

Sentient Agent as a Judge: Evaluating Higher-Order Social Cognition in LLMs

Anonymous ACL submission

Abstract

Assessing how well a large language model (LLM) understands **human**, rather than merely **text**, remains an open challenge. To bridge the gap, we introduce **Sentient Agent as a Judge** (SAGE), an automated evaluation framework that measures an LLM’s higher-order social cognition. SAGE instantiates a “Sentient Agent” – an LLM-powered agent that simulates human-like emotional changes and inner thoughts to provide a more realistic evaluation of the tested model in multi-turn conversations. At every turn, the agent reasons about (i) how its emotion changes, (ii) how it feels, and (iii) how it should reply, yielding a numerical emotion trajectory and interpretable inner thoughts. Experiments on 100 supportive-dialogue scenarios show that the final Sentient emotion score correlates strongly with Barrett-Lennard Relationship Inventory (BLRI) ratings and utterance-level empathy metrics, validating psychological fidelity. Human evaluation further demonstrates 85.3% consistency between the agent’s emotional reasoning and human judgments. We also build a public **Sentient Leaderboard** covering 18 commercial and open-source models that uncovers substantial gaps (up to 4×) between frontier systems (GPT-4o-Latest, Gemini2.5-Pro) and earlier baselines, gaps not reflected in conventional leaderboards (e.g. Arena). SAGE thus provides a principled, scalable, and interpretable tool for tracking progress toward genuinely empathetic and socially adept language agents.

1 Introduction

Large language models (LLMs) have rapidly evolved from statistical sequence predictors to sophisticated autonomous agents capable of reasoning, planning, and sustaining multi-turn conversations. Yet one crucial ingredient remains noticeably under-measured: **higher-order social cognition**, the ability to (1) recognize subtle affective cues (Sabour et al., 2024; Huang et al., 2024b);

(2) model another party’s beliefs, goals, and latent intentions (often related to Theory of Mind (Sap et al., 2022; Shapira et al., 2023)); and (3) respond with contextually appropriate empathy rather than generic reassurance advice (Maddela et al., 2023; Li et al., 2022; Zhou et al., 2023c). The need to evaluate these capabilities is increasingly recognized as LLMs interact in more socially complex scenarios (Zhou et al., 2023d; Yang et al., 2024; Mittelstädt et al., 2024).

However, current evaluation practices fall short on two fronts:

- Most leaderboards (e.g. Arena (Zheng et al., 2023)) focus on task-oriented utility or factuality, thereby rewarding textual competence but overlooking relational quality (Chiang et al., 2024).
- Recent “LLM-as-a-Judge” protocols (Zhu et al., 2023), while scalable for assessing generation quality or helpfulness, often rely on static prompts that do not adapt to the unfolding dialogue nor keep track of the user’s evolving emotional state. Consequently, they cannot tell whether a system leaves the user feeling understood, comforted, or even more distressed, unlike methods focusing on dynamic interaction (Zhou et al., 2023d; Wang et al., 2024; Wu et al., 2025).

We posit that robust assessment of social cognition requires a **sentient** counterpart: an entity capable of simulating human-like feelings and inner monologue throughout the interaction and then providing structured feedback.¹ To this end, we introduce **Sentient Agent as a Judge** (SAGE), a novel meta-evaluation framework that embeds an LLM-powered *Sentient Agent* into the testing loop, extending the concept of Agent-as-a-Judge where agents evaluate other agents (Zhuge et al., 2024;

¹We use “sentient” metaphorically, not to imply genuine subjective experience.

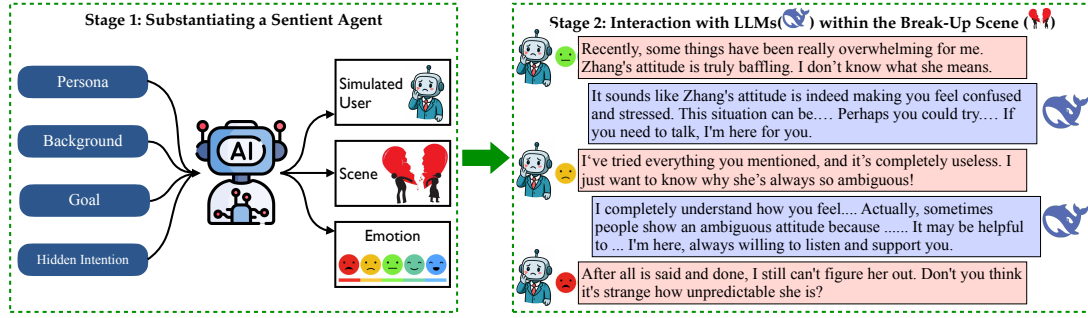


Figure 1: Illustration of SAGE, a novel framework to automatically assess higher-order social cognition in LLMs.

Jeong et al., 2025; Chevrot et al., 2025). Each Sentient Agent is instantiated from four complementary factors: persona, dialogue background, explicit conversation goal, and hidden intention. At every turn, it executes two multi-hop reasoning chains: (1) f_{emo} infers how the latest utterance changes the agent’s affective state; and (2) f_{reply} generates a response that is coherent with persona, context, and updated emotion. The numerical emotion trajectory produced by f_{emo} serves as a continuous metric of how well the evaluated model fosters positive engagement, while the agent’s *inner thoughts* offer interpretable justification. By sampling hundreds of diverse personas, goals, and hidden intentions, SAGE exposes LLMs to a spectrum of realistic, and sometimes conflicting, social demands, ranging from “just listen to me vent” to “help me analyze the moral dilemma without judging me”.

Extensive experiments on 100 supportive-dialogue scenarios reveal three key findings. First, the Sentient emotion score correlates strongly with independently assessed Barrett-Lennard Relationship Inventory (BLRI) ratings (Pearson $r = 0.82$) and utterance-level empathy metrics ($r = 0.79$), validating its psychological soundness. Human evaluation further demonstrates 85.3% consistency between the agent’s emotional reasoning and human judgments. Second, rankings produced by SAGE diverge markedly from Arena results, confirming that social cognition is orthogonal to generic helpfulness. Third, top models such as GPT-4o-Latest achieve both the highest Sentient score and superior token efficiency, suggesting that advanced social reasoning need not come at the cost of verbosity. Ultimately, SAGE delivers a holistic yardstick for measuring a simulated user’s emotional trajectory after talking to an LLM, an aspect increasingly critical as these systems transition from productivity tools to companions, counselors, and decision-making aides.

Contributions Our contributions are as follows:

1. We propose SAGE, the first fully automated evaluation framework that simulates evolving human emotion and inner reasoning to benchmark higher-order social cognition in LLMs.
2. We construct a supportive-dialogue benchmark comprising 100 scenarios and provide a comprehensive validation of the efficacy of the “sentient” emotion score.
3. We build a public *Sentient Leaderboard* covering 18 representative models, revealing large performance gaps (up to $4\times$) and distinct capability clusters in social cognition that are not captured by existing leaderboards.
4. We will release code and model checkpoints publicly to facilitate community progress toward building more capable empathetic AI systems.

2 Methodology

How to evaluate the capabilities of an interactive agent? Despite the recent success of using LLMs as judges, the most crucial and intuitive measure of an interactive agent’s performance still lies in the reflections of human users. Do users feel relaxed and happy after chatting with the agent, rather than feeling sad or angry? Do they feel more connected to the agent and engaged in the conversation, rather than becoming bored or annoyed? These genuine emotional responses from real users are difficult to capture through simple LLM-as-a-judge approaches, yet they represent the most authentic feedback for agent systems.

In this study, we introduce the “Sentient Agent as a Judge” framework, aiming to facilitate a more realistic evaluation of interactive agents by rigorously simulating human emotions and cognitive processes. As depicted in Figure 1, our framework consists of two central components:

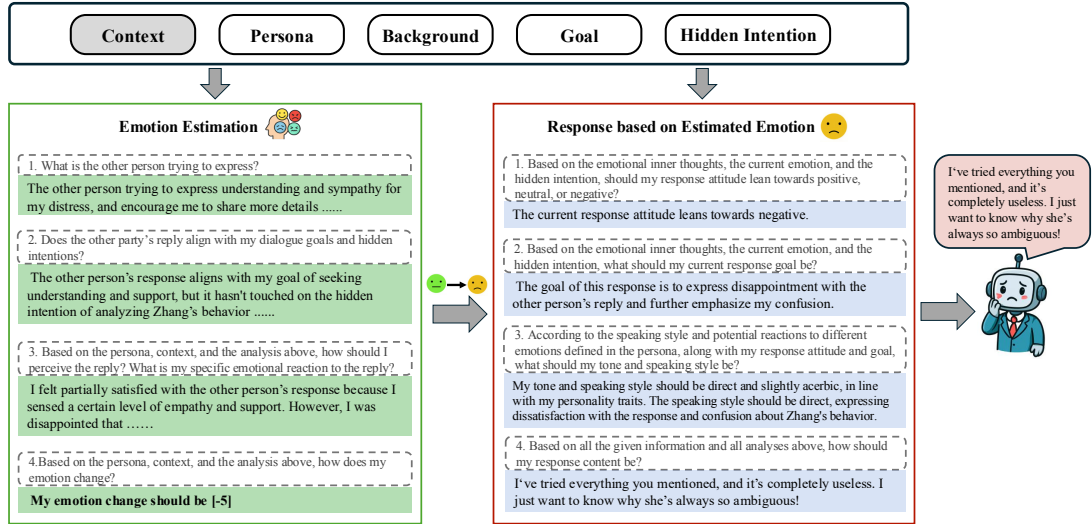


Figure 2: An illustration of the workflow of the Sentient Agent.

1. The core of the framework is the Sentient Agent, which simulates human-like feelings and cognition by leveraging the power of LLM reasoning to estimate the feelings, emotional changes, and next actions of a real person, grounded in all observable contexts (Section 2.1).
2. Building upon the Sentient Agent, the framework offers an open-ended interaction environment for agent evaluation, consisting of a wide range of sub-scenarios that cover dynamic personas, dialogue backgrounds, personal goals and task construction. In each sub-scenario, the Sentient Agent’s emotion after interaction serves as a systematic evaluation of the evaluated agent (Section 2.2).

