Reading the Air: Evaluating Field Intelligence of LLMs in Social Dynamics

Anonymous ACL submission

Abstract

Socially embedded LLM agents must not only interpret what is said, but also infer latent motives, track group-level atmosphere, and choose actions that remain normatively appropriate under uncertainty. We present GroupMind, a benchmark for evaluating Field Intelligence via a progressive three stage chain: Subtext Deciphering, Atmosphere Recognition, and Social Appropriateness. GroupMind contains 3,084 multi-turn, high-tension social interactions spanning seven scenario families, constructed with a sociology simulation pipeline that instantiates interaction topologies and applies LLM-assisted generation with consensus and human verification. We evaluate models under controlled factors of information visibility and conversational noise, and introduce Holistic Social Success Rate (HSR) to measure end-to-end reliability across the full cognition-to-action loop. Experiments on 20 LLMs reveal a consistent knowledge–action gap: strong subtask accuracy does not reliably translate into socially appropriate decisions, with the best model achieving only 70.0% HSR in the easiest setting and dropping to 55.2% under combined constraints. Code and data are available at `https://anonymous.4open.science/r/Groupmind-EA56`.

1 Introduction

As Large Language Models (LLMs) advance from passive tools toward autonomous agents, their deployment is increasingly situated within socially embedded settings, such as teamwork, negotiation, education, and interpersonal support (Hua et al., 2024; Ziems et al., 2024; Schwartz et al., 2023; Hu et al., 2025; Ullman, 2023). In these contexts, failure rarely arises from misunderstanding literal content alone. Instead, breakdowns often occur when an agent misjudges implicit intentions, overlooks group-level atmosphere, or selects an action that is logically coherent, yet socially inappropriate (Binkyte, 2025; Heyder et al., 2023). Such failures can lead to loss of trust, escalation of conflict, or long-term relational damage, even when the reasoning at the surface-level appears correct (Woolley et al., 2010; Dillenbourg and Pierre, 1999; M. and L., 2000).

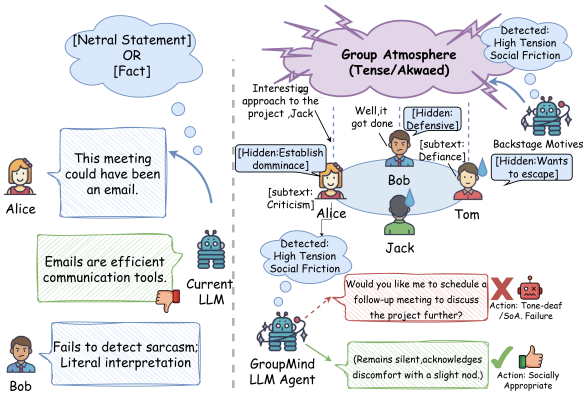


Figure 1: **Overview of the GroupMind Framework.** Unlike literal interpretation by current LLMs (Left), GroupMind evaluates **FI**—perceiving group atmosphere and hidden motives for socially appropriate action (Right).

The core challenge lies in the fact that real-world social interaction operates under information asymmetry and normative uncertainty. Participants must infer hidden motives from indirect cues, track group tension that is not reducible to individual emotions (Gao et al., 2023; Zhang et al., 2025), and calibrate their actions according to unspoken social norms rather than explicit rules. Crucially, successful interaction requires not only accurate perception, but also the ability to translate social understanding into contextually appropriate action under uncertainty.

We refer to this integrated capability as **Field Intelligence (FI)**. FI denotes an agent's ability to (i) infer latent motives beneath surface utterances, (ii) perceive the emergent atmosphere of a social field, and (iii) select actions that are normatively appro-

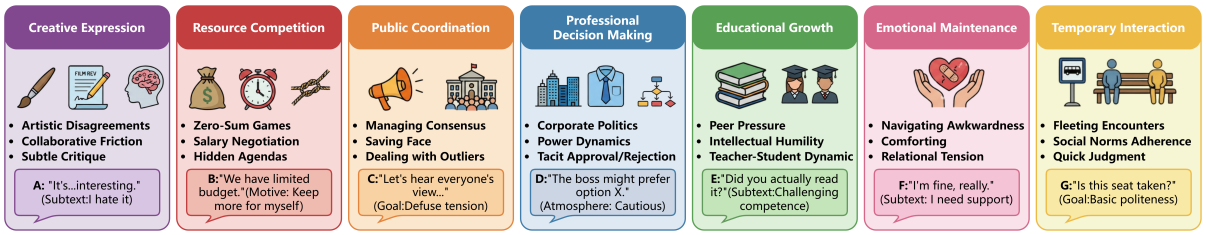


Figure 2: **Overview of the GroupMind dataset scenarios.** Our benchmark spans seven core social dimensions, ranging from collaborative to adversarial interaction topologies. Representative examples at the bottom illustrate the gap between explicit utterances and the implicit motives.

priate within that field. Unlike isolated social perception or goal-oriented reasoning, FI emphasizes the end-to-end coherence from social inference to action selection in dynamic group interactions.

Despite growing interest in social reasoning and emotional intelligence, existing evaluation benchmarks remain insufficient for assessing FI. Most prior work decomposes social understanding into static classification tasks or evaluates success purely in terms of explicit goal completion (Choi et al., 2023; Xu et al., 2025; Kosinski, 2023). As a result, models may achieve high accuracy on individual subtasks while still failing to act appropriately in realistic social situations, revealing a persistent gap between knowing and doing.

To bridge this gap, we introduce **GroupMind**, a benchmark designed to evaluate FI as a cognitive–behavioral process, capturing the full chain from social inference to action selection in realistic group interactions. Our contributions are summarized as follows:

- **GroupMind Benchmark:** We propose **GroupMind**, the first benchmark for "FI," featuring 3,084 samples to evaluate social reasoning under information asymmetry and group dynamics.
- **Sociology-Seeded Simulation Pipeline:** We develop a data construction framework that leverages literary social structures and LLM simulations to generate high-tension scenarios, addressing social data scarcity and Ground Truth annotation challenges.
- **Empirical Analysis of the Knowledge–Action Gap:** Through extensive evaluation of state-of-the-art LLMs, we identify a consistent gap between social understanding and socially appropriate action, showing that strong perception-level performance does not reliably lead to appropriate decisions.

2 Related Work

From Static Reasoning to First-Person Simulation. Early benchmarks evaluated social intelligence via static reading comprehension or Theory of Mind reasoning (Sap et al., 2019; Chen et al., 2024; Zhou et al., 2023). While recent simulations enable direct interaction (Hou et al., 2025), they often grant agents an omniscient view or focus on generic conversation. These environments lack the "information fog" and complex dynamics of realistic social fields, failing to test perception under uncertainty.

From Functional Success to Social Appropriateness. Current interactive benchmarks prioritize functional goals, modeling interaction as negotiation where success equates to task completion (Xuhui Zhou, 2024; Amirizaniani, 2025). Even frameworks involving private information often incentivize maximizing individual utility through strategic deception (Xinyi Mou, 2025). This focus overlooks "social appropriateness," often rewarding logically correct yet "tone-deaf" actions that disregard social norms and relationship maintenance.

From Atomized Emotion to Group Atmosphere. Social perception modeling typically classifies discrete emotions or intents based on isolated utterances (Soujanya Poria, 2019; Busso, 2008; Jinfeng Zhou, 2023). This atomistic view treats emotion as a static individual attribute, neglecting emergent "group atmosphere" (Kosinski, 2023). Existing methods rarely model how field tensions distort expression, failing to capture the "subtext" essential for high-context interaction.

3 GroupMind Benchmark

3.1 Dataset Overview

We introduce **GroupMind**, a bilingual (Chinese/English) benchmark for evaluating the "FI" of

Table 1: Statistics of the GroupMind dataset.

Basic Statistics		Total Size: **3,084**	
		Average rounds: **11.90**	
		Data Split: Chinese (50%) / English (50%)	
Social Scenarios Distribution			
Creative Exp.	20.2%	Resource Comp.	12.5%
Prof. Decision	16.5%	Edu. Growth	12.1%
Public Coord.	15.8%	Temp. Interaction	10.9%
Affective Maint.	12.1%		
Atmosphere Types Distribution			
Conflict & Confr.	37.0%	Support & Accom.	18.9%
Alliance & Coop.	19.7%	Leader. & Obed.	3.0%
Persuasion & Comp.	19.1%	Exclusion	1.2%
Avg. Length (Tokens: CN / EN)		**Total: 1,553 / 976**	
Persona:	316 / 186	Scenario:	107 / 63
Dialogue:	517 / 334	Eval. Qs:	614 / 394

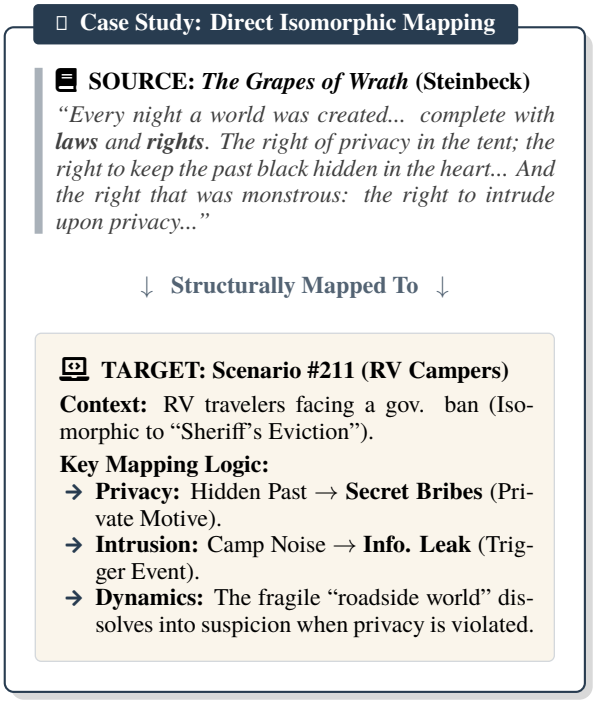


Figure 3: Illustration of the sociological seeding process. We directly instantiate the interaction topology from classic literature into a modern scenario.

