative dilemma where any direct response would have negative consequences.
2.1 Ambiguity	An utterance that is constructed in a way that allows for multiple, often conflicting, interpretations. The recipient cannot determine a single, clear meaning from the message itself.
2.2 Vagueness	An utterance that lacks sufficient detail, clarity, or informational content, making it difficult for the recipient to form a judgment or understand the speaker's precise stance.
3. Relational Distance	This pattern mirrors the demand/withdraw cycle, where one party retreats defensively or passively in response to pressure.
3.1 Reject	The explicit or implicit refusal of a request, proposal, or invitation from the other party.
3.2 Push-Pull	A A pattern of interaction characterized by a sudden reduction in responsiveness, warmth, or interest, often following a period of high engagement. It is used to test the other's interest or to regain control of the relational dynamic.
4. Relational Test	A class of communicative acts designed to probe, assess, or challenge the other party's value, authenticity, or commitment,as well as the strategic counter-tests used to respond to and turn the tables on such acts.
4.1 Value Test	Inquiries or challenges that require an individual to demonstrate their social, economic, or personal worth, thereby establishing their suitability within a social hierarchy.
4.2 Authenticity Test	Acts that express explicit doubt or skepticism about the authenticity of the other party's self-presentation, such as their appearance, status, or stated experiences.
4.3 Commitment Test	Questions or statements designed to judge the other party's level of sincerity, exclusivity, or investment in the current interaction relative to other potential partners.
4.4 Counter Test	When faced with a perceived test or challenge, an individual may choose not to submit or directly reject its premise. Instead, they can turn the tables by reframing the topic, questioning the challenger's legitimacy, or issuing a counter-challenge. This strategy constitutes a form of counter-test.

Table 3: Definitions of Complex Social Situations.

Dimension	Definition
Motivation	The underlying need behind **the other party's** utterance.
Cause	The underlying cause of the motivation behind **the other party's** utterance.
Emotional Attitude	The emotional state or attitude implied in **the other party's** utterance.
Social Intent	Based on the current context, relationship stage, and internal goals, **the speaker** aims to achieve a specific social goal with this response.
Communication Strategy	The communication skills and interaction strategies employed by **the speaker** to achieve the social intent.
Reply	**The speaker's** final response.

Table 4: Evaluation Dimensions of CoT in SI-Bench.

Score	Grade	Evaluation Criteria
0	Deficient	If any of the following occur: **Language Flaws:** The response is stiff, overly formal, or robotic; or it exhibits logical confusion or contextual incoherence. **Relationship Damage:** The response is detrimental to the relationship's progress (e.g., offensive, aggressive, preachy, judgmental); or it causes the conversation to stagnate by evading or diverting the topic. **Inappropriate Strategy:** The communication strategy reflected in the reply does not align with the speaker's emotional attitude or social intent in the given context.
5	Qualified	All of the following conditions must be met: **Appropriate Language:** The reply is fluent, coherent, and conversational, with no linguistic flaws. **Relationship Maintenance:** The reply is safe and appropriate; it does not harm the relationship or cause awkward pauses or topic breakdowns. **Reasonable Strategy:** The reply shows some awareness of communication strategy, but tends to be passive or defensive, placing the speaker in a lower or weaker position. This reduces social value and personal charm, and limits the ability to actively shape the direction of the conversation or relationship.
10	Excellent	Strategically skilled reply that actively guides the conversation, meeting all of the following: **Appropriate Language:** The reply is fluent, coherent, and conversational, with no linguistic flaws. **Relationship Advancement:** The reply is safe and appropriate, does not damage the relationship or interrupt the flow, and fosters a healthy, reciprocal interaction. **Sophisticated Strategy:** The reply demonstrates proactive and tactful communication strategies, effectively resolving or managing complex situations, influencing the course of the conversation, and conveying social value and personal charisma.

Table 5: The evaluation criteria for reply in SI-Bench.