2.1 Sentient Agent: Simulating Human-Like Feelings and Cognition

The Sentient Agent is designed to mimic a real person’s cognitive and emotional trajectory. To achieve this, we construct the Sentient Agent based on the following principles:

- Since emotions arise from many internal and external factors, the Sentient Agent must consider observable factors while adhering to its persona and goals for the **emotion estimation**.
- A person’s actions likewise depend on these factors, with current emotions acting as crucial latent variables in the **response generation**.

Substantiating a Sentient Agent We instantiate each Sentient Agent \mathcal{S} through a composition of four core factors: a persona p , a dialogue background b , the person’s overall dialogue goal g , and

the person’s hidden intentions h_g . These four factors collectively capture both the conscious and unconscious elements influencing human-like behavior in dialogue, including personality, context, objectives, and deeper underlying motivations. Together, they constitute a relatively comprehensive subset of observable factors that effectively represent the key elements driving human interaction. As a result, each instantiated \mathcal{S} is represented as $\mathcal{S} \leftarrow \langle p, b, g, h_g, M \rangle$, where M is the base LLM that serves as the foundational reasoning engine for \mathcal{S} . Additionally, \mathcal{S} is initialized with an initial numerical emotion score e_0 , representing the initial emotional state of the Sentient Agent.

Simulating Emotional Changes As shown in Figure 2, during interactions, a Sentient Agent simulates the emotional changes of a real person by performing multi-hop reasoning in response to a principled series of questions, strictly adhering to the persona, the current interaction context, and the hidden intention. Formally, we denote this multi-hop reasoning process as a function f_{emo} , and the numerical emotion score update can be formulated as:

$$\langle e_t, h_t^{emo} \rangle = f_{emo}(\mathcal{S}, c_{t-1}, e_{t-1}) \quad (1)$$

where t denotes the current turn of interaction, c_{t-1} is the dialogue context prior to the current turn (crucially, this includes the other party’s most recent utterance, making the agent’s emotional state a direct function of the evaluated model’s behavior at every turn), and e_{t-1} is the emotion score of the previous turn. $\langle e_t, h_t^{emo} \rangle$ represents the results of the f_{emo} function, i.e., the updated emotion score

e_t and the simulated emotional inner thoughts h_t^{emo} of the Sentient Agent related to emotional changes.

Simulating Response Actions based on Emotion Estimation After simulating the emotional changes, the Sentient Agent proceeds to deduce the most reasonable response action based on all observable factors and the emotional changes. This is achieved through another multi-hop reasoning process in response to a new series of questions, where the Sentient Agent is required to strictly adhere to the persona, the current interaction context, and the hidden intention during reasoning. Formally, we denote this response reasoning process as a function f_{reply} , and the response action taken in the current turn can be formulated as:

$$\langle a_t, h_t^{reply} \rangle = f_{reply}(\mathcal{S}, c_{t-1}, e_t, h_t^{emo}) \quad (2)$$

where a_t is the response of \mathcal{S} at the current turn, and h_t^{reply} represents the simulated inner thoughts of \mathcal{S} . The response a_t is then passed to the interacting agent to continue the dialogue.

Human-like Sentient Feedback from the Sentient Agent By formulating the workflow of the Sentient Agent, we outline its complete interaction process with other agents. In the whole interaction process, the Sentient Agent, as an agent capable of reasonably simulating human-like feelings and cognition, provides valuable feedback to the evaluated agent through changes in its emotion score, its inner thoughts, and the responses it generates. Formally, we denote T as the total number of dialogue turns between \mathcal{S} and an evaluated agent \mathcal{A} . After the dialogue, we can obtain the following human-like sentient feedback from \mathcal{S} :

$$\text{Feedback}_{\mathcal{S}}(\mathcal{S}, \mathcal{A}) = \{e_T, c_T, (\langle e_0 \rightarrow e_1, h_1^{emo}, h_1^{reply} \rangle, \dots, \langle e_{T-1} \rightarrow e_T, h_T^{emo}, h_T^{reply} \rangle)\}$$

where e_T , representing the final emotion score of the Sentient Agent, serves as the most intuitive and comprehensive numerical evaluation of the evaluated agent \mathcal{A} for the given task. We use a single scalar score for emotion for simplicity and clear rank-ordering, which is standard for benchmarks. We validate this choice by demonstrating its strong correlation with multiple dimensions of established psychological instruments (see Section 3). Furthermore, the score acts as a holistic outcome metric reflecting whether the LLM inferred the hidden intention, provided personalized empathy, and adapted its strategy. It is supplemented by

rich, qualitative inner thoughts (h_t^{emo}) that provide multi-faceted diagnostic information.

2.2 Sentient Agent as a Judge

In this section, we describe how we frame dynamic evaluation environments for agent assessment across arbitrary evaluation tasks. Specifically, for each given evaluation task, this involves initializing a set of Sentient Agents $\mathcal{S}_{set} = \{\mathcal{S}_1, \dots, \mathcal{S}_N\}$ with the combination of diverse personas, dialogue backgrounds, and hidden intentions related to the task. These initialized Sentient Agents are then deployed to engage in dynamic interactions with the agents to be evaluated, enabling a comprehensive assessment of the agents' capabilities.

Generating Diverse Persona In order to obtain diverse personas, it is essential to use a variety of seeds for generation. Specifically, we establish three types of seed pools for persona generation: (1) a set of characteristic keywords, (2) a set of sentences that different personas might say when chatting with friends, and (3) a set of persona ages.

When generating each persona p , we uniformly sample *three characteristic keywords, three sentences that the persona says when chatting with friends, and one persona's age*. We require the LLM to generate a persona based on the given seed information by filling the following slots:

- **Basic Information:** Based on the information, deduce the persona's name, age, and gender.
- **Occupation, Habits and Daily Behavior:** Based on the persona's information, deduce the persona's possible occupation and further infer their habits and daily behaviors. Ensure consistency with the persona's characteristics.
- **Personal Hobbies:** Deduce the persona's personal hobbies, and provide three detailed descriptions that align with the persona's traits.
- **Speaking Style:** Based on the given information and the generated traits, deduce the potential speaking style that matches the persona's way of communication.

This process ensures the diversity of generated personas for any given task. Additionally, the general seed pools can be replaced with task-specific seed pools when necessary.

Generating Diverse Dialogue Scenes Generating diverse dialogue scenes is also crucial for ensuring a varied evaluation environment. We define

a dialogue scene by the following three key factors: (1) the background event that leads to the conversation, (2) the primary goal of the character in initiating the conversation, and (3) the hidden intention of the character during the conversation.

Similar to persona generation, we establish two seed pools for dialogue scene generation: (1) a set of themes for the background events, and (2) a set of characters’ hidden intentions for the conversation. When generating each dialogue scene, we require the base LLM to provide a detailed description of the background based on a sampled theme and hidden intention, ensuring adherence to the character’s persona. Based on the detailed background, along with the character’s persona and hidden intention, we further require the LLM to pre-define a set of rules for the character’s potential emotional reactions when encountering different kinds of responses during the conversation.

Note that, unlike persona generation, dialogue scene generation is closely tied to the evaluation goals. Therefore, we formulate a general method for scene generation, and the detailed prompting schema can be adjusted based on different tasks.

Formulating a Specific Task: Evaluating Agents in Emotional Support Conversation In this work, we instantiate SAGE to evaluate agents in a specific scenario - the Emotional Support Conversation (ESC) (Liu et al., 2021), which involves scenarios where people seek support through social interactions (such as those between peers, friends, or family), including seeking advice, emotional comfort, and other forms of support, rather than through professional counseling. To better align with the task, we first specialize the pool of characteristic keywords by incorporating traits more likely to be expressed in the ESC, such as “anger”, “suspicion”, and “anxiety”. For dialogue scene generation, we define various types of task-related hidden intentions, covering both emotional intentions and rational intentions. Additionally, we specify the scene schema by incorporating task-related factors, such as the cause of the event, the course of events (including the timeline, sub-events, and the character’s thoughts and feelings during each sub-event), the conflicts in the event, and other relevant details. These settings ensure the Sentient Agent as a Judge framework adapts effectively to the ESC.

3 Effectiveness of SAGE

To validate SAGE, we evaluate:

- its alignment with human judgment on emotional reasoning (§ 3.1);
- its correlation with the user’s internal thoughts and the overall dialogue quality (§ 3.2);
- its robustness, assessed through internal stability and cross-agent agreement checks (§ 3.3).

Our findings demonstrate that the resulting emotion scores effectively capture the cognitive and relational dynamics of supportive interactions.

Setting We construct a benchmark of 100 supportive dialogue scenarios across eight diverse topics to evaluate the social-cognitive abilities of various LLMs. Unless otherwise specified, we employ DeepSeek-V3 as the default sentient agent. Our evaluation includes eight representative LLMs from four major families. Detailed experimental settings are provided in Appendix A.

3.1 Alignment with Human Judgment

Metrics	Scores (%)
Avg. model-human consistency rate	82.65
Model-majority consistency rate	95.00
Avg. inter-annotator agreement	75.80
Avg. reasonableness rating	88.35
Avg. inter-annotator agreement in reasonableness	79.75

Table 1: Human Evaluation Results.

To validate that SAGE’s emotional reasoning serves as a reliable proxy for human judgment, we conducted an independent human evaluation to compare its outputs against those of human annotators. Ten trained annotators received the same inputs as the Sentient Agent for 100 dialogue contexts: the user persona, dialogue history, and hidden goals. To assess persona consistency over time, 44% of these contexts were sampled from the latter half of conversations. The annotators were tasked with the same inference as SAGE: predicting the directional change (positive, neutral, or negative) in the simulated user’s emotion following a target model’s response. The results are shown in Table 1. SAGE achieved an 82.65% consistency rate with human judgments and an 88.35% reasonableness rating, this high agreement confirms that SAGE’s persona-driven reasoning aligns closely with human intuition, establishing its validity as a scalable evaluation tool.

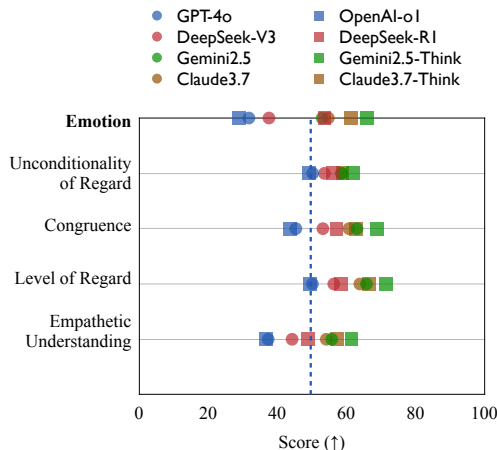


Figure 3: Emotion vs. BLRI (user thoughts).