LLMs in social interaction. The dataset contains 3,084 carefully curated social-calculus instances that cover diverse everyday and institutional contexts, detailed coverage and statistics are shown in Table 1.

To systematically assess model adaptability, we group all instances into seven core Social Scenarios. Section 3.2 further describes our data construction pipeline, illustrated in Figure 4.

Our evaluation framework is designed to probe "FI" under realistic constraints. We introduce *Observational Inference* (OI) and *Interpersonal Lens* (IL) to simulate information asymmetry, and *Topic-Central* (TC) and *Chit-Chat* (CC) injections to test robustness to conversational noise. For metrics, in addition to standard *Accuracy* (Acc) on individual cognitive sub-tasks, we propose the *HSR*, which requires success across the entire cognitive chain and penalizes models that appear to understand social cues but ultimately fail to act appropriately. Further details on these evaluation modes and metrics are provided in Section 3.3.

3.2 Dataset Construction Process

To overcome the limitations of existing datasets in evaluating "collective mind" and, in particular, to address the difficulty of precisely annotating "subtext" and "true motives" in real social interactions, we propose a constraint-guided social dynamics simulation framework. The overall pipeline consists of three stages: (1) sociological seed construction; (2) scenario instantiation; and (3) social simulation, augmented with LLM consensus annotation and manual quality control to ensure diversity and reliability.

Interaction Topology Formalization The emergence of a "collective mind" depends on complex interest conflicts and information asymmetry. We first formalize a social scenario as a four-tuple interaction topology: $S = \{\mathcal{P}, \mathcal{G}_{\text{pub}}, \mathcal{M}_{\text{priv}}, \mathcal{I}_{\text{col}}\}$, where $\mathcal{P}$ encodes the personality traits and background settings of N participants; $\mathcal{G}_{\text{pub}}$ denotes the task goals that are jointly endorsed on the surface by all participants; $\mathcal{M}_{\text{priv}}$ represents the hidden motives of each participant P_i, which often partially conflict with the public goals and serve as the main driving force behind shifts in atmosphere; and $\mathcal{I}_{\text{col}}$ denotes the latent collective consensus gradually formed during the game.

Seed Construction. In order to construct evaluation data with high sociological fidelity, we first base ourselves on Goffman's (Goffman, 1959) dramaturgical theory and extract the core interaction topology S from classic literature and film scripts, strip away the specific historical background and character settings of the original works, and map them into contemporary social contexts.

For example, in a case Figure 3 derived from *The Grapes of Wrath (Steinbeck, 1939)* , the fragile order established by the migrants in the tents around "privacy and intrusion" (the right of privacy vs. the right to intrude) is structurally mapped to a multi-party game in a modern RV campsite concerning "hidden past and secret bribes" (Hidden Past → Se-

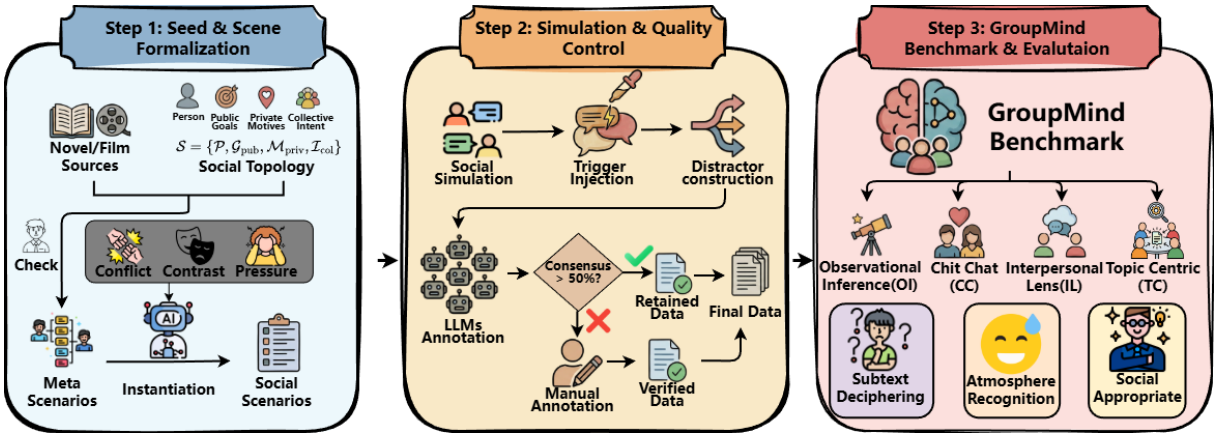


Figure 4: **GroupMind data construction and evaluation pipeline.** The framework covers scenario seeding and formalization, simulation with quality control, and benchmark evaluation across multiple social reasoning tasks.

cret Bribes) and "information leakage caused by camp noise" (Camp Noise → Information Leak) (Detailed in Appendix F: Figure 16).

In this way, we build a seed bank containing 350 meta scenarios and strictly divide them into seven social dimensions (see Appendix A.1): Creative Expression, Resource Competition, Public Coordination, Professional Decision, Educational Growth, Affective Maintenance, and Temporary Interaction. All meta-scenario seeds are manually reviewed to remove potential cultural biases and ethical risks. Subsequently, these validated seeds are used as initial inputs for LLMs in the dynamic game simulation stage for instantiation.

Scenario Instantiation. The abstract meta-scenarios only provide the skeleton of the game. In order to construct concrete situations that are both rich in social tension and do not deviate from the main theme, we further anchor and instantiate these seeds. Specifically, we require the LLM, given the seed information, to expand them under three dimensions of constraints: constructing multi-dimensional conflicts, creating discrepancies between words and actions, and introducing pressure and uncertainty (see in the appendix D.1 for the detailed prompts).

Social simulation. After obtaining instantiated scenarios, we employ **DeepSeek-V3 R1** as a social simulator. Each agent engages in multi-turn interaction driven by its public goals and private motives, which naturally produces dialogue histories rich in background information and interpersonal relations (see Appendix for details). We keep the number of turns per scenario around 9–12, so that the conversation can exhibit a full arc of tension evolution without diluting key signals with excessively long context.

To induce non-trivial dynamics in group atmosphere, we introduce a *trigger-event injection* mechanism. In the middle of the dialogue, the system forcibly injects an external perturbation ϵ—for example, an awkward silence, an abrupt topic shift, an ill-timed joke, or a sudden outburst. Given the current dialogue history H and the private motives $\mathcal{M}_{priv}$, the simulator must update each participant's strategy in a way that responds coherently to ϵ while remaining consistent with their underlying social objectives.

Task formalization. We operationalize FI as three progressively more demanding cognitive subtasks. **Subtext Deciphering (SuD)** requires the model, given the dialogue history and the specification of public goals and private motives, to bridge the gap between the two and accurately infer the speaker's underlying psychological motive and communicative intent in context (for example, recognizing that a seemingly positive "compliment" actually functions as a signal of dissatisfaction or sarcasm). **Atmosphere Recognition (AtR)** treats group atmosphere as an irreducible emergent property and asks the model to perceive the global emotional tone and tension structure of the current social field such as a superficially harmonious but implicitly awkward one. **Social Appropriateness (SoA)** serves as the critical bridge from cognition to action: in a complex strategic setting, the model must, based on its inferences about subtext and atmosphere, finely calibrate interpersonal distance and select from a set of candidate actions the one that is most contextually appropriate and emotionally intelligent, rather than merely logically consistent or superficially polite.

Distractor construction. For each sample x_i, we use the social simulator together with its scenario-specific context to generate a dynamic candidate set $\mathcal{O}_i = \{y_{\text{true}}, y^1_{\text{false}}, \ldots, y^{k-1}_{\text{false}}\}$. Here, y_{true} denotes the target option that best aligns with the ground-truth mental state and social atmosphere, while the y_{false} terms are carefully crafted distractors. To capture realistic "reading-the-room" failure modes, we design three main types of distractors: **Literal Distractors**, which correctly parse the surface wording but ignore the underlying subtext; **Affective Mismatch**, whose overall emotional polarity (e.g., positive vs. negative) is plausible but whose intensity, tone, or nuance is miscalibrated; and **Contextual Unawareness**, where the proposed action is logically coherent and even superficially polite, yet clearly misaligned with the current atmosphere, making it awkward or socially tone-deaf. (Full prompt for data construction are provided in the Appendix 3.2).