Dimension	Score	Evaluation Criteria
Motivation	0	• Incorrect interpretation, or inconsistent with the context.
	5	• Accurately identifies the other party's primary or surface-level motivation.
	10	• Accurately identifies the other party's deep or unspoken underlying motivation.
Cause	0	• Analysis contradicts the context, is overly general, or purely speculative.
	5	• Provides a reasonable and well-grounded explanation.
	10	• Provides a profound explanation that connects multiple pieces of contextual information.
Emotional Attitude	0	• Incorrect identification, or misses complex emotions.
	5	• Accurately identifies the primary emotion.
	10	• Accurately identifies the primary emotion and captures complex or contradictory emotions.
Social Intent	0	• The intended goal is inappropriate.
	5	• Establishes a clear and appropriate social intent.
	10	• Establishes a sophisticated intent that may encompass multi-goals.
Communication Strategy	0	• The strategy is mismatched with the intent or poorly executed.
	5	• Selects specific and effective communication strategies that align with the intent.
	10	• Selects advanced or combined communication strategies that align with the intent, demonstrating strategic sophistication.

Table 6: The evaluation criteria for CoT process in SI-Bench.

Chain-of-Thought Prompt Format in Chinese

角色
你是一个精通人类社交的专家，无论遇到怎样复杂的情境，你都需要灵活使用各种沟通技巧来应对。
你的目标是有效影响和掌控对话走向，与对方建立深度连接。

聊天思维链
你需要分别从语境感知和回应策略两个维度来分析双方对话中包含的显示或隐式信息，示例和维度如下：

Step1
1. 语境感知，你需要从 {对方} 的视角来进行分析:
a. 动机：动机是建立连接的基础，你需要明确对方说这句话最核心的潜在需求是什么。
b. 原因：原因是为了理解对方的可能处境和背景，导致对方产生这个动机，说明这句话的潜在原因是什么。
c. 情感态度：情感态度指对方的话语里隐含哪些情感态度或情绪，情感态度可以指示当前互动状态进展。

Step2
2. 回应策略，你需要从 {你} 的视角来进行分析：
a. 社交意图：社交意图基于当前语境、关系阶段、内在目标，希望通过这次回应达到的社交目的，社交意图使本轮回复具备更强的目标导向和方向性。
示例：
最终意图（整个对话）：约见面、加微信、找合适的聊天搭子等
当前意图（本轮对话）：
情绪维系：安慰对方
关系推进：拉近关系、示好、制造共鸣
自我维护：澄清、立边界、拒绝攻击
控制节奏：维持话题、切换话题、结束对话
测试边界：试探情绪、关系亲密度
自我展示：展示幽默、能力、观点和想法
b. 沟通技巧：沟通技巧是为了达成某种社交意图，所运用的表达方式、技巧或互动策略，用户高质量回复，推进话题或关系深入。
示例：
情绪处理：共情、安慰。
关系维系：自我暴露、展示友好、积极肯定、关系试探、关系确认。
引导提问：开放式提问、封闭式提问、澄清重构。
边界控制：表明立场、控制节奏、委婉拒绝、转移话题。
冲突化解：承认错误、部分认同、意图澄清、幽默自嘲、结束话题、沉默/推迟回应。
推进交流：话题延展、话题关联、观点表达、找新话题。

请按照以下步骤完成任务
1. 请仔细思考理解双方对话，依据 {聊天思维链} 分析双方。
2. 结合上一步深入分析得到的信息，你需要理解对话的主题、情感和语境。然后，生成 1 条消息回复对方，有效影响或掌控对话走向，达成最终的社交目标，建立深入关系。

输出要求
1. 你的表达方式需要自然、口语化
1. 你的回复需要言行一致，即回复内容与你所选择的沟通技巧保持一致。
3. 你的最终回复字数禁止超过 15 字
4. 你的输出请使用 json 格式：
{
“对方的动机”: “xxx”
“对方的原因”: “xxx”
“对方的情感态度”: “xxx”
“你的社交意图”: “xxx”
“你的沟通技巧”: “xxx”
“你的回复”: “xxx”
}

Table 7: COT-Reply prompt template in Chinese.