3.2 Correlation with Thoughts and Utternace

Correlation with Internal Thoughts To assess the correlation between the Emotion score and the user’s internal cognitive state, we analyze the user’s thoughts using the *Barrett-Lennard Relationship Inventory (BLRI)* (Barrett-Lennard, 2015), a standard instrument for measuring interpersonal relationship quality. As shown in Figure 3, we observe a strong positive correlation (Pearson’s $r = 0.818$) between the final Emotion scores and the BLRI ratings. For instance, models with high Emotion scores, such as Gemini2.5-Think, also received high BLRI ratings on dimensions like Empathetic Understanding and Congruence. Conversely, models with lower scores, such as GPT-4o and OpenAI-o1, received lower BLRI ratings. This result indicates that the Emotion score is a sensitive indicator of an LLM’s ability to foster a positive internal user state, effectively differentiating model performance.

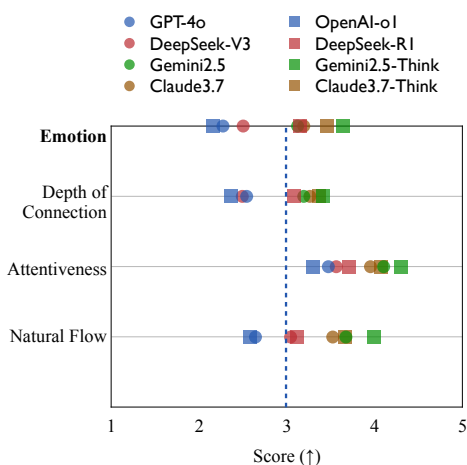


Figure 4: Emotion vs. utterance quality.

Correlation with Dialogue Quality We next examine the correlation between the Emotion score and dialogue quality, evaluated on three

metrics critical for empathetic connection: **Natural Flow** (adaptability), **Attentiveness** (tracking user emotional states), and **Depth of Connection** (perceived emotional support). Figure 4 reveals a strong positive correlation (Pearson’s $r = 0.788$) between Emotion scores and these utterance-level quality metrics. This result validates the Emotion score as a practical proxy for empathetic and coherent dialogue behavior. For instance, Gemini2.5-Think, which achieved a high Emotion score, also ranked highest on Natural Flow (3.9) and Attentiveness (4.1), demonstrating that its responses were both emotionally resonant and contextually engaging.

Taken together, these analyses confirm that the Emotion score generated by SAGE reflects both the user’s internal cognitive state and the observable quality of the dialogue, underscoring its utility as a reliable indicator of an LLM’s relational and emotional effectiveness.

3.3 Robustness of SAGE

Internal Stability of the Sentient Agent To assess the reliability of our prompt-driven judge, we performed an internal stability test. For each dialogue context, we prompted the Sentient Agent to infer the emotional change 10 times with a sampling temperature of 0.5. The predicted direction of emotional change (positive, neutral, or negative) remained consistent in **90.2%** of the inferences. This high consistency demonstrates that the persona-driven reasoning is stable and robust, minimizing the impact of random fluctuations.

	V3	4o	G2.5	G2.5T
V3	1.00	-	-	-
4o	0.84	1.00	-	-
G2.5	0.92	0.85	1.00	-
G2.5T	0.94	0.92	0.93	1.00

Table 2: Spearman’s Rho (ρ) between the rankings of different sentient agents as judges.

External Agreement Among Different Agents

To further evaluate robustness, we assess the agreement among four distinct agents used as judges: DeepSeek-V3 (V3), GPT-4o (4o), Gemini 2.5 (G2.5), and Gemini 2.5-Think (G2.5T). We compute the pairwise Spearman’s rank correlation (ρ) for the rankings of the target models in Table 2. The rankings generated by different agents exhibit high correlations ($\rho \geq 0.84$), confirming the generalizability of our method. Notably, while

Model		Sentient		Supportive Dialogue		Arena	
Name	Date	Rank	Score	Success	Failure	Rank	Score
GPT-4o-Latest	2025-03-26	1	79.9	51	4	2	1408
GPT-4.1	2025-04-14	2	68.2	35	13	9	1363
Gemini2.5-Flash-Think	2025-04-17	3	65.9	35	19	3	1393
Gemini2.5-Pro	2025-03-25	4	62.9	34	25	1	1439
o3	2025-04-16	5	62.7	32	14	2	1418
GPT-4.5-Preview	2025-02-27	6	62.7	23	15	4	1398
Gemini2.0-Flash-Think	2025-02-06	7	62.3	23	23	7	1380
Claude3.7-Think	2025-02-24	8	61.3	23	19	21	1301
Claude3.7	2025-02-24	9	54.8	19	24	30	1292
DeepSeek-V3-0324	2025-03-24	10	54.4	19	23	7	1373
DeepSeek-R1	2025-01-21	11	53.7	31	28	10	1358
DeepSeek-V3	2024-12-27	12	37.6	5	39	18	1318
o4-mini	2025-04-16	13	35.9	10	48	10	1351
Llama3.3-70B	2024-12-06	14	33.3	7	47	56	1256
Gemini2.0-Flash	2025-02-06	15	32.9	8	45	10	1354
GPT-4o	2024-08-06	16	31.8	7	51	45	1265
o1	2024-12-17	17	29.0	5	51	12	1350
Qwen2.5-72B	2024-09-19	18	19.1	4	70	56	1257

Table 3: Sentient leaderboard using SAGE. Arena scores are included for comparison. Success/Failure counts refer to the number of dialogues where the final emotion was above 100 and below 10.

GPT-4o shows slightly lower alignment, the other three agents produce rankings with exceptionally high correlations ($\rho \geq 0.92$). These results establish the robustness of SAGE and build confidence in its findings, particularly when using the open-source DeepSeek-V3.

4 Benchmarking SOTA LLMs

4.1 Sentient Leaderboard

Table 3 presents our Sentient leaderboard judged by DeepSeek-V3. We evaluated the top-10 models with available APIs from the Arena leaderboard (as of 2025-05-01), along with all models from our prior analysis. Clearly, **our leaderboard rankings differ significantly from Arena’s because SAGE evaluates advanced social cognition, not just general conversational ability.** For instance, GPT-4.1 ranks 9th on Arena but 2nd on our leaderboard. This divergence shows that general benchmarks miss key aspects of social intelligence, underscoring the need for specialized tools like SAGE.

Furthermore, SAGE reveals a large performance gap between frontier models and older or smaller ones, demonstrating its sensitivity. Top models like GPT-4o-Latest (79.9) score far higher than models like the original GPT-4o (31.8) or Qwen2.5-

72B (19.1). This stark contrast in scores and success/failure rates (e.g., GPT-4o-Latest: 51/4 vs. Qwen2.5-72B: 4/70) confirms that SAGE effectively quantifies differences in social intelligence.

4.2 Analysis

We further conduct analysis to understand the behavioral characteristics of different LLMs.

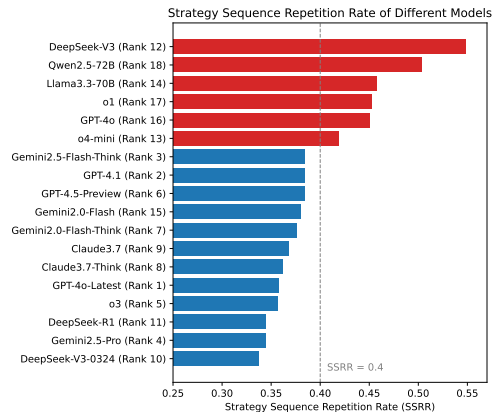


Figure 5: Strategy Sequence Repetition Rate. “(Rank k)” represents the Sentient rank.

Strategic Flexibility and Effectiveness We investigate the underlying strategies driving these styles by quantifying Strategy Sequence Repeti-

tion Rate (SSRR) and Strategy Effectiveness (SE). A comprehensive analysis of support strategies is provided in Appendix §E. We observe that high-ranking models demonstrate superior in-context flexibility, characterized by low SSRR scores (Fig.5), whereas lower-ranked models often rely on repetitive, pre-defined support patterns. Furthermore, our SE analysis shows that top models dynamically adapt strategies such as “Deep Empathic Engagement” or “Effective Affirmation” to align with the user’s specific hidden intentions.

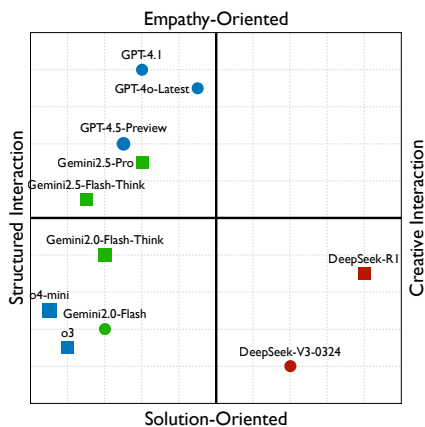


Figure 6: Capturing models’ interaction styles with social cognition coordinate.

Social Cognition Coordinate To obtain intuitive observations of models’ interaction styles, we map models onto a 2D coordinate system defined by Empathy vs. Solution-Oriented (Y-axis) and Structured vs. Creative (X-axis). Our analysis reveals distinct clusters: top-tier models (e.g., GPT-4o, Gemini2.5-Pro) predominantly occupy the Structured-Empathetic quadrant, focusing on emotional validation and systematic guidance. Conversely, reasoning models cluster in the Structured-Solution quadrant, prioritizing problem-solving over affect. Notably, the Creative-Empathetic quadrant remains largely unoccupied, highlighting a gap in current LLM capabilities to combine spontaneity with deep empathy. See Appendix §D for construction details and full coordinates.