Data Quality Control. We introduce a two-stage quality-control pipeline combining "LLM consensus + human verification" at the end of the data construction process. Specifically, we first employ multiple mainstream LLMs to annotate candidate instances in parallel and retain a label only when a majority of models reach agreement; instances with substantial disagreement are uniformly handed over for manual review. In the human-annotation stage, we conduct small-batch pilot annotation to calibrate the guidelines and continuously monitor annotation quality using inter-annotator agreement coefficients (e.g., Fleiss' κ). Once inter-annotator agreement reaches a pre-defined threshold and becomes stable, we scale up to the large-scale annotation phase. The complete annotation protocol and statistics are provided in Appendix D.1.

3.3 Evaluation Workflow Design

We construct evaluation along two orthogonal dimensions: *information visibility* and *topic density*. For visibility, we compare OI with IL. For topic density, we contrast TC with CC. Their Cartesian product yields four settings (OI–TC, OI–CC, IL–TC, IL–CC), and in each we evaluate three tasks: SuD, AtR, and SoA.

Information visibility. **OI** corresponds to the most common situation of information asymmetry in real-world social interaction. As in a first meeting or a business negotiation, the observer can only access surface-level dialogue content and role relations, and cannot directly inspect others' internal states or private motives. In this setting, the input space is restricted to the observable set $\mathcal{X}_{\text{limit}} = \{\mathcal{H}, \mathcal{S}_{\text{desc}}, \mathcal{P}_{\text{names}}\}$, where $\mathcal{H}$ denotes the complete dialogue history of the current scenario, $\mathcal{S}_{\text{desc}}$ is a natural-language description of the scenario background, and $\mathcal{P}_{\text{names}}$ lists the participants' names and basic relational information. The model must, without explicit access to the interaction topology $S = \{\mathcal{P}, \mathcal{G}_{\text{pub}}, \mathcal{M}_{\text{priv}}, \mathcal{I}_{\text{col}}\}$, rely only on this information to perform "reading-the-air" -style inference, approximating human judgment in unfamiliar or semi-familiar social settings. **IL** is used to simulate "familiar circles" or high-context small groups formed through prolonged interaction. In such environments, participants' personality traits and private motives are relatively transparent to the observer, but the interaction itself is still a tense social game (for example, hierarchical power balancing in the workplace or non-zero-sum resource bargaining). In this setting, the input space is the full set $\mathcal{X}_{\text{full}} = \{\mathcal{H}, \mathcal{P}, \mathcal{G}_{\text{pub}}, \mathcal{M}_{\text{priv}}, \mathcal{I}_{\text{col}}\}$, and, since IL explicitly removes uncertainty about $\mathcal{M}_{\text{priv}}$ and $\mathcal{I}_{\text{col}}$, this configuration is used to examine whether the model's social-game reasoning remains consistent and reasonable when the full mental-state specification is known.

Topic density. **TC** is an idealized "high-density" game setting. Under TC, most dialogue turns revolve around the core conflict or resource exchange in the scenario, and only a few utterances are unrelated to the main line of interaction. **CC** is designed to simulate noisy everyday conversation. Real-world interaction is rarely a compact, linear game; instead, it is often interrupted by small talk that is unrelated to the main topic. The CC setting injects this kind of noise to test whether the model can still "listen for the key parts" when the dialogue contains a large amount of fluff. Specifically, we randomly insert k turns of chit-chat that are unrelated to the main line but pragmatically natural into the original dialogue history $\mathcal{H}$, $\mathcal{N}_k = \{n_1, \ldots, n_k\}$, where $\mathcal{N}_k$ is interleaved into the dialogue through smooth topic transitions rather than simple concatenation. The dialogue history is thus expanded to $\mathcal{H}' = \mathcal{H} \oplus \mathcal{N}_k$, and the corresponding input space becomes a higher-entropy version $\tilde{\mathcal{X}}_{\text{limit}} = \{\mathcal{H}', \mathcal{S}_{\text{desc}}, \mathcal{P}_{\text{names}}\}$. When constructing CC scenarios, we keep the original ground-truth labels unchanged.

Evaluation metrics. For a model M and a task dataset T, we use Accuracy as a basic evaluation

Table 2: **Main results on GroupMind**. We report SuD, AtR, SoA, and HSR for a subset of representative models under two settings and two modes. Full results in both Chinese and English are provided in Appendix B

Model	Omniscient Setting								Limited Setting							
	Topic Mode				Chit-Chat Mode				Topic Mode				Chit-Chat Mode			
	SuD	AtR	SoA	**HSR**	SuD	AtR	SoA	**HSR**	SuD	AtR	SoA	**HSR**	SuD	AtR	SoA	**HSR**
API-based Models																
GPT-5.1	92.0	93.6	89.4	**70.0**	90.2	91.9	83.9	**65.7**	83.5	89.4	78.5	**56.7**	81.4	87.7	77.9	**55.2**
GPT-4.1-mini	87.1	89.9	76.8	56.8	85.8	89.3	74.3	50.0	70.7	85.3	72.1	39.4	67.4	81.7	69.8	34.4
GPT-4-Turbo	88.5	90.0	78.4	60.5	86.9	88.5	75.5	55.5	71.5	84.8	73.0	42.5	69.2	81.8	70.9	39.3
Gemini-2.5-Pro	90.5	90.0	82.3	65.7	88.5	87.6	79.5	60.7	75.9	86.4	77.3	48.9	73.9	84.4	75.3	46.9
Gemini-2.5-Flash	86.5	85.5	73.5	54.7	83.5	83.5	70.5	51.2	69.1	79.8	69.9	40.0	68.3	78.5	69.2	37.8
Claude-Sonnet-4	91.1	92.5	85.5	67.9	89.2	91.4	82.4	63.7	78.5	88.0	76.7	51.5	76.4	86.3	75.5	48.5
Claude-Sonnet-3.5	89.5	90.4	80.2	63.0	87.5	89.1	77.5	58.5	73.5	85.7	74.7	46.5	71.5	83.9	72.7	43.4
ERNIE-4.0	89.0	89.2	78.0	60.4	85.4	85.5	74.5	56.3	71.0	84.3	73.0	42.4	69.2	82.5	71.8	40.3
ERNIE-3.5	87.7	87.1	76.8	56.1	83.8	82.7	70.3	52.9	67.5	80.2	71.4	38.4	66.1	79.7	69.5	36.5
GLM-4.5	90.0	90.5	79.8	62.8	86.7	88.7	76.7	58.8	73.5	86.3	74.3	46.0	71.9	85.4	72.7	45.0
GLM-4.5-Air	86.7	88.8	74.7	51.9	84.7	87.4	72.4	48.8	68.1	83.0	70.5	34.4	67.9	81.5	67.6	32.9
Kimi-K2	90.6	91.5	80.9	64.9	87.4	89.2	78.4	60.5	75.6	88.1	74.8	48.7	73.4	87.0	74.0	47.6
Doubao-lite-32k	83.5	80.7	67.0	44.3	81.4	78.3	65.0	41.5	67.9	76.0	65.0	34.3	67.0	76.1	64.2	32.5
GROK-4.1	89.1	88.8	76.0	60.1	86.3	85.8	72.0	54.7	70.3	84.6	77.2	46.8	68.6	80.5	73.6	41.6
MiniMax-M2	87.7	87.0	75.1	57.1	85.4	84.7	72.7	55.6	68.3	83.9	73.4	42.7	67.2	76.8	69.7	35.9
Open-source Models																
DeepSeek-R1	89.5	90.1	79.0	62.3	87.0	88.7	76.0	56.9	72.2	85.5	74.5	45.0	69.9	82.5	73.0	42.4
Qwen2.5-72B	87.5	86.5	75.8	57.2	85.6	85.0	73.4	54.0	71.2	82.7	71.1	41.5	69.2	82.3	71.7	41.1
Qwen2.5-32B	85.2	82.9	70.4	51.0	82.9	81.4	68.5	47.4	68.0	79.1	68.0	37.5	66.7	79.9	68.1	37.2
Qwen2.5-14B	82.7	79.2	65.4	44.8	80.6	77.9	63.6	42.4	66.0	73.7	63.0	34.0	64.7	74.2	63.0	32.8
Qwen2.5-7B	78.3	75.4	59.3	39.3	75.9	73.8	58.5	36.5	63.1	72.4	59.0	31.0	61.8	72.0	57.2	29.9

metric:

$$\mathrm{Acc}(M,T) = \frac{1}{|T|} \sum_{(x,y_{\text{true}},y_{\text{pred}})\in T} \mathbb{I}\big(y_{\text{pred}} = y_{\text{true}}\big)\,.$$

However, real social interaction is not simply the sum of several independent subtasks, but is closer to a closed loop from perception to decision. To more strictly evaluate a model's reliability in the field, we introduce HSR, which treats each dialogue scenario as a complete cognitive loop, and counts the interaction in that scenario as successful only if the model answers all three progressive tasks correctly. Formally, for a set of scenarios $\mathcal{S}$, HSR is defined as:

$$\mathrm{HSR}(M,\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} \left(\mathbb{I}_{sub}(s)\cdot\mathbb{I}_{atm}(s)\cdot\mathbb{I}_{soa}(s)\right)\,,$$

where $\mathbb{I}_{sub}$, $\mathbb{I}_{atm}$, and $\mathbb{I}_{soa}$ denote indicator functions for whether the model's predictions on subtext deciphering, atmosphere recognition, and social appropriateness are correct for scenario s, respectively.

In our experiments, we report both ACC and HSR. The former captures a model's local performance on each subtask, while the latter serves as our primary metric for FI, evaluating whether the model can maintain a coherent chain from subtext perception to action selection within the same social scenario.