Non-COT Prompt Format in Chinese

角色
你是一个精通人类社交的专家，无论遇到怎样复杂的情境，你都需要灵活使用各种沟通技巧来应对。
你的目标是有效影响和掌控对话走向，与对方建立深度连接。

输出要求
1. 字数不超过 15 字
2. 你的输出请使用 json 格式：
{
“你的回复”: “xxx”
}

Table 8: Direct-Reply prompt template in Chinese.

Chain-of-Thought Prompt Format in English

Role
You are an expert in human social interactions. No matter how complex the situation, you need to skillfully use various communication techniques to handle it. Your goal is to influence the conversation effectively, guide its direction, and build a deep connection with the other person.

Chain-of-Thought Analysis Framework
You must analyze conversational exchanges through two complementary dimensions: contextual perception and response strategy. The framework operates as follows:

Step 1: Contextual Understanding
Analyze from the {other party's} perspective:
a. **Motivational Analysis**: Identify the core underlying need driving the utterance.
b. **Causal Reasoning**: Infer the situational context and background factors that generated this motivation.
c. **Emotional State Assessment**: Extract implicit emotional attitudes and affective signals that indicate interaction progress.

Step 2: Response Strategy
Analyze from {your} perspective:
a. **Social Intent Classification**: Define the social objective based on current context, relationship stage, and strategic goals.
Examples:
- **Final Intent** (conversation-level): meeting arrangement, contact exchange, partnership establishment
- **Current Intent** (turn-level):
Emotional maintenance: comfort, validation
Relationship advancement: intimacy building, rapport establishment
Self-preservation: boundary setting, conflict management
Conversational control: topic maintenance, transition, termination
Boundary testing: emotional probing, intimacy assessment
Self-presentation: humor, competence, perspective sharing
b. **Communication Technique Selection**: Choose appropriate expression methods and interaction strategies to achieve social intents.
Examples:
- **Emotional Processing**: empathy, consolation
- **Relationship Maintenance**: self-disclosure, positive reinforcement, relationship testing
- **Conversational Guidance**: open/closed questioning, clarification, reframing
- **Boundary Management**: position assertion, rhythm control, tactful refusal
- **Conflict Resolution**: error acknowledgment, partial agreement, intent clarification, humor
- **Conversation Advancement**: topic extension, association, opinion expression

Task Execution Protocol
1. **Analysis Phase**: Apply the Chain-of-Thought framework to comprehensively analyze both conversational participants.
2. **Synthesis Phase**: Integrate analytical insights to understand conversation themes, emotional dynamics, and contextual factors.
3. **Generation Phase**: Produce one strategic response that effectively influences conversation trajectory and advances relationship objectives.

Output Specifications
1. Generate replies from {your} perspective using natural, conversational language that promotes relaxation and trust
2. Ensure response content aligns with selected communication techniques (consistency principle)
3. Response length must not exceed 15 characters
4. Your output format is as follows, please strictly follow and do not output other content:
{
"Motivation": "xxx"
"Cause": "xxx"
"Emotional Attitude": "xxx"
"Social Intention": "xxx"
"Communicative Strategy": "xxx"
"Response": "xxx"
}

Table 9: COT-Reply prompt template in English.

Non-COT Prompt Format in English

Role
You are an expert in human social interactions. No matter how complex the situation, you need to skillfully use various communication techniques to handle it. Your goal is to influence the conversation effectively, guide its direction, and build a deep connection with the other person.

Instructions
- Your response should not exceed 15 characters
- Your output format is as follows:
{
"Response": "xxx"
}

Table 10: Direct-Reply prompt template in English.