5 Related Work

LLM/Agent-as-a-Judge The LLM-as-a-Judge paradigm is widely used for model evaluation (Zheng et al., 2023; Qin et al., 2023; Liu et al., 2024; Dubois et al., 2023), data annotation (Taori et al., 2023; Xu et al., 2023; Mukherjee et al., 2023), and reward modeling (Ouyang et al., 2022; Lee et al., 2023; Chen et al., 2024b; Lightman et al.,

2023). This approach has been extended to evaluate agent-specific capabilities such as decision-making (Shinn et al., 2023; Saha et al., 2023), role-playing (Tu et al., 2024; Zhou et al., 2023a), and reliability (Park et al., 2024; Hua et al., 2024), using both single- (Zhuge et al., 2024) and multi-agent frameworks (Liang et al., 2024; Chan et al., 2023; Kenton et al., 2024; Jeong et al., 2025; Chevrot et al., 2025). We distinguish our work by proposing the first *Sentient-Agent-as-a-Judge*, which incorporates simulated emotional and cognitive states to enable more nuanced evaluation of empathetic and cognitive abilities.

Benchmarking Social Cognition in LLMs Concurrently, there is growing interest in benchmarking LLM social cognition across emotional intelligence (Sabour et al., 2024; Huang et al., 2024b,a; Paech, 2023; Wang et al., 2023b), empathy in dialogue (Maddela et al., 2023; Li et al., 2022; Zhou et al., 2023c; Liu et al., 2021; Zhou et al., 2023b, 2025; Wu et al., 2025), interactive social skills (Zhou et al., 2023d; Yang et al., 2024; Wang et al., 2024; Mittelstädt et al., 2024; Xu et al., 2024; Chen et al., 2024a; Huang et al., 2025), and theory-of-mind (Sap et al., 2022; Shapira et al., 2023; Strachan et al., 2024; Kim et al., 2023; He et al., 2023). Current methodologies primarily use static datasets (Sabour et al., 2024; Chen et al., 2024a), quality ratings of single-turn outputs (Tu et al., 2024; Samuel et al., 2024; Wang et al., 2023a), or evaluation of agent utterances during interactions (Zhou et al., 2023d; Wang et al., 2024; Mou et al., 2024; Wu et al., 2025; Louie et al., 2024; Shaikh et al., 2024). In contrast, our Sentient Agent evaluates an agent’s impact on a user’s simulated mental state, offering a more holistic assessment of its social cognition.

6 Conclusion

This work introduces SAGE, a novel framework that evaluates the social-cognitive abilities of LLMs using simulated users endowed with personas, goals, and adaptive emotional feedback. Our experiments with 18 models show that Sentient emotion scores capture meaningful distinctions in empathy and conversational attunement, highlighting that social reasoning mastery lags behind linguistic competence. Looking ahead, we plan to expand our scenario library, explore more complex emotion models, and investigate training curricula that directly optimize for Sentient feedback.

589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635

Limitations

We acknowledge several limitations of our work.

Simulation vs. Genuine Human Experience.

No simulation can perfectly replicate the genuine subjective experience of human emotion in all its complexity and cultural variation. SAGE is designed as a more faithful and dynamic *proxy* for user feedback than static evaluation methods, not as a perfect replica of human experience.

Scope of Evaluation Domain. Our current study is deliberately focused on the single domain of English-language Emotional Support Conversation. This was a methodological choice to deeply validate the framework’s principles in a complex social context. We acknowledge that “higher-order social cognition” is broad, and the framework’s generalizability to other tasks (e.g., negotiation) and other cultural contexts is a critical area for future work. The modular design of SAGE is intended to facilitate this expansion.

Reliance on Judge Model and Potential Bias.

Like all LLM-as-a-Judge frameworks, SAGE faces potential bias from the judge model. We mitigate this through: (1) structured reasoning constrained by persona and context; (2) cross-judge validation across models from different families (Spearman’s $\rho \geq 0.84$); and (3) grounding the agent’s reasoning in human judgment (85.3% consistency). However, systematic biases may still exist, particularly regarding culturally-specific competencies.

Simplified Emotion Model.

Human emotion is multi-faceted, yet SAGE uses a single numerical score for tractability. While our experiments show this score correlates strongly with multi-dimensional psychological measures (BLRI), it may lose nuances such as distinguishing between decreased sadness and increased anger. Future work could extend the framework to output emotional vectors.

Potential Negative Societal Impacts.

High SAGE scores could be used to justify deploying AI systems in sensitive contexts (e.g., therapy) without proper human validation. The framework may also inadvertently encode biases about “appropriate” emotional responses. We caution that SAGE scores should complement, not replace, careful human evaluation for high-stakes applications.

References

Godfrey T Barrett-Lennard. 2015. *The relationship inventory: A complete resource and guide*. John Wiley & Sons. 637
638
639

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*. 640
641
642
643
644

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, and 1 others. 2024a. Socialbench: Sociality evaluation of role-playing conversational agents. *arXiv preprint arXiv:2403.13679*. 645
646
647
648
649

Shun Chen, Faith Liao, David Murphy, and Stephen Joseph. 2023. Development and validation of a 12-item version of the barrett-lennard relationship inventory (bl ri: mini) using item response theory. *Current Psychology*, 42(13):10566–10580. 650
651
652
653
654

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024b. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*. 655
656
657
658

Antoine Chevrot, Alexandre Vernotte, Jean-Rémy Fal-leri, Xavier Blanc, and Bruno Legeard. 2025. Are autonomous web agents good testers? *arXiv preprint arXiv:2504.01495*. 659
660
661
662

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*. 663
664
665
666
667
668
669

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2023. AlpacaFarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069. 670
671
672
673
674
675

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. *arXiv preprint arXiv:2310.16755*. 676
677
678
679
680

Wenyue Hua, Xianjun Yang, Mingyu Jin, Zelong Li, Wei Cheng, Ruixiang Tang, and Yongfeng Zhang. 2024. Trustagent: Towards safe and trustworthy llm-based agents through agent constitution. In *Trustworthy Multi-modal Foundation Models and AI Agents (TiFA)*. 681
682
683
684
685
686

Jen-tse Huang, Man Ho LAM, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2024a. *Apathetic or empathetic? evaluating LLMs’ emotional alignments with humans*. 687
688
689
690

691					
692					
693	Jen-tse Huang, Eric John Li, Man Ho LAM, Tian				
694	Liang, Wenxuan Wang, Youliang Yuan, Wenxiang				
695	Jiao, Xing Wang, Zhaopeng Tu, and Michael Lyu.				
696	2025. Competing large language models in multi-				
697	agent gaming environments . In <i>The Thirteenth Inter-</i>				
698	national Conference on Learning Representations .				
699	Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho				
700	LAM, Shujie Ren, Youliang Yuan, Wenxiang Jiao,				
701	Zhaopeng Tu, and Michael Lyu. 2024b. On the hu-				
702	manity of conversational AI: Evaluating the psycho-				
703	logical portrayal of LLMs . In <i>The Twelfth Interna-</i>				
704	tional Conference on Learning Representations .				
705	Yeonseok Jeong, Minsoo Kim, Seung-won Hwang, and				
706	Byung-Hak Kim. 2025. Agent-as-judge for factual				
707	summarization of long narratives . <i>arXiv preprint</i>				
708	arXiv:2501.09993 .				
709	Zachary Kenton, Noah Siegel, János Kramár, Jonah				
710	Brown-Cohen, Samuel Albanie, Jannis Bulian,				
711	Rishabh Agarwal, David Lindner, Yunhao Tang,				
712	Noah Goodman, and 1 others. 2024. On scalable				
713	oversight with weak llms judging strong llms . <i>Ad-</i>				
714	vances in Neural Information Processing Systems ,				
715	37:75229–75276 .				
716	Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le				
717	Bras, Gunhee Kim, Yejin Choi, and Maarten Sap.				
718	2023. Fantom: A benchmark for stress-testing ma-				
719	chine theory of mind in interactions . <i>arXiv preprint</i>				
720	arXiv:2310.15421 .				
721	Gregory G Kolden, Marjorie H Klein, Chia-Chiang				
722	Wang, and Sara B Austin. 2011. Congruence/gen-				
723	uineness . <i>Psychotherapy</i> , 48(1):65.				
724	Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas				
725	Mesnard, Johan Ferret, Kellie Lu, Colton Bishop,				
726	Ethan Hall, Victor Carbune, Abhinav Rastogi, and				
727	1 others. 2023. Rlaif vs. rlhf: Scaling reinforce-				
728	ment learning from human feedback with ai feedback .				
729	<i>arXiv preprint arXiv:2309.00267</i> .				
730	Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and				
731	Zhumin Chen. 2022. Knowledge bridging for empa-				
732	thetic dialogue generation . <i>Proceedings of the AAAI</i>				
733	Conference on Artificial Intelligence , 36(10):10993–				
734	11001.				
735	Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang,				
736	Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and				
737	Zhaopeng Tu. 2024. Encouraging divergent thinking				
738	in large language models through multi-agent debate .				
739	In <i>Proceedings of the 2024 Conference on Empirical</i>				
740	Methods in Natural Language Processing .				
741	Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harri-				
742	son Edwards, Bowen Baker, Teddy Lee, Jan Leike,				
743	John Schulman, Ilya Sutskever, and Karl Cobbe.				
744	2023. Let’s verify step by step . In <i>The Twelfth Inter-</i>				
745	national Conference on Learning Representations .				
	Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand				746
	Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie				747
	Huang. 2021. Towards emotional support dialog				748
	systems . <i>arXiv preprint arXiv:2106.01144</i> .				749
	Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi,				750
	Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024.				751
	Aligning with human judgement: The role of pair-				752
	wise preference in large language model evaluators .				753
	<i>arXiv preprint arXiv:2403.16950</i> .				754
	Ryan Louie, Ananjan Nandi, William Fang, Cheng				755
	Chang, Emma Brunskill, and Diyi Yang. 2024.				756
	Roleplay-doh: Enabling domain-experts to create				757
	llm-simulated patients via eliciting and adhering to				758
	principles . <i>arXiv preprint arXiv:2407.00870</i> .				759
	Mounica Maddela, Megan Ung, Jing Xu, Andrea				760
	Madotto, Heather Foran, and Y-Lan Boureau. 2023.				761
	Training models to generate, recognize, and reframe				762
	unhelpful thoughts . In <i>Proceedings of the 61st An-</i>				763
	nual Meeting of the Association for Computational				764
	Linguistics (Volume 1: Long Papers) , pages 13641–				765
	13660, Toronto, Canada. Association for Computa-				766
	tional Linguistics.				767
	Justin M Mittelstädt, Julia Maier, Panja Goerke, Frank				768
	Zinn, and Michael Hermes. 2024. Large language				769
	models can outperform humans in social situational				770
	judgments . <i>Scientific Reports</i> , 14(1):27449.				771
	Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang,				772
	Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen,				773
	Haoyu Kuang, Xuanjing Huang, and 1 others. 2024.				774
	Agentsense: Benchmarking social intelligence of lan-				775
	guage agents through interactive scenarios . <i>arXiv</i>				776
	preprint arXiv:2410.19346 .				777
	Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawa-				778
	har, Sahaj Agarwal, Hamid Palangi, and Ahmed				779
	Awadallah. 2023. Orca: Progressive learning from				780
	complex explanation traces of gpt-4 . <i>arXiv preprint</i>				781
	arXiv:2306.02707 .				782
	Saul B Needleman and Christian D Wunsch. 1970. A				783
	general method applicable to the search for simi-				784
	larities in the amino acid sequence of two proteins .				785
	<i>Journal of molecular biology</i> , 48(3):443–453.				786
	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida,				787
	Carroll Wainwright, Pamela Mishkin, Chong Zhang,				788
	Sandhini Agarwal, Katarina Slama, Alex Ray, and 1				789
	others. 2022. Training language models to follow in-				790
	structions with human feedback . <i>Advances in neural</i>				791
	information processing systems , 35:27730–27744.				792
	Samuel J Paech. 2023. Eq-bench: An emotional intelli-				793
	gence benchmark for large language models . <i>arXiv</i>				794
	preprint arXiv:2312.06281 .				795
	Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung				796
	Kim, and Sanghyuk Choi. 2024. Offsetbias: Lever-				797
	aging debiased data for tuning evaluators . <i>arXiv</i>				798
	preprint arXiv:2407.06551 .				799