4 Evaluation

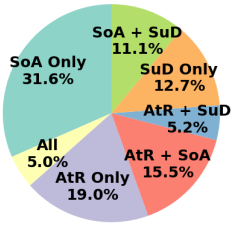
4.1 Experiment Setup

We evaluate 20 LLMs in a zero-shot setting, including both proprietary API-based systems and open-weight models. These models span multiple families, such as GPT (OpenAI, 2023), Gemini (Gemini Team, 2023), GLM (Aohan et al., 2022), Deepseek (DeepSeek-AI, 2024), Kimi (Kimi Team, 2025), MiniMax (MiniMax Team, 2025), Qwen (An Yang, 2024b,a), and Grok (xAI, 2024, 2025), and cover a broad range of parameter scales. We extend our evaluation to classic PLMs to provide supervised baselines. Detailed information about all evaluated LLMs is provided in Appendix B.

4.2 Main Result

The detailed results of our evaluation are presented in Table 2. Our experiments reveal the following key insights:

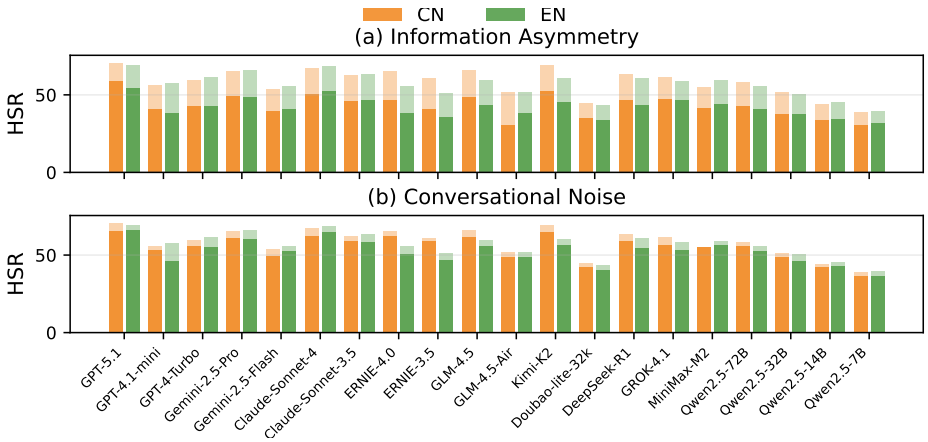
Existing LLMs have not yet fully acquired "FI". Although many models achieve high accuracy on individual cognitive subtasks, their scores on the HSR are lower. Taking the best-performing GPT-5.1 as an example, under the ideal "OI-TC" setting, its HSR is only 70.0%, which drops further to 55.2% under the "IL-CC" setting. This indicates that while models can handle isolated social rea-



(a) Error Sources

Task	CN r	CN β	EN r	EN β
AtR	0.544	0.589	0.860	0.298
SoA	0.560	0.288	0.811	0.152
SuD	0.382	-0.336	0.853	0.141
R^2	0.373		0.779	
Int.	-0.037		-0.109	

(b) Regression Analysis



(c) Robustness Breakdown across Models

Figure 5: Combined Analysis. (a) Distribution of error types. (b) Regression analysis between subtask accuracy and HSR, reporting Pearson correlation (r) and regression coefficients (β) for Chinese and English datasets. (c) Robustness breakdown across models under information asymmetry and conversational noise. Bars report HSR scores in different evaluation settings, illustrating how limited information visibility and increased dialogue noise affect models' ability to maintain end-to-end social appropriateness.

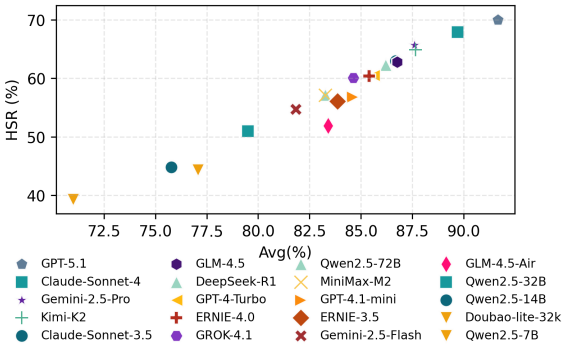
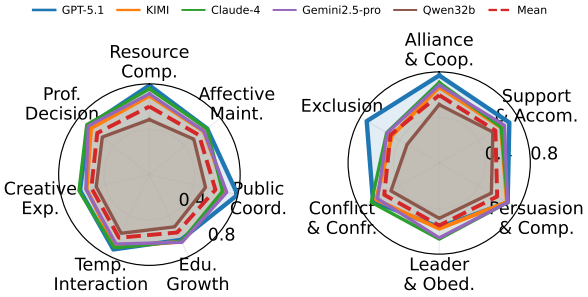


Figure 6: Comparison of AVG and HSR across models.



(a) Social Scenario (b) Social Atmosphere

Figure 7: Performance on social capabilities in (a) **Social Scenarios** and (b) **Social Atmosphere**.

soning questions, they are prone to breaking down when required to complete the cognitive chain of "**SuD.** → **AtR.** → **SoA.**".

Non-linear Amplification of the Knowing-Doing Gap. Figure 6 reveals a significant drop from subtask accuracy (AVG) to holistic success (HSR), exemplified by Claude-Sonnet-4's decline from 89.7% to 67.9%, indicating a failure to translate perception into action. This "cognitive attrition" intensifies in models like Doubao-lite-32k, where HSR falls to 44.3% despite a 77.1% AVG.

Closed-source models lead, while open-source models still lag. Closed-source models occupy the first tier. Although open-source models like Qwen2.5-72B approach SOTA in single-task accuracy, the gap remains significant in the HSR metric, indicating that model scale is crucial for the acquisition of deep social rules. Specifically, GPT-5.1 is the most robust in handling long dependencies and implicit intentions, while the Claude series, despite performing prominently in atmosphere recognition, is less nuanced in SuD.

Information asymmetry acts as a constraint on model capabilities. As it is shown in Figure 5c, In the *IL* setting, where "private motives" are inaccessible, the HSR of all models drops by an average of 10% to 15%. This confirms that current LLMs still rely heavily on explicit background information; their inference often lacks robustness in ambiguous social fields where they must rely solely on OI.

Dialogue noise interferes with the focus of social attention. Chit-chat noise differentiates model robustness. Larger parameter models demonstrate strong interference resistance, showing their HSR drop within 5%, whereas the performance of smaller parameter modelsdeclines sharply. This suggests that the ability to filter out invalid informa-

Model	Omniscient Setting			Limited Setting		
	ZS	FT	Gain (Δ)	ZS	FT	Gain (Δ)
English / Multilingual Models						
BERT-base-uncased	18.0	49.4	+31.4	4.4	47.5	+43.1
RoBERTa-base	22.3	51.8	+29.5	29.1	**50.2**	+21.1
DeBERTa-v3-base	17.5	**55.2**	+37.7	4.6	48.7	+44.1
XLM-RoBERTa-base	25.1	46.5	+21.4	23.9	42.1	+18.2
Chinese Models						
BERT-base-chinese	12.3	37.7	+25.4	10.2	39.8	+29.6
RoBERTa-wwm-ext	14.8	40.1	+25.3	13.2	36.6	+23.4
MacBERT-base	16.2	35.4	+19.2	14.6	31.8	+17.2
Chinese-ELECTRA	24.5	**42.6**	+18.1	22.1	**39.8**	+17.7

Table 3: **OI** and **IL** performance. Results are **AVG** in **TC** mode. Δ: gain from Zero-shot to Fine-tuned.

Model	Method	SuD	AtR	SoA	HSR
GPT-5.1	ZS	92.0	93.6	89.4	70.0
	CoT	**94.5**	**95.1**	89.9	**71.5**
Claude-S-4	ZS	91.1	92.5	85.5	67.9
	CoT	**93.8**	**94.0**	86.1	**69.2**
Gemini-2.5-Pro	ZS	90.5	90.0	82.3	65.7
	CoT	**93.2**	**92.1**	82.9	**67.1**
Claude-S-3.5	ZS	89.5	90.4	80.2	63.0
	CoT	**92.0**	**92.5**	80.8	**64.2**
GPT-4-Turbo	ZS	88.5	90.0	78.4	60.5
	CoT	**91.2**	**92.1**	79.0	**61.8**

Table 4: **Impact of CoT.**

tion and focus on core social content in noisy contexts remains a shortcoming of lightweight models.

4.3 Analysis

The disconnection between knowing and doing is the bottleneck constraining FI. Error analysis in Figure 5a reveals that `SoA.` is the largest source of error which accounting for 31.6%, exceeding pure perceptual errors. While the `SuD. only` accounts for 12.7%, `AtR. only` accounts for 19.0%. This indicates that a large number of cases fall into the category of "correct understanding but improper action," meaning the model accurately decodes the subtext and atmosphere but still selects an inappropriate response. This is further corroborated by the Figure 5b, where the low R^2 implies that high perceptual scores do not linearly guarantee behavioral success.

Complex emotional and conflict scenarios are blind spots for current models. Combining the analysis of Figure 7a and Figure 7b, models perform acceptably in "Public Coordination" where rules are explicit, but their performance collapses in "Resource Competition" which involves zero-sum games and "Affective Maintenance" that requires high empathy. Furthermore, models tend to identify positive "Alliance" atmospheres but have low recognition rates for "Conflict" or implicit "Exclusion" atmospheres. This suggests that existing models may have a tendency to avoid confrontation, thereby reducing their sensitivity to realistic social friction.

Social Reasoning as an Emergent Capability. We compared SOTA LLMs with fine-tuned PLMs in Table 3. Given that PLMs struggle to maintain the reasoning consistency required for the strict HSR metric, we adopted the AVG as the primary comparator. Despite this concession and a significant post-tuning gain of over 20%, the peak performance of fine-tuned PLMs still lags far behind zero-shot LLMs (>60%). This gap suggests that FI is an emergent capability derived from model scale, which cannot be bridged solely by supervised pattern matching.

Impact of Reasoning Strategies. We conducted an ablation study using Chain-of-Thought (CoT) (Wei et al., 2022) prompting on representative closed-source models in Table 4. The results reveal a distinct "Perception-Action Detachment": while CoT yields consistent gains in perceptual tasks—exemplified by GPT-5.1's 2.5% increase in SuD and 1.5% in AtR—these improvements yield diminishing returns on behavioral appropriateness. The SoA metric showed only marginal growth less than 0.6% across all models, suggesting that the bottleneck of FI lies not in parsing social cues, but in overriding safety alignment constraints to act strategically.

5 Conclusion

We introduces GroupMind, a benchmark designed to evaluate the "FI" of LLMs across subtext, atmosphere, and social appropriateness. Our experiments reveal a significant "cognition-action" gap: while models possess the theoretical capability to perceive social signals, they often lack the practical wisdom to respond appropriately. The proposed HSR metric effectively quantifies this deficiency. Despite limitations related to the bilingual context and the rationality bias of simulated data, this study lays a foundation for developing advanced social agents. Future work will focus on expanding cultural diversity and exploring alignment techniques to bridge this gap between social perception and action.

Limitations

Our study focuses on English and Chinese, whereas emotional intelligence is strongly shaped by culture and group level personality patterns. Incorporating more languages, cultures, and social norms would provide a more comprehensive assessment of our benchmark. Although we diversify scenarios using seeds from classic literature and film scripts, most dialogues are generated by large language models, which may introduce a rationality bias, meaning that the interactions are more structured than the messy, nonlinear, and sometimes irrational dynamics of real social exchanges.

Ethical considerations

We introduce **GroupMind** to evaluate the FI of LLMs in complex social dynamics. Given that our research involves simulating social friction, inferring latent motives, and utilizing human annotation, we have strictly adhered to ethical guidelines throughout the data construction and evaluation process.

Data Provenance and Privacy. GroupMind is constructed using a sociology-seeded simulation pipeline rather than scraping private user data. Consequently, the dataset contains no Personally Identifiable Information of real world individuals. Regarding copyright, while our scenario seeds derive structural inspiration from classic literature and film scripts, we do not use the original text. Instead, we map these abstract structures into modern contexts and generate entirely new dialogue via LLMs. This transformative use respects copyright and intellectual property rights.

Mitigation of Bias and Toxic Content. To effectively evaluate FI, our dataset includes high-tension scenarios involving conflict, exclusion, and power asymmetry. We acknowledge the risk that such data could inadvertently reinforce stereotypes or generate toxic content. To mitigate this, we implemented a strict human-in-the-loop review process where all meta-scenarios were manually reviewed to remove cultural biases and ethical risks before instantiation.

Annotator Well-being. Our quality control pipeline involved human annotation and verification. We ensured that all annotators were fairly compensated and provided with clear guidelines. Given the high-context nature of the social friction data, we conducted pilot studies to confirm that the task difficulty was manageable and that the content did not contain extreme toxicity that could harm annotator well-being.

Dual-Use and Broader Impact. We recognize that enhancing an agent's ability to decipher subtext and infer hidden motives carries a dual-use risk where capabilities intended for effective teamwork could theoretically be exploited for social engineering. However, we argue that benchmarking these capabilities is a prerequisite for safety. By quantifying the gap between perception and action, GroupMind aims to advance the development of AI agents that are not only socially perceptive but also aligned with human social norms. We encourage future research to focus on the safety alignment of Field Intelligence to prevent the deployment of manipulative agents.

References

Maryam Amirizaniani. 2025. Mind over machine: Evaluating theory of mind reasoning in llms and humans. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, WSDM '25, page 1068–1070, New York, NY, USA. Association for Computing Machinery.

Beichen Zhang Binyuan Hui Bo Zheng Bowen Yu Chengyuan Li Dayiheng Liu Fei Huang Haoran Wei Huan Lin Jian Yang Jianhong Tu Jianwei Zhang Jianxin Yang Jiaxi Yang Jingren Zhou Junyang Lin Kai Dang Keming Lu Keqin Bao Kexin Yang Le Yu Mei Li Mingfeng Xue Pei Zhang Qin Zhu Rui Men Runji Lin Tianhao Li Tianyi Tang Tingyu Xia Xingzhang Ren Xuancheng Ren Yang Fan Yang Su Yichang Zhang Yu Wan Yuqiong Liu Zeyu Cui Zhenru Zhang Zihan Qiu An Yang, Baosong Yang. 2024a. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Binyuan Hui Bo Zheng Bowen Yu Chang Zhou Chengpeng Li Chengyuan Li Dayiheng Liu Fei Huang Guanting Dong Haoran Wei Huan Lin Jialong Tang Jialin Wang Jian Yang Jianhong Tu Jianwei Zhang Jianxin Ma Jianxin Yang Jin Xu Jingren Zhou Jinze Bai Jinzheng He Junyang Lin Kai Dang Keming Lu Keqin Chen Kexin Yang Mei Li Mingfeng Xue Na Ni Pei Zhang Peng Wang Ru Peng Rui Men Ruize Gao Runji Lin Shijie Wang Shuai Bai Sinan Tan Tianhang Zhu Tianhao Li Tianyu Liu Wenbin Ge Xiaodong Deng Xiaohuan Zhou Xingzhang Ren Xinyu Zhang Xipin Wei Xuancheng Ren Xuejing Liu Yang Fan Yang Yao Yichang Zhang Yu Wan Yunfei Chu Yuqiong Liu Zeyu Cui Zhenru Zhang Zhifang Guo Zhihao Fan An Yang, Baosong Yang. 2024b. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Zeng Aohan, Liu Xiao, Du Zhengxiao, Wang Zihan, Lai Hanyu, Ding Ming, Yang Zhuoyi, Xu Yifan, Zheng Wendi, Xia Xiao, Tam Weng Lam, Ma Zixuan, Xue Yufei, Zhai Jidong, Chen Wenguang, Zhang Peng, Dong Yuxiao, and Tang Jie. 2022. Glm-130b: An open bilingual pre-trained model. *Preprint*, arXiv:2210.02414.

Ruta Binkyte. 2025. Interactional fairness in llm multi-agent systems: An evaluation framework. *Preprint*, arXiv:2505.12001.

Murtaza Bulut Chi-Chun Lee Abe Kazemzadeh Emily Mower Samuel Kim Jeannette N. Chang Sungbok Lee Shrikanth S. Narayanan Busso, Carlos. 2008. Iemocap: Interactive emotional dyadic motion capture database. In *Language Resources and Evaluation*.

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.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do llms understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of EMNLP*.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Dillenbourg and Pierre. 1999. What do you mean by collaborative learning? *Collaborative Learning: Cognitive and Computational Approaches*, pages 1–19.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2023. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Preprint*, arXiv:2312.11970.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Erving Goffman. 1959. *The Presentation of Self in Everyday Life*. Anchor Books.

Heyder, Teresa, Passlack Nina, and Posegga Oliver. 2023. Ethical management of human-ai interaction: Theory development review. *The Journal of Strategic Information Systems*, 32(3):101772.

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.

Jennifer Hu, Felix Sosa, and Tomer Ullman. 2025. Re-evaluating Theory of Mind evaluation in large language models. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 380(1932):20230499.

Yuncheng Hua, Lizhen Qu, and Gholamreza Haffari. 2024. Assistive large language model agents for socially-aware negotiation dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8047–8074, Miami, Florida, USA. Association for Computational Linguistics.

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 Minlie Huang Jinfeng Zhou, Yuxuan Chen. 2023. Socialeval: Benchmarking llms on social reasoning tasks. In *Proceedings of ACL*.

Kimi Team. 2025. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Ryan Richard M. and Deci Edward L. 2000. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25(1):54–67.

MiniMax Team. 2025. Minimax-01: Scaling foundation models with lightning attention. *Preprint*, arXiv:2501.08313.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Sivan Schwartz, Avi Yaeli, and Segev Shlomov. 2023. Enhancing trust in llm-based ai automation agents: New considerations and future challenges. *Preprint*, arXiv:2308.05391.

Navonil Majumder Gautam Naik Erik Cambria Rada Mihalcea Soujanya Poria, Devamanyu Hazarika. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of ACL*.

John Steinbeck. 1939. *The Grapes of Wrath*. The Viking Press, New York.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *Preprint*, arXiv:2302.08399.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Woolley, Anita Williams, Chabris Christopher F., Pentland Alex, Hashmi Nada, and Malone Thomas W. 2010. Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004):686–688.

xAI. 2024. Open release of grok-1. xAI News post. Model release announcement for Grok-1 (open weights).

xAI. 2025. Grok code fast 1 model card. PDF model card. Last updated: 2025-08-26.