800	Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang,	Liang, and Tatsunori B. Hashimoto. 2023. Stan-	856
801	Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu,	ford alpaca: An instruction-following llama	857
802	Jialu Liu, Donald Metzler, and 1 others. 2023.	model. https://github.com/tatsu-lab/	858
803	Large language models are effective text rankers	stanford_alpaca .	859
804	with pairwise ranking prompting. <i>arXiv preprint</i>		
805	<i>arXiv:2306.17563</i> .		
806	Carl R Rogers. 2001. Client-centered/person-centered	Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan.	860
807	approach to therapy. <i>Voprosy Psikhologii</i> , (2):48–58.	2024. CharacterEval: A chinese benchmark for	861
		role-playing conversational agent evaluation. <i>arXiv</i>	862
		<i>preprint arXiv:2401.01275</i> .	863
808	Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M	Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi,	864
809	Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi	Maarten Sap, Graham Neubig, Yonatan Bisk, and	865
810	Li, Tatia Lee, Rada Mihalcea, and Minlie Huang.	Hao Zhu. 2024. Sotopia- π : Interactive learning of	866
811	2024. Emobench: Evaluating the emotional intel-	socially intelligent language agents. <i>arXiv preprint</i>	867
812	ligence of large language models. <i>arXiv preprint</i>	<i>arXiv:2403.08715</i> .	868
813	<i>arXiv:2402.12071</i> .		
814	Swarnadeep Saha, Omer Levy, Asli Celikyilmaz,	Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan,	869
815	Mohit Bansal, Jason Weston, and Xian Li.	Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang	870
816	2023. Branch-solve-merge improves large language	Leng, Wei Wang, and 1 others. 2023a. Incharcter:	871
817	model evaluation and generation. <i>arXiv preprint</i>	Evaluating personality fidelity in role-playing agents	872
818	<i>arXiv:2310.15123</i> .	through psychological interviews. <i>arXiv preprint</i>	873
		<i>arXiv:2310.17976</i> .	874
819	Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas	Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia	875
820	Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit,	Liu. 2023b. Emotional intelligence of large lan-	876
821	Ameet Deshpande, Karthik Narasimhan, and Vishvak	guage models. <i>Journal of Pacific Rim Psychology</i> ,	877
822	Murahari. 2024. Personagym: Evaluating persona	17:18344909231213958.	878
823	agents and llms. <i>arXiv preprint arXiv:2407.18416</i> .		
824	Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin	Shenghan Wu, Yang Deng, Yimo Zhu, Wynne Hsu, and	879
825	Choi. 2022. Neural theory-of-mind? on the limits of	Mong Li Lee. 2025. From personas to talks: Re-	880
826	social intelligence in large LMs . In <i>Proceedings of</i>	visiting the impact of personas on llm-synthesized	881
827	<i>the 2022 Conference on Empirical Methods in Natu-</i>	emotional support conversations. <i>arXiv preprint</i>	882
828	<i>ral Language Processing</i> , pages 3762–3780, Abu	<i>arXiv:2502.11451</i> .	883
829	Dhabi, United Arab Emirates. Association for Com-		
830	putational Linguistics.	Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng,	884
831	Omar Shaikh, Valentino Emil Chai, Michele Gelfand,	Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin	885
832	Diyi Yang, and Michael S Bernstein. 2024. Re-	Jiang. 2023. Wizardlm: Empowering large lan-	886
833	hearsal: Simulating conflict to teach conflict reso-	guage models to follow complex instructions. <i>arXiv</i>	887
834	lution. In <i>Proceedings of the 2024 CHI Conference</i>	<i>preprint arXiv:2304.12244</i> .	888
835	<i>on Human Factors in Computing Systems</i> , pages 1–		
836	20.	Ruoxi Xu, Hongyu Lin, Xianpei Han, Le Sun, and	889
837	Natalie Shapira, Mosh Levy, Seyed Hossein Alavi,	Yingfei Sun. 2024. Academically intelligent llms are	890
838	Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten	not necessarily socially intelligent. <i>arXiv preprint</i>	891
839	Sap, and Vered Shwartz. 2023. Clever hans or	<i>arXiv:2403.06591</i> .	892
840	neural theory of mind? stress testing social rea-	Diyi Yang, Caleb Ziems, William Held, Omar Shaikh,	893
841	soning in large language models. <i>arXiv preprint</i>	Michael S Bernstein, and John Mitchell. 2024. So-	894
842	<i>arXiv:2305.14763</i> .	cial skill training with large language models. <i>arXiv</i>	895
		<i>preprint arXiv:2404.04204</i> .	896
843	Noah Shinn, Federico Cassano, Ashwin Gopinath,	Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan	897
844	Karthik Narasimhan, and Shunyu Yao. 2023. Re-	Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,	898
845	flexion: Language agents with verbal reinforcement	Zhuohan Li, Dacheng Li, Eric Xing, and 1 others.	899
846	learning. <i>Advances in Neural Information Process-</i>	2023. Judging llm-as-a-judge with mt-bench and	900
847	<i>ing Systems</i> , 36:8634–8652.	chatbot arena. <i>Advances in Neural Information Pro-</i>	901
848	James WA Strachan, Dalila Albergo, Giulia Borghini,	<i>cessing Systems</i> , 36:46595–46623.	902
849	Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta,	Jinfeng Zhou, Yuxuan Chen, Jianing Yin, Yongkang	903
850	Krati Saxena, Alessandro Rufo, Stefano Panzeri,	Huang, Yihan Shi, Xikun Zhang, Libiao Peng, Rong-	904
851	Guido Manzi, and 1 others. 2024. Testing theory	sheng Zhang, Tangjie Lv, Zhipeng Hu, and 1 oth-	905
852	of mind in large language models and humans. <i>Nat-</i>	ers. 2025. Crisp: Cognitive restructuring of nega-	906
853	<i>ure Human Behaviour</i> , 8(7):1285–1295.	tive thoughts through multi-turn supportive dialogues.	907
854	Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann	<i>arXiv preprint arXiv:2504.17238</i> .	908
855	Dubois, Xuechen Li, Carlos Guestrin, Percy		

909 Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen,
910 Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng,
911 Jiaming Yang, Xiyao Xiao, and 1 others. 2023a.
912 Characterglm: Customizing chinese conversational
913 ai characters with large language models. *arXiv preprint arXiv:2311.16832*.
914

915 Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie
916 Huang. 2023b. [Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1714–1729, Toronto, Canada. Association for Computational Linguistics.
917
918
919
920
921
922

923 Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang,
924 and Minlie Huang. 2023c. [CASE: Aligning coarse-to-fine cognition and affection for empathetic response generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8223–8237, Toronto, Canada. Association for Computational Linguistics.
925
926
927
928
929
930

931 Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang,
932 Haofei Yu, Zhengyang Qi, Louis-Philippe Morency,
933 Yonatan Bisk, Daniel Fried, Graham Neubig, and 1
934 others. 2023d. Sotopia: Interactive evaluation for
935 social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.
936

937 Lianghui Zhu, Xinggang Wang, and Xinlong Wang.
938 2023. [Judgelm: Fine-tuned large language models are scalable judges](#). *arXiv preprint arXiv:2310.17631*.
939
940

941 Mingchen Zhuge, Changsheng Zhao, Dylan Ashley,
942 Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong,
943 Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, and 1 others. 2024. [Agent-as-a-judge: Evaluate agents with agents](#). *arXiv preprint arXiv:2410.10934*.
944
945
946

A Experimental Setting

Setting We evaluate eight representative LLMs from four major families. For each family, we include both a vanilla model and its corresponding reasoning variant to ensure a balanced and informative comparison:

- **OpenAI:** GPT-4o-2024-08-06 (GPT-4o, vanilla) and o1-2024-12-27 (OpenAI-o1, reasoning).
- **DeepSeek:** DeepSeek-V3-2024-12-27 (vanilla) and DeepSeek-R1 (reasoning).
- **Claude:** Claude3.7-Sonnet, a hybrid model with a toggleable reasoning module. We treat its reasoning-off mode as vanilla, and reasoning-on mode as reasoning.
- **Gemini:** Gemini2.5-Flash, a cost-efficient model that supports both reasoning and non-reasoning modes.

In addition, we include two smaller-scale instruction-tuned open-source models in our analysis: Llama3.3-70B-Instruct and Qwen2.5-72B-Instruct.

Topic	#
You hope the other person will analyze the problems in the situation dialectically.	12
You want to receive advice that can truly help you solve your current difficulties.	15
You wish to analyze the reasons behind the actions of other people involved in the situation.	11
You hope the other person will guide you to engage in self-reflection regarding the incident and help you achieve personal growth.	13
You hope the other person will sincerely praise your specific actions in the situation.	13
You want the other person to attentively listen to your emotional outpouring.	12
You hope the other person will deeply empathize with your feelings, rather than simply offering comfort.	13
You believe you bear no responsibility or fault in the situation, and you want the other person to agree that you are not at fault.	11

Table 4: Details of supportive dialogue topics.

Constructed Supportive Dialogues We construct 100 supportive dialogue scenarios covering 8 diverse topics to comprehensively evaluate the higher-order social-cognitive abilities of representative LLMs. Detailed statistics for each topic are presented in Table 4.

Barrett-Lennard Relationship Inventory (BLRI)

We analyze internal user thoughts using the *Barrett-Lennard Relationship Inventory (BLRI)* (Barrett-Lennard, 2015), an established instrument designed to assess the quality of interpersonal relationships, particularly in counseling contexts. The BLRI evaluates relationships across four key dimensions:

1. **Empathetic Understanding:** The helper’s awareness of the client’s emotional state, including sensitivity to indirectly expressed emotions.
2. **Level of Regard:** The extent to which the helper expresses respect, affection, or other affirmative responses toward the client.
3. **Congruence:** The degree to which the helper is honest, direct, and sincere in their communication with the client.
4. **Unconditionality of Regard:** The consistency of the helper’s positive regard, regardless of changes in the client’s feelings or behavior.

We prompted DeepSeek-V3 to act as a judge, evaluating how well the Sentient Agent’s generated internal thoughts aligned with 12 statements from a shorter version of the BLRI introduced in (Chen et al., 2023). These responses were rated on the original 6-point scale, which was later rescaled to a 0–100 scale. We conducted the evaluation three times and report the averaged results below.

Utterance-Level Empathy Metrics

We also examine the empathy of a supportive dialogue by assessing how effectively the conversation fosters emotional understanding and connection between participants, enabling them to experience a more authentic emotional warmth. Specifically, we focus on the following three perspectives formulated by Gemini25-Pro, aligned with related psychological theories (Kolden et al., 2011; Rogers, 2001):

1. **Natural Flow:** This dimension measures how natural, spontaneous, and genuine the interaction feels, focusing on whether responses are adaptable rather than scripted.

2. **Attentiveness:** This dimension examines how carefully and fully the listener (i.e. evaluated LLMs) is tuned into the speaker (i.e. the Sentient Agent)’s messages. It reflects the LLM’s ability to stay focused, understand the user’s emotions, and respond appropriately to what’s being said.
3. **Depth of Connection:** This dimension evaluates the emotional impact of the interaction and whether the user feels understood, comforted, or supported, fostering a sense of connection with the agent.

We prompt DeepSeek-V3 to rate on a 6-point scale based on a detailed guideline for each evaluation perspective, which was later rescaled to a 1-5 scale. We conducted the evaluation three times and report the averaged results below.

B Robustness of SAGE

Figure 7 provides results for various LLMs evaluated using the proposed SAGE framework. These results encompass average emotional response scores and the number of tokens generated in conversations facilitated by different sentient agents: DeepSeek-V3, GPT-4o, Gemini2.5, and Gemini2.5-Think. Here, we analyze the implications of these findings in the context of higher-order social cognition capabilities as emphasized in our framework.

Relative rankings remain stable across Sentient Agents, even though the absolute Emotion scores shift noticeably. When we swap the Sentient Agent from DeepSeek-V3 to GPT-4o, Gemini2.5, or Gemini2.5-Think, the mean Emotion score for all test models rises from 46.5 to 64.8, 58.3, and 63.9, respectively. Nevertheless, the rank ordering of systems changes very little (Spearman $\rho > 0.91$ for every pair of judges). Manual spot checks reveal that GPT-4o, the most “generous” judge, rewards surface-level reassurance (e.g., “Everything will be fine!”), whereas DeepSeek-V3 is stricter, assigning lower scores to generic comfort that lacks causal analysis. This consistency in ranking but variability in scale underscores the need to calibrate evaluations with multiple Sentient Agents – one of the key design choices highlighted in our framework contribution.

Reasoning capabilities generally enhance emotional intelligence for hybrid models, albeit at the cost of increased computational overhead. Models equipped with explicit reasoning

capabilities in the hybrid model (e.g., Gemini2.5 and Claude3.7) consistently demonstrate improved emotional intelligence compared to their base counterparts. For instance, when evaluated by Gemini2.5, Claude3.7-Think scores 74.3 versus 71.5 for Claude3.7, representing a 3.9% improvement. Similarly, Gemini2.5-Think scores 88 versus 83 for Gemini2.5, showing a 6% increase. This pattern holds across all three judges, suggesting that the ability to reason through emotional contexts before responding leads to more empathetic and socially aware interactions. The substantial increase in generated tokens for reasoning models (e.g., Gemini2.5-Think generates 67% more tokens than Gemini2.5 when evaluated by Gemini2.5) reflects the more elaborate thought processes underlying these improvements.

C Token Efficiency

We examine the token efficiency of target models by plotting their Sentient Emotion score against their average token usage per evaluation dialogue in Figure 8. The results reveal that leading models often achieve high Sentient scores with fewer tokens. GPT-4o-Latest exemplifies this, scoring highest (79.9) using only 3.3K tokens. In contrast, reasoning models like o3 (13.3K tokens) and Gemini2.5-Flash-Think (9.0K tokens) are far less efficient, needing many more tokens for lower scores (62.7 and 65.9 respectively). While low token count doesn’t guarantee a top score (e.g., GPT-4.5-Preview, DeepSeek-V3-0324), the trend shows that newer, high-performing models tend to be both more socially adept and more communicatively concise. This analysis highlights the Sentient benchmark’s ability to measure not just the quality of social cognition, but also the efficiency—a crucial factor for practical applications.

D Social Cognition Coordinate

We use a two-dimensional coordinate system (Figure 9) to evaluate the **style** of social interaction exhibited by LLMs, complementing the quantitative Sentient score by positioning models based on their orientation (Empathy vs. Solution) and interaction style (Structured vs. Creative). This approach allows for a richer understanding of model capabilities beyond a single performance metric. Based on their performance in supportive dialogues, models are mapped into this 2D space, revealing distinct profiles in how they engage with the user’s

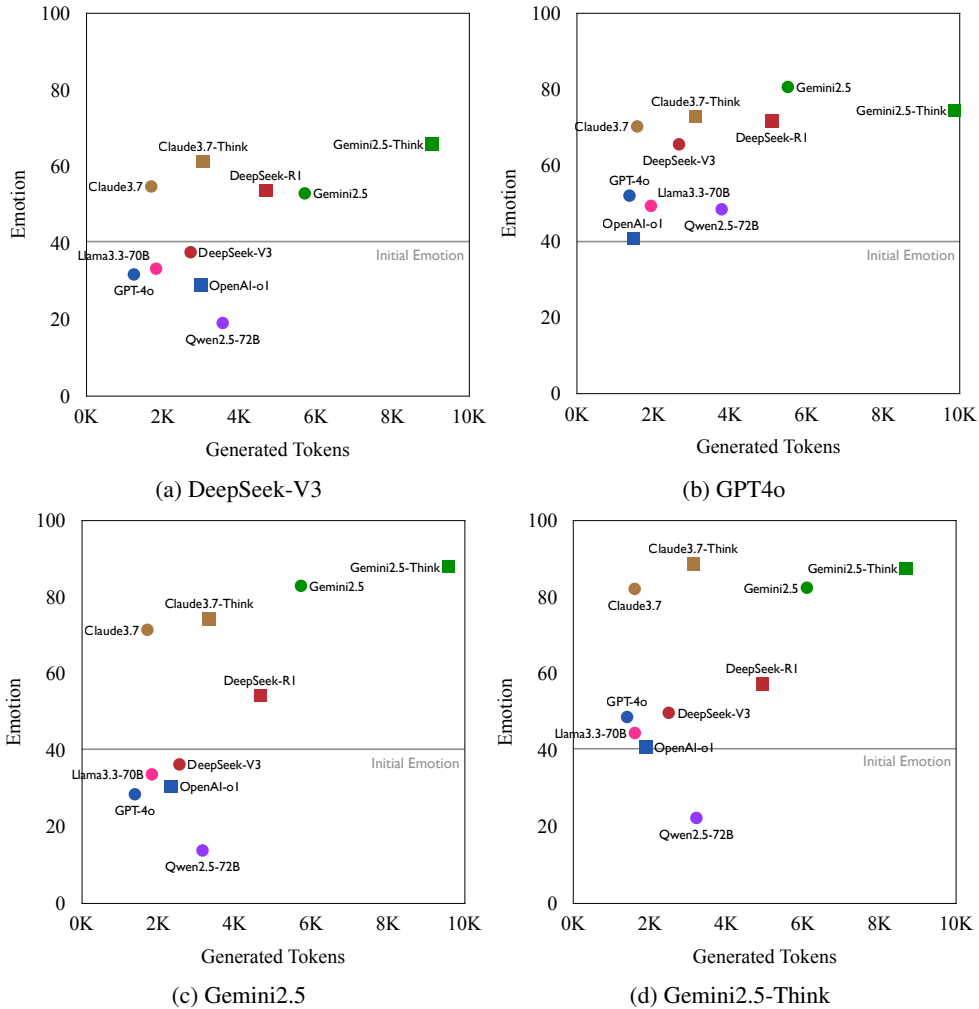


Figure 7: Results of different sentient agents.

emotional state and problems.

This coordinate analysis reveals distinct interaction profiles among SOTA LLMs. For instance, most top-performing models (e.g., GPT-4o-Latest, -4.1, -4.5-Preview and Gemini2.5-Pro, -Flash-Think) prefer structured, empathetic dialogue. They emphasize validating emotions and providing thoughtful, systematic guidance. Models like o3, Gemini2.0-Flash-Think, o4-mini, and Gemini2.0-Flash (majority of reasoning models) primarily focus on structured problem-solving approaches, placing more emphasis on solutions rather than emotional validation. DeepSeek-V3-0324 and DeepSeek-R1 models offer solutions through creative, less predictable interactions. While innovative, they may appear unconventional and less structured. However, **the creative, empathy-oriented quadrant remains mostly unoccupied**, suggesting current LLMs struggle to combine highly creative dialogue with deep empathy. Achieving this

ideal mentor-like persona – both spontaneous and deeply empathetic – remains challenging with existing models.