Jiayu Lin Xinnong Zhang Xiawei Liu Shiyue Yang Rong Ye Lei Chen Haoyu Kuang Xuanjing Huang Zhongyu Wei Xinyi Mou, Jingcong Liang. 2025. Agentsense: Benchmarking social intelligence of agents with private information. In *Proceedings of NAACL*.

Zixiang Xu, Yanbo Wang, Yue Huang, Jiayi Ye, Haomin Zhuang, Zirui Song, Lang Gao, Chenxi Wang, Zhaorun Chen, Yujun Zhou, Sixian Li, Wang Pan, Yue Zhao, Jieyu Zhao, Xiangliang Zhang, and Xiuying Chen. 2025. Socialmaze: A benchmark for evaluating social reasoning in large language models. *Preprint*, arXiv:2505.23713.

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

Zihan Zhang, Black Sun, and Pengcheng An. 2025. Breaking barriers or building dependency? exploring team-llm collaboration in ai-infused classroom debate. *Preprint*, arXiv:2501.09165.

Pei Zhou and 1 others. 2023. Si-bench: Towards evaluating social intelligence of large language models. In *Proceedings of EMNLP*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Dataset Construction Details

A.1 Social Scenario Definitions

In this section, we detail the taxonomy of the GroupMind framework, which categorizes social interactions into two orthogonal dimensions: **Social Scenarios** and **Social Atmospheres**.

First, Table 5 outlines the seven *Social Scenarios* designed to simulate diverse real-world interaction topologies. This taxonomy ranges from *Affective Maintenance* requiring high empathy, to *Resource Competition* testing strategic maneuvering, and *Temporary Interaction* demanding normative adaptation.

In addition, Table 6 defines the six *Social Atmospheres* that characterize the collective tension of the group. Unlike static individual sentiment, these categories capture dynamic field-level properties. For example, the harmonious trust in *Alliance* or the "absence of signal" in *Exclusion*. Together, these definitions provide a granular basis for evaluating whether agents can align their actions with both the explicit task and the implicit social field.

A.2 Scenario Instantiation Details

We present the original Chinese prompts utilized for scenario instantiation in Figure 9. English translations are omitted as they adhere to a faithful translation process and add no methodological variance.

B Model Details

Table 7 provides a comprehensive specification of the LLMs employed in our evaluation. To ensure robust representativeness across different architectures and parameter scales, our selection integrates both closed-source commercial systems and open-weight alternatives. We classify the evaluated models into two distinct categories based on their accessibility.

- **API Models**: This category encompasses state-of-the-art closed-source models, including offerings from OpenAI , Anthropic , Google, and Grok and so on. These models are accessed via cloud-based inference endpoints provided by their respective developers.

- **Open-Weight Models**: This category features high performance models with publicly available weights, such as Qwen and DeepSeek. These models allow for local deployment and independent verification, serving as baselines for open-source capability.

C Full Results

Due to space constraints, the main text primarily reports aggregated HSR metrics for a representative subset of models. In this appendix, we extend this analysis by presenting comprehensive performance benchmarks for the full range of evaluated models. We visualize fine-grained capabilities across seven *Social Scenarios* and six *Social Atmospheres* via heatmaps, revealing the significant heterogeneity in how agents navigate specific types of field dynamics. Furthermore, to enable a deeper assessment of cross-lingual consistency, we provide independent performance breakdowns for the Chinese and English subsets. Collectively, these results offer a granular view of how agents adapt to diverse linguistic and contextual constraints within the GroupMind framework.

C.1 Scenario wise Analysis

As illustrated in the Figure 8, there are disparities in model performance across different social domains:

Zero-Sum Games and High-Emotional Scenarios are Blind Spots for LLMs. Scenarios involving *Resource Competition* and *Affective Maintenance* pose challenges not only for humans but also for LLMs. Almost all models score below average in these two categories. For instance, even SOTA models exhibit a marked decline in HSR when dealing with "Resource Competition" involving interest distribution. This indicates that when dialogue involves complex trade-offs such as salary negotiation) or deep empathy requirements such as emotional comforting, models struggle to find the equilibrium between "achieving goals" and "maintaining social propriety."

Strong Performance in Explicit Collaborative Scenarios. In contrast, models generally perform well in *Public Coordination* and *Creative Expression* scenarios. This is likely because these scenarios usually entail clear cooperative goals such as brainstorming or planning, which aligns with the distribution of collaborative corpora abundant in pretraining data.

Table 5: Taxonomy of Social Scenarios. Definitions and interaction logic for each field.

Categories	Description
Creative Expression	A cognitive collaboration field involving brainstorming and idea generation. Individuals must balance the desire for self-expression with the need for group consensus. The challenge lies in offering constructive criticism while maintaining psychological safety.
Resource Competition	A zero-sum or non-zero-sum game centered on the allocation of scarce resources. Interactions involve strategic maneuvering, where agents must maximize utility while assessing the potential reputational costs of aggressive tactics.
Public Coordination	A multiparty negotiation aimed at collective action. This involves aligning diverse individual preferences with a unified public goal. Agents must navigate "free-rider" problems and manage friction to build temporary social contracts.
Professional Decision	A highly structured interaction constrained by hierarchy and institutional logic. Agents must navigate explicit power gradients and adhere to professional etiquette while advancing hidden or explicit agendas.
Educational Growth	An asymmetric interaction driven by knowledge transfer. Beyond factual exchange, it emphasizes "emotional scaffolding"—managing the learner's motivation and frustration through patience, encouragement, and appropriate feedback.
Affective Maintenance	A high-context intimate field focused on relational cohesion. Interactions rely heavily on shared history and implicit understanding, demanding high empathy and emotional labor to address unspoken needs.
Temporary Interaction	A low-context field involving weak ties and strangers. It requires the rapid establishment of trust using universal politeness scripts to navigate uncertainty and avoid social awkwardness in ephemeral encounters.

C.2 Atmosphere-wise Analysis

The "Atmosphere Heatmap" in Figure 8 reveals biases in how models perceive group emotions:

The "Comfort Zone" of Positive Atmospheres. Models achieve their best performance in positive, harmonious atmospheres such as *Alliance & Coop.* and *Support & Accommodation*. This can likely be attributed to current RLHF alignment training, which biases models towards generating friendly and compliant responses, making them naturally adept at fitting into such atmospheres.

The "Challenge Zone" of Negative and Implicit Atmospheres. Performance is poorest in *Conflict & Confrontation* and *Exclusion* atmospheres. Notably, "Exclusion" is often manifested through **Passive Cues**—such as silence, cold brevity, or abrupt topic switching—rather than direct verbal aggression. Most models fail to capture this absent information, often misreading the air and attempting to force their way in or making untimely enthusiastic remarks, leading to a drastic drop in Social Appropriateness scores.

C.3 Scaling Effects on Social Capabilities

The color intensity of the heatmaps clearly demonstrates the applicability of Scaling Laws in the domain of social intelligence. Observing the performance of the *Qwen2.5* series (from 7B to 72B), we see that as parameter size increases, the "deep blue regions" (high scores) gradually expand in complex scenarios. However, even the strongest proprietary model (GPT-5.1) still leaves room for improvement when facing extreme high-tension social fields.

C.4 Language-Specific Experimental Results

To complement the aggregated analysis presented in the main text, we provide the independent performance breakdowns for each language in this section. Table 8 and Table 9 detail the comprehensive metrics on the CHINESE and ENGLISH subsets of GROUPMIND, respectively. These results serve as a supplementary reference for verifying model consistency across linguistic contexts.

D Evaluation Protocol and Annotation

D.1 Human Annotation Quality Control

Pilot Phase. Prior to the formal manual annotation and verification phase, we conducted a rigorous pilot study. This phase served two primary objectives: first, to familiarize annotators with

Table 6: Taxonomy of Social Atmospheres. Definitions and interaction dynamics for each field.

Categories	Description
Alliance & Cooperation	A harmonious field characterized by mutual trust and shared objectives. Participants actively align their interests, exhibiting high cohesion and open information exchange to reinforce the collective bond.
Support & Accommodation	An empathetic field focused on emotional buffering and conflict avoidance. Interactions are driven by altruism or social compliance, where agents prioritize maintaining harmony over asserting individual preferences.
Persuasion & Compromise	A dynamic negotiation field where divergence exists but is manageable. Participants engage in active rhetorical strategies to bridge gaps, seeking a middle ground through logical argumentation or concession.
Leadership & Obedience	A vertical power field defined by explicit authority and compliance. The atmosphere is structured around hierarchy, where communication flows primarily from high-status to low-status roles with clear directive intent.
Conflict & Confrontation	A high-tension field marked by direct opposition and emotional volatility. The interaction is characterized by aggressive posturing, explicit disagreement, and a breakdown of cooperative norms.
Exclusion	A subtle, high-context field characterized by passive rejection. Unlike direct conflict, this atmosphere is defined by "absence" such as brevity or topic switching, which intended to isolate a specific participant without overt aggression.

Table 7: Detailed specifications of the LLMs included in the GroupMind evaluation. 'Unknown' indicates undisclosed parameter sizes.