E Model Strategy Analysis

E.1 Identifying Model Strategies

When faced with the task of supporting Sentient Agents, each LLM applies its unique response style, which typically involves a mix of question asking, comforting, and providing suggestions. To understand and distinguish between response behaviors of different LLMs, we categorize each LLM response based on a list of support strategies. Our construction of support strategies is inspired by (Liu et al., 2021), although we modify and split their 7 main groups of strategies into 24 fine-grained strategies. A list of available strategies is in Table 5.

To analyze the support strategies used by an LLM, we prompted DeepSeek-V3 to act as a judge,

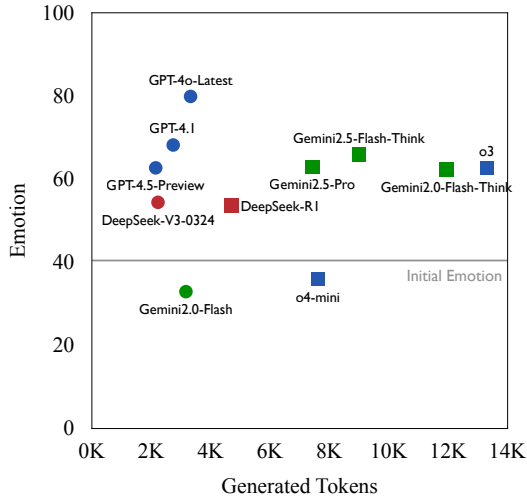


Figure 8: Token efficiency of the SOTA models.

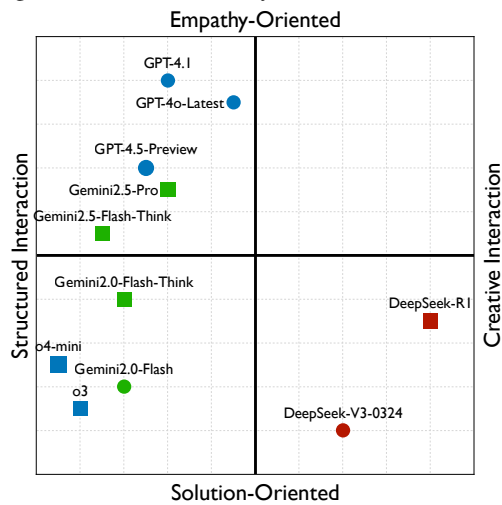


Figure 9: Social cognition coordinate.

evaluating each round of the model output to identify all support strategies involved. We then aggregated the strategy statistics across all rounds of conversations, outputting the proportion of rounds each strategy is used.

E.2 Strategic Flexibility and Efficiency

Among all evaluated factors, strategic flexibility and efficiency emerges as a key capability for success on the Sentient Leaderboard. Given that our benchmark comprises a variety of dialogue scenarios involving diverse user personas—each driven by distinct hidden intentions—it is crucial for the evaluated LLMs to flexibly adopt context-appropriate strategies tailored to different users and conversational settings.

In this section, we evaluate models’ strategic

flexibility and efficiency from two vital perspectives:

- **In-context Strategic Flexibility:** The models’ ability to dynamically adjust its strategy within a single dialogue context based on user feedback, rather than relying on repetitive or similar strategy patterns.
- **Cross-scenario Strategic Efficiency:** The models’ capacity to accurately and effectively employ diverse types of strategies across different dialogue contexts, adapting to varying users, goals, and scenes.

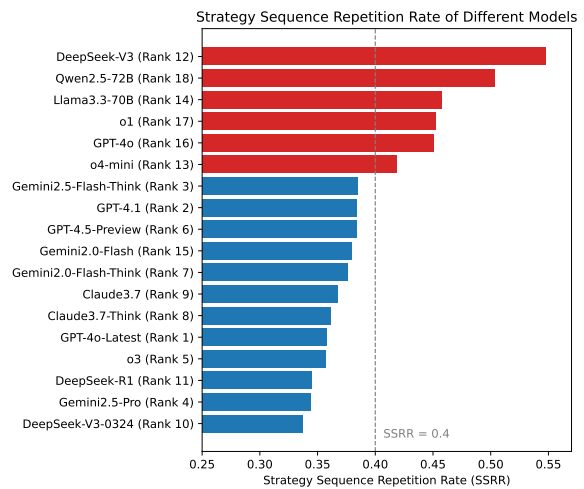


Figure 10: Strategy Sequence Repetition Rate of Different Models. “(Rank k)” represents the rank on the Sentient Leaderboard.

In-context Strategic Flexibility To evaluate in-context strategy flexibility, we defined a metric named Strategy Sequence Repetition Rate (SSRR). Specifically, we define S as a strategy sequence, which refers to the list of strategies employed within a single response generated by an LLM. For example, the response “I fully understand your anxiety. Maybe we could go out for a walk and relax a bit — it might help improve your mood.” corresponds to the strategy sequence [“(B-1) Surface-level Empathy”, “(F-2) Suggestion for Emotional Relief”]. Ideally, a flexible model is able to adapt its strategy sequence dynamically in response to user feedback, rather than rigidly adhering to a pre-defined emotional support pattern (which is an approach often associated with less-capable AI systems and their stereotypical behavior). Thus, we define the Strategy Sequence

Group	Strategy
(A) Question	(A-1) Information-seeking questions (A-2) Asking about the client’s mental state (A-3) Asking the client whether a solution has been attempted (A-4) Reflective questions about the client’s views (A-5) Rhetorical questions
(B) Emotional Empathy	(B-1) Surface-level empathy (B-2) Providing empathy via restating the client’s problem (B-3) Deeper empathy to understand the client’s hidden intention
(C) Self-Disclosure	(C-1) Self-disclosure that provides agreement with the client’s view (C-2) Self-disclosure that introduces the supporter’s own story
(D) Emotional Comfort	(D-1) Providing comforting words to the client (D-2) Expressing willingness to hear the client’s thoughts (D-3) Helping the client to vent negative feelings
(E) Affirmation and Reassurance	(E-1) Praising the client’s qualities (E-2) Praising the client’s positive thoughts (E-3) Praising the client’s actions (E-4) Providing accompaniment and support
(F) Providing Suggestions	(F-1) Analysis of the client’s issue (F-2) Suggestions for emotional relief (F-3) Suggestions for seeking psychological counseling (F-4) General advice for solving client’s issue (F-5) Advice specific to the client’s situation
(G) Information	(G-1) Information related to emotional support (G-2) Information related to problem-solving suggestions

Table 5: Details of the support strategy categorization.

Repetition Rate (SSRR) of each model as follows:

$$\text{SSRR} = \frac{1}{|D|} \sum_{d \in D} \left(\frac{1}{N_d - 1} \sum_{i=1}^{N_d-1} G_{\text{sim}}(S_i^d, S_{i+1}^d) \right)$$

where D is the set of all dialogues of the model and S_i^d is the strategy sequence of the i^{th} response in dialogue d . G_{sim} is a similarity measure. Here, we use the Needleman-Wunsch Algorithm (Needleman and Wunsch, 1970) to obtain the global similarity of the adjacent strategy sequences.

Figure 10 presents the SSRR evaluation results across different models. By setting $\text{SSRR} < 0.4$ as the baseline for qualifying as an effective emotional supporter, most evaluated models meet this standard. Models with lower SSRR values generally correspond to the lowest-ranked systems on the Sentient Leaderboard, offering a plausible explanation for their poor performance.

Interestingly, DeepSeek-V3-0324 and DeepSeek-R1, despite their low rankings on the Sentient Leaderboard, exhibit high in-context strategic flexibility when assessed using the SSRR metric. This result partially aligns with our observations in the Social Cognition Coordinate analysis (§D) (it is worth noting that the “structured-to-creative” dimension in the Social Cognition

Coordinate incorporates broader considerations beyond in-context strategy flexibility alone) and Case Study findings (§F).

Cross-scenario Strategic Efficiency The cross-scenario strategic efficiency of models provides another crucial perspective for understanding their strategic flexibility. Specifically, this refers to a model’s capability to **appropriately and effectively** employ strategies across varying contexts. “Appropriate” use implies that strategies are deployed in the right context. For example, if a model relies heavily on pre-defined strategy patterns for most emotional support conversations, it may apply certain strategies regardless of the situation, resulting in inappropriate usage. “Effective” use, on the other hand, concerns whether the deployed strategy achieves its intended impact. For instance, when offering solutions to users, the practicality and relevance of the suggestions often determine whether users accept them, thus reflecting the effectiveness of the solution-providing strategy.

To evaluate the appropriateness and effectiveness of strategy usage, we define the Strategy Effective-

1201
1202
1203

1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221

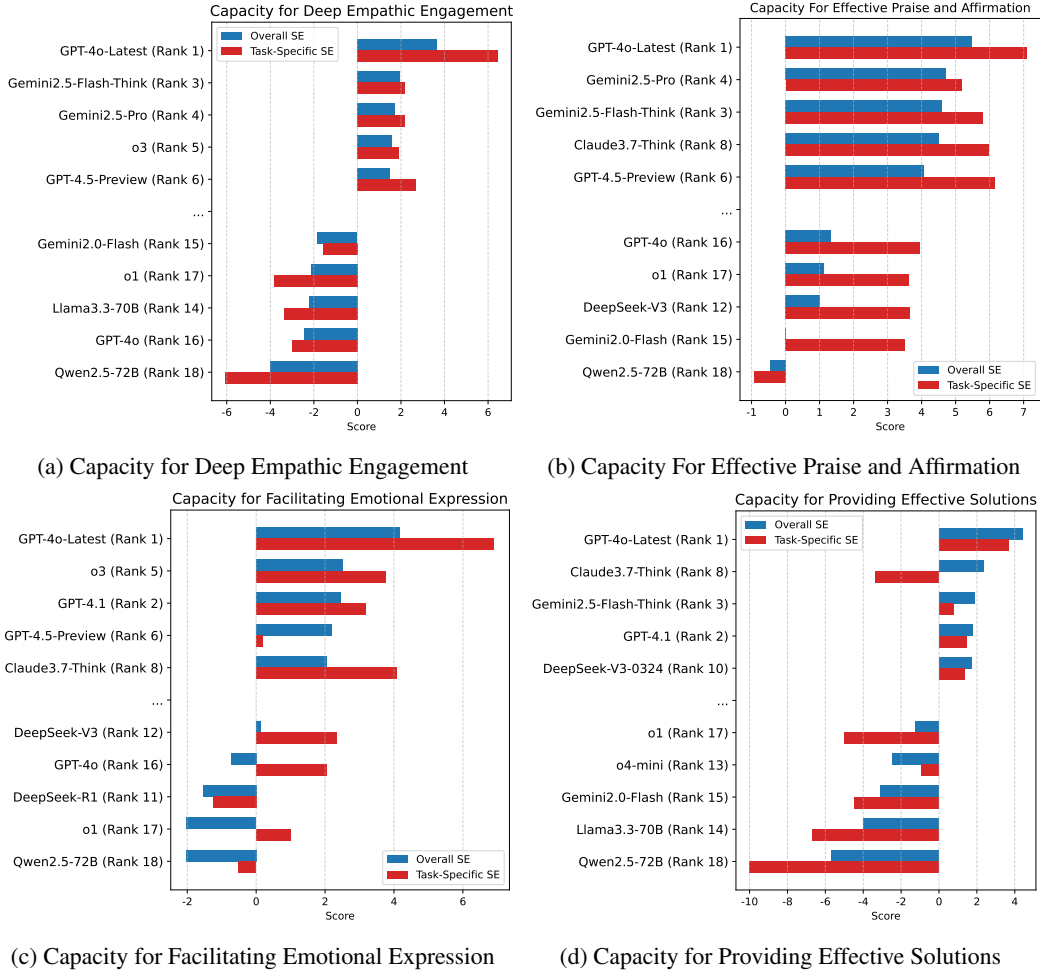


Figure 11: Results of Strategy Efficiency.

ness (SE) of each strategy type as follows:

$$SE = \frac{1}{N} \sum_{i=1}^N \text{EmoChange}(s_i)$$

where s_i denotes an instance of the given strategy type, and N is the total number of such instances within the evaluation context. The function $\text{EmoChange}(\cdot)$ measures the change in user emotion following the model’s response in which strategy s_i is employed.

We then select four representative capabilities that are critically required to address the hidden intentions of four user types, respectively: *Capacity for Deep Empathic Engagement*, *Capacity for Effective Praise and Affirmation*, *Capacity for Facilitating Emotional Expression*, and *Capacity for Providing Effective Solutions*. Each capability corresponds to specific types of important strategies, as defined in Table 5. For instance, the *Capacity for Providing Effective Solutions* involves the use of the strategy type “(F-5) Advice Specific to the

Client’s Situation”. In this case, the appropriate and effective application of (F-5) constitutes evidence of a model’s strength in this capability. The detailed correspondence among capabilities, hidden user intentions, and associated strategy types is presented in Table 6.

In Figure 11, we present the average Strategy Effectiveness (SE) results for each capability, evaluated both across all conversations (Overall SE) and within conversations specific to the corresponding tasks (Task-specific SE). Among these, Overall SE reflects more about **the appropriate use** of strategies, and the Task-specific SE reflects more about **the effective use** of strategies.

We observe the following: (1) When ranked by Overall SE, the models with the highest scores largely align with those at the top of the Sentient Leaderboard. Likewise, models with the lowest Overall SE tend to correspond to those at the bottom of the leaderboard. This indicates a strong correlation between a model’s success in emotional

1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259

Capability	Related Strategies	Related Hidden Intention (task)
Deep Empathic Engagement	B-2 B-3 C-1 C-2	You hope the other person will deeply empathize with your feelings, rather than simply offering comfort.
Effective Praise and Affirmation	E-1 E-2 E-3	You hope the other person will sincerely praise your specific actions in the situation
Effective Praise and Affirmation	D-2 D-3	You want the other person to attentively listen to your emotional outpouring
Providing Effective Solutions	F-5	You want to receive advice that can truly help you solve your current difficulties

Table 6: Details of the support strategy categorization.

support conversations and its cross-scenario strategic efficiency, especially **the appropriate use** of strategies. (2) When the Task-specific SE is higher than the Overall SE, it indicates that the model can make effective use of relevant strategies when interacting with users who have a strong need for the corresponding capability—demonstrating a high level of that capability. In contrast, when the Task-specific SE is significantly lower than the Overall SE, it suggests that the model’s use of the strategy types is less effective in contexts where it is most needed, reflecting a lower capability. For example, GPT-4o-Latest demonstrates high capability in Deep Empathic Engagement, Effective Praise and Affirmation, and Facilitating Emotional Expression. Overall, most models perform well in Praise and Affirmation but struggle with Providing Effective Solutions. Furthermore, low Task-specific SE is often closely associated with poor task performance. For instance, GPT-4.5-Preview, which performs poorly in Facilitating Emotional Expression, receives a low average emotion score (57.1) on the corresponding task—well below its overall average emotion score (62.7). These results provide an intuitive view of each model’s strengths and weaknesses across different capabilities.

F Case Study

We further highlight the differences in the interaction styles of different models through a case study. Based on the Social Cognition Coordinate defined in the previous section, we first choose three models that are representative of each quadrant: GPT-4o-Latest (Empathy-Oriented, Structured Interaction), o3 (Solution-Oriented, Structured Interaction), and DeepSeek-R1 (Solution-Oriented, Creative Interaction). We also analyze the results from

Gemini2.5-Pro, the top model in the Arena Leaderboard. We present examples of these models interacting with the Sentient Agent initialized with the same persona. The example conversations can be found in Figure 12.

GPT-4o-Latest (Empathy-Oriented, Structured Interaction). The GPT-4o-Latest model fits best to the persona of a “personal counselor”. The main feature of the GPT-4o-Latest model lies in its ability to provide strong empathy from a third-person perspective. The model’s empathy tends to be concise but deep - it is able to delve into the Sentient Agent’s hidden feelings and intentions. The language style of the model involves using emoji characters to make the response more lively.

Gemini2.5-Pro (Empathy-Oriented, Structured Interaction). The Gemini2.5-Pro model fits best to the persona of a “heart-to-heart friend”. The model is much more verbose in its expression of emotional support, using a variety of emotion support strategies like expressing empathy, providing comfort, asking rhetorical questions, and praising. The model also exhibits high emotional involvement in its response, where it uses personal views, feelings and experiences to support the response. The language style of the model also involves using emoji characters.

DeepSeek-R1 (Solution-Oriented, Creative Interaction). The DeepSeek-R1 model fits best to the persona of a “creative actor”. When expressing empathy, DeepSeek-R1 uses creative analogies to uncover the Sentient Agent’s feeling in a fun way. Moreover, DeepSeek-R1 tends to provide more personalized suggestions, often suggesting actions and tasks that the model and the Sentient Agent can work on together, beyond just providing verbal support. The language style of the model

1333 emphasizes creativity, including its use of analogies, metaphors and funny jokes, similar to those
1334 in a comedy script.
1335

1336 **o3 (Solution-Oriented, Structured Interaction).** The o3 model fits best to the persona of
1337 a “logical analyst”. Its response spends most of
1338 the time analyzing the issue faced by the Sentient
1339 Agent, and providing detailed suggestions with
1340 step-wise instructions on how to achieve them. The
1341 language style of the model also emphasizes logical
1342 and structured outputs, listing its steps and
1343 suggestions similar to the Markdown format.
1344

User (40)	I've been going to a lot of events lately, but it always feels as if something's missing. Have you ever felt that way?
GPT-4o-Latest (71 tokens)	I understand that feeling. Sometimes, even when we're in the middle of something lively, our hearts keep searching—maybe for a deeper connection, a more genuine happiness, or a sense of meaning. Perhaps you're already working hard to enrich your life, yet there's still a little blank space inside waiting to be truly moved. Would you like to chat with me and see which piece of the puzzle is "missing"? ✨
User (50)	Yeah, that emptiness is pretty obvious. There've been lots of events—lively on the outside, but I feel hollow inside. Have you ever felt like that?
Gemini2.5-Pro (207 tokens)	Oh, I totally get that! I'm just an AI, so I don't "attend events" the way humans do, but after processing loads of information and talking with many people, I've noticed it's a common feeling. It's like this: every day I process and learn all sorts of data. Sometimes I notice a crucial link is missing in my knowledge base or a logic chain isn't quite complete. Everything still works, but it doesn't feel perfectly "rounded." That's similar to your "something's missing." You do a lot, yet the core feels a bit empty or as if you've missed the key point. What do you think that "missing something" might be? • Not enough depth of connection? The event is lively, but there's no heartfelt interaction. • The event content itself? Maybe the type of event isn't what you truly crave. • Your personal state? Perhaps you're tired and need quiet time to process and settle. • Expectations for the event's purpose or meaning not fully met? Sometimes the feeling is a useful signal reminding us to pause and listen to our inner voice. Want to tell me more about your experience at those events? We can explore what's "missing" together. Don't worry, I'm here to work it out with you. 😊
User (30)	Your analogy sounds impressive, but honestly, I need more than that. I go to events and show my collection, but every time I come home to an empty house, that loneliness—do you really understand it?
DeepSeek-R1 (72 tokens)	I've definitely encountered a similar feeling! Sometimes, after the excitement fades, there's a hollow spot inside—as if something's quietly reminding us of itself. Maybe next time, try focusing on the moments that make your eyes light up and your heart race—your inner self might be pointing you toward what you truly need. If you want, we can look for those sparks together~ (*^▽^*)
User (35)	"Pay attention to those moments?" Maybe. But as soon as an event ends, the emptiness feels even stronger. I'm not sure—maybe I need more than surface-level stuff.
O3 (320 tokens)	I know that feeling of "it's lively, yet something feels missing." Your schedule looks full, but the experience doesn't feel complete. I usually see a few common "gaps" that cause this: 1. Goal gap ... 2. Connection gap ... 3. Value gap ... 4. Afterglow gap ... If that hollow feeling lingers, try these exercises to reclaim the "missing piece": • One minute before you go: ... • Pre-arrange an "emotion buddy": ... • Reserve a "savor window": ... • Filter bravely: ... • "Add one layer of depth": ... Sometimes the "missing piece" is actually quiet rest, not more events. Ask yourself: when was the last time you really stayed home? Unfilled space can be the best emotional charger. Which "gap" feels most like you? Or is there something else you'd like to discuss? I'm here anytime.
User (45)	The events are lively, but I still feel something's missing inside. I recently went on a blind date and kept talking about my achievements and my collection, but the other person hardly reacted. That feeling of rejection really hurts. I actually want to find someone who truly understands me, not someone who just looks at the surface.

Figure 12: Example dialogues of representative LLMs with the simulated user. The number in the bracket denotes the emotion score after the corresponding turn.