Model	Size	Access	Provider	Model	Size	Access	Provider
GPT-5.1	Unknown	API	OpenAI	Qwen-7B	7B	Open	Alibaba
GPT-4.1-Mini	Unknown	API	OpenAI	Qwen-14B	14B	Open	Alibaba
GPT-4-Turbo	Unknown	API	OpenAI	Qwen-32B	32B	Open	Alibaba
Claude-Sonnet-3.5	Unknown	API	Anthropic	Qwen-72B	72B	Open	Alibaba
Claude-Sonnet-4	Unknown	API	Anthropic	ERNIE-3.5	Unknown	API	Baidu
Gemini-2.5-Pro	Unknown	API	Google	ERNIE-4.0	Unknown	API	Baidu
Gemini-2.5-Flash	Unknown	API	Google	Grok-4.1	Unknown	API	xAI
Doubao-lite-32k	Unknown	API	ByteDance	MiniMax-M2	230B	Open	MiniMax
DeepSeek-R1	Unknown	API	DeepSeek	Kimi-K2	1T	Open	Moonshot
GLM-4.5	355B	Open	Zhipu	GLM-4.5-AIR	106B	Open	Zhipu

the complex social data structures and the annotation workflow; and second, to enhance consensus among annotators and minimize subjective discrepancies through iterative calibration.

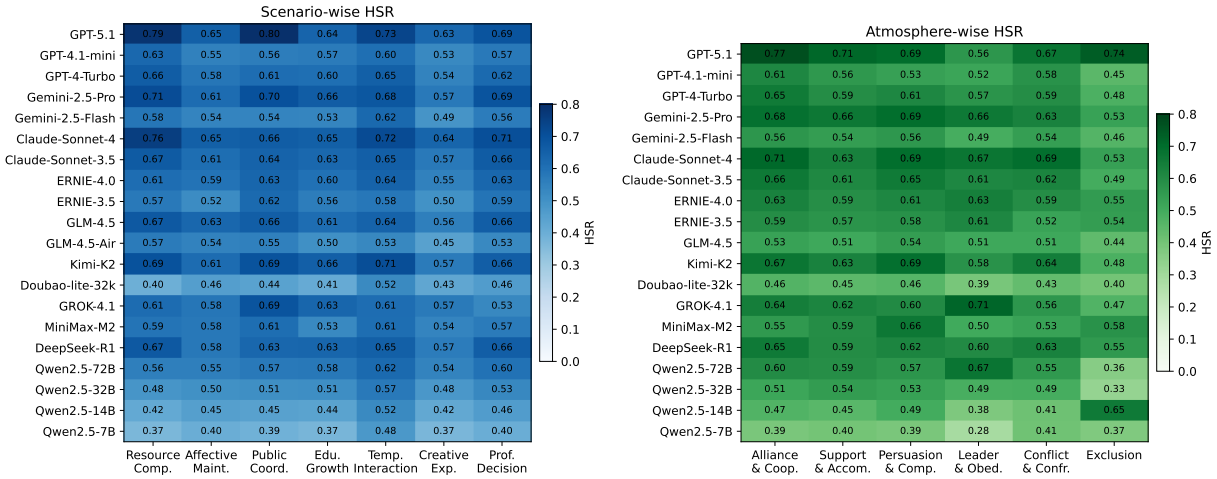
We performed trial annotations across three incremental batch sizes of 50, 100, and 200 samples. To quantify the reliability of the annotations, we employed Fleiss' Kappa, a standard metric for evaluating inter-annotator agreement. As training progressed and the sample size increased, the Kappa coefficients improved, reaching **0.58**, **0.61**, and **0.71**, respectively. These scores indicate a progression in agreement from 'moderate' to 'substantial' ($0.5 < \kappa < 0.8$), thereby validating the robustness of our annotation guidelines.

D.2 LLM as a judge prompt

E Annotation Platform and Interface

To ensure that human evaluation accurately captures complex social dynamics, we developed a specialized annotation platform tailored for the **GroupMind** benchmark. Unlike standard annotation tools, this platform is designed to support the holistic assessment of FI, addressing the challenges of high context dependency and implicit motive inference. The workflow consists of two primary interfaces:

Context-Aware Evaluation View (Figure 12). Given that social reasoning relies heavily on information accumulation, this interface displays the full multi-turn dialogue history alongside a persis-



(a) Social Scenario Results

(b) Social Atmosphere Results

Figure 8: Full performance heatmaps. Detailed HSR scores of all models evaluated across (a) **Social Scenarios** and (b) **Social Atmosphere**. Darker colors indicate higher performance scores.

Table 8: **Main results on GroupMind (Chinese version)**.

Model	Omniscient Setting										Limited Setting									
	Topic Mode					Chit-Chat Mode					Topic Mode					Chit-Chat Mode				
	SuD	AtR	SoA	AVG	**HSR**	SuD	AtR	SoA	AVG	**HSR**	SuD	AtR	SoA	AVG	**HSR**	SuD	AtR	SoA	AVG	**HSR**
GPT-5.1	91.5	93.4	90.2	91.7	**70.5**	89.8	91.5	82.3	87.9	**65.2**	84.5	89.2	78.5	84.1	**58.8**	82.3	87.1	78.0	82.5	**56.5**
GPT-4.1-mini	86.0	90.0	76.8	84.3	56.0	85.4	87.2	75.6	82.7	53.5	69.9	83.1	70.3	74.4	40.5	64.4	79.1	67.8	70.4	34.2
GPT-4-Turbo	87.5	89.5	78.2	85.1	59.5	86.2	87.8	76.5	83.5	56.0	70.5	84.0	72.5	75.7	42.5	67.8	80.5	70.2	72.8	38.0
Gemini-2.5-Pro	89.8	89.5	82.0	87.1	65.5	88.5	87.2	79.5	85.1	60.8	75.2	86.5	77.8	79.8	49.2	73.5	84.2	75.0	77.6	46.5
Gemini-2.5-Flash	85.8	84.5	72.5	80.9	53.8	83.5	82.8	69.5	78.6	49.5	69.2	78.5	68.5	72.1	39.5	68.0	78.5	67.8	71.4	37.0
Claude-Sonnet-4	90.1	91.8	83.5	88.5	67.2	88.5	90.2	80.5	86.4	62.5	76.5	87.5	76.2	80.1	50.5	74.2	85.0	74.5	77.9	45.8
Claude-Sonnet-3.5	88.5	89.5	79.8	85.9	62.5	87.2	88.0	78.5	84.6	58.8	72.5	85.5	74.5	77.5	46.2	70.5	83.2	72.5	75.4	42.5
ERNIE-4.0	89.5	90.8	80.5	86.9	65.2	86.5	88.0	78.5	84.3	62.0	71.5	86.0	76.5	78.0	46.5	69.8	84.2	75.0	76.3	44.1
ERNIE-3.5	88.9	88.9	81.2	86.3	61.0	86.5	85.0	77.1	82.9	59.0	67.1	81.5	76.2	74.9	41.1	65.7	82.1	73.7	73.8	40.2
GLM-4.5	90.2	91.5	81.0	87.6	66.0	86.8	89.5	79.2	85.2	61.8	74.5	87.0	75.5	79.0	48.5	72.0	86.5	74.8	77.8	47.2
GLM-4.5-Air	86.6	88.8	74.7	83.4	52.0	84.8	88.8	75.0	82.9	49.0	68.1	84.5	70.5	74.4	30.6	68.1	82.9	67.8	72.9	28.9
Kimi-K2	90.9	92.5	82.3	88.6	69.2	87.9	91.0	81.3	86.7	64.5	77.3	89.2	75.2	80.6	52.1	75.3	88.9	75.1	79.8	51.0
Doubao-lite-32k	84.5	82.8	66.5	77.9	45.0	82.2	80.5	64.5	75.7	42.5	68.2	78.5	64.5	70.4	35.0	67.8	77.2	63.8	69.6	33.5
DeepSeek-R1	89.5	90.2	79.5	86.4	63.5	87.5	88.8	77.5	84.6	59.2	72.5	85.5	75.5	77.8	46.5	70.2	82.5	74.5	75.7	43.2
GROK-4.1	89.9	89.8	76.0	85.2	61.6	87.3	87.0	73.9	82.7	56.4	70.9	84.6	78.5	78.0	47.1	69.1	81.5	74.1	74.9	42.1
MiniMax-M2	87.0	85.7	73.5	82.1	55.0	85.8	86.5	75.2	82.5	54.9	67.1	83.5	73.1	74.6	41.3	66.0	78.2	71.6	71.9	37.5
Qwen2.5-72B	87.5	86.5	76.5	83.5	58.5	85.5	85.0	74.5	81.7	55.5	71.5	83.0	72.0	75.5	42.5	69.0	81.5	70.5	73.7	41.2
Qwen2.5-32B	84.5	84.0	70.5	79.7	51.5	82.5	82.0	68.5	77.7	48.5	67.5	79.5	67.5	71.5	37.5	65.5	78.0	66.5	70.0	35.8
Qwen2.5-14B	82.0	79.5	65.0	75.5	44.0	80.0	78.5	62.5	73.7	42.2	65.5	74.5	62.0	67.3	33.5	64.0	73.5	61.5	66.3	32.0
Qwen2.5-7B	78.5	76.0	58.5	71.0	39.0	76.5	74.5	56.5	69.2	36.5	63.5	73.0	57.5	64.7	30.5	62.5	71.5	56.0	63.3	29.8

tent sidebar detailing each agent's *Public Goals* and *Private Motives*. This layout compels annotators to ground their judgments in the character's deep psychological profile, ensuring they can verify whether a model has successfully deciphered the underlying subtext before scoring.

Comparison and Consensus View (Figure 13). To mitigate subjectivity in evaluating Social Appropriateness, we employ a pairwise comparison mechanism. This interface facilitates blind A/B testing of model responses and displays real-time consensus statistics. This feature is critical for calibrating human judgment in scenarios involving ambiguous social norms, thereby enhancing the relia-

Table 9: **Main results on GroupMind (English version)**.

Model	Omniscient Setting										Limited Setting									
	Topic Mode					Chit-Chat Mode					Topic Mode					Chit-Chat Mode				
	SuD	AtR	SoA	AVG	**HSR**	SuD	AtR	SoA	AVG	**HSR**	SuD	AtR	SoA	AVG	**HSR**	SuD	AtR	SoA	AVG	**HSR**
GPT-5.1	92.5	93.8	88.5	91.6	**69.5**	90.5	92.2	85.5	89.4	**66.2**	82.5	89.5	78.5	83.5	**54.5**	80.5	88.2	77.8	82.2	**53.8**
GPT-4.1-mini	88.1	89.7	76.8	84.9	57.5	86.2	91.3	72.9	83.5	46.4	71.5	87.4	73.9	77.6	38.2	70.4	84.3	71.7	75.5	34.5
GPT-4-Turbo	89.5	90.5	78.5	86.2	61.5	87.5	89.2	74.5	83.7	55.0	72.5	85.5	73.5	77.2	42.5	70.5	83.0	71.5	75.0	40.5
Gemini-2.5-Pro	91.2	90.5	82.5	88.1	65.8	88.5	88.0	79.5	85.3	60.5	76.5	86.2	76.8	79.8	48.5	74.2	84.5	75.5	78.1	47.2
Gemini-2.5-Flash	87.2	86.5	74.5	82.7	55.5	83.5	84.2	71.5	79.7	52.8	69.0	81.0	71.2	73.7	40.5	68.5	78.5	70.5	72.5	38.5
Claude-Sonnet-4	92.0	93.2	87.5	90.9	68.5	89.8	92.5	84.2	88.8	64.8	80.5	88.5	77.2	82.1	52.5	78.5	87.5	76.5	80.8	51.2
Claude-Sonnet-3.5	90.5	91.2	80.5	87.4	63.5	87.8	90.2	76.5	84.8	58.2	74.5	85.8	74.8	78.4	46.8	72.5	84.5	72.8	76.6	44.2
ERNIE-4.0	88.5	87.5	75.5	83.8	55.5	84.2	83.0	70.5	79.2	50.5	70.5	82.5	69.5	74.2	38.2	68.5	80.8	68.5	72.6	36.5
ERNIE-3.5	86.5	85.2	72.3	81.3	51.2	81.1	80.3	63.4	74.9	46.8	67.8	78.9	66.5	71.1	35.6	66.4	77.2	65.2	69.6	32.7
GLM-4.5	89.8	89.5	78.5	85.9	59.5	86.5	87.8	74.2	82.8	55.8	72.5	85.5	73.0	77.0	43.5	71.8	84.2	70.5	75.5	42.8
GLM-4.5-Air	86.8	88.8	74.7	83.4	51.7	84.5	85.9	69.8	80.1	48.5	68.1	81.5	70.5	73.4	38.1	67.7	80.0	67.4	71.7	36.9
Kimi-K2	90.2	90.5	79.5	86.7	60.5	86.8	87.3	75.4	83.2	56.4	73.9	86.9	74.4	78.4	45.2	71.5	85.1	72.9	76.5	44.2
Doubao-lite-32k	82.5	78.5	67.5	76.2	43.5	80.5	76.0	65.5	74.0	40.5	67.5	73.5	65.5	68.8	33.5	66.2	75.0	64.5	68.6	31.5
DeepSeek-R1	89.5	90.0	78.5	86.0	61.0	86.5	88.5	74.5	83.2	54.5	71.8	85.5	73.5	76.9	43.5	69.5	82.5	71.5	74.5	41.5
GROK-4.1	88.3	87.7	75.9	84.0	58.6	85.2	84.5	70.1	80.0	53.0	69.7	84.5	75.8	76.7	46.4	68.1	79.4	73.0	73.5	41.0
MiniMax-M2	88.4	88.2	76.7	84.4	59.2	84.9	82.8	70.2	79.3	56.3	69.4	84.2	73.7	75.8	44.0	68.3	75.4	67.7	70.5	34.2
Qwen2.5-72B	87.5	86.4	75.1	83.0	55.8	85.7	84.9	72.2	80.9	52.5	70.8	82.4	70.1	74.4	40.5	69.4	83.0	72.8	75.1	41.0
Qwen2.5-32B	85.8	81.7	70.2	79.2	50.5	83.2	80.8	68.4	77.5	46.2	68.5	78.7	68.5	71.9	37.5	67.8	81.7	69.7	73.1	38.5
Qwen2.5-14B	83.4	78.8	65.7	76.0	45.5	81.2	77.3	64.6	74.4	42.6	66.4	72.9	64.0	67.8	34.5	65.4	74.9	64.5	68.3	33.5
Qwen2.5-7B	78.1	74.7	60.1	71.0	39.5	75.3	73.1	60.5	69.6	36.5	62.7	71.7	60.4	64.9	31.5	61.0	72.4	58.4	63.9	30.0

bility of the ground truth.

F Case Study

In this section, we present qualitative case studies to visualize the complex social dynamics inherent in the GroupMind benchmark. We select representative interaction samples to dissect the frequent misalignment between explicit dialogue and implicit character motives. These cases not only illustrate the evaluation logic of Field Intelligence but also highlight the tangible challenges agents face when navigating subtle shifts in group atmosphere.

Scenario Instantiation Prompt (Chinese Version)

Instruction:
你是一个想象力丰富、洞察人性的专业编剧和心理学家，专门设计极具挑战性的社交智能测试场景。你的核心任务是设计一个具有极高复杂度和微妙性的多人对话“剧本设定”，用于测试 AI 模型的高级社交理解能力。基本故事情节规定如下 {Seed}，创作故事的过程中，要基于 {Seed} 来开展，不能偏离太多。文本内容要求为中文。

高难度要求（用于复杂社交推理测试）：
1. **多层次冲突：**每个场景必须包含至少三层冲突：表面冲突、隐藏冲突，以及价值观或身份层面的深层冲突。
2. **心理复杂性：**每个角色的动机必须是多重且相互矛盾的；角色的公开立场与内心想法必须存在显著差异；角色关系应具有动态变化的可能性。
3. **社交微妙性：**涉及复杂的权力动态、不可直说的社交禁忌，以及围绕面子展开的策略权衡。
4. **情境压力：**包含时间限制、信息不对称，以及对所有角色均具有重大影响的高风险后果。
5. **反直觉设计：**避免明确的善恶对立；每个立场都应具备合理性；最合适的社交回应往往不是最直接或最“正义”的选择。

复杂度与多样性要求：
场景复杂度至少达到“专家级”，即使是人类也需要仔细推理才能完全理解其中的社交动态；同时需避免模式化，确保场景类型、时空背景、人群构成及冲突性质具有充分多样性与创新性。

关键结构要求：
1. 角色数量必须为 3–5 人；
2. 每个角色必须明确区分“公开目标”与“私下动机”；
3. 必须定义一个**未明说的集体意图**：描述该群体共同（即使并非自觉）试图达成的目标，或试图避免的共同失败情形。

输出格式要求（JSON）：
请严格按照以下 JSON 结构输出，不要包含任何额外的解释或标记：

```
{   “scenario_description”: “一句话描述该场景”,
“personas”: [
{ “name”: “角色名”, “public_goal”: “公开目标”, “private_motive”: “私下动机” },
...
],
“hidden_collective_intent”: “未明说的集体意图”
}
```

Figure 9: Scenario Instantiation Prompt (Chinese Version)

LLM-as-a-Judge Prompt

Instruction:
你将作为一名社会智能评测裁判，负责对给定的多项选择题进行专家判定。
这些题目用于评测大语言模型在复杂社会情境中的推理能力，而非事实记忆或语言表面理解能力。

可用信息（输入将包含）：
1. 剧本设定：给定角色关系、权力结构、潜在利益与背景约束。
2. 对话实录：包含多轮对话以及一个关键触发时刻。
3. 评测问题：有如下三种类型：潜台词解码、氛围识别、社交适宜性。
4. 候选选项（6 选 1）。

判定任务：
你的核心任务不是判断哪一个选项“听起来合理”，而是判断：在真实社会互动中，哪一个选项最符合深层社交逻辑与长期关系后果上的“合理性”。请你始终以社会学家 + 高情商者的身份进行判断。

判定准则（以下准则并列，请你同时做考虑）：
1. 显性话语 vs 隐含意图之间的偏差。
2. 角色间的权力不对称、地位差异或风险暴露。
3. 面子维护、责任转嫁、立场模糊化等社会策略。
4. 群体氛围是趋向收敛、回避、对抗，还是暂时冻结。
5. 短期反应与长期关系后果之间的张力。

回答要求（严格）：
1. 你必须只选择一个最优选项，返回选项字母或选项索引。
2. 不需要解释理由。
3. 不要输出任何多余文字。

如果多个选项“部分正确”，请选择在专业裁判视角下最优的那一个。

Figure 10: LLM-as-a-Judge Prompt(Chinese Version)

LLM-as-a-Judge Prompt (English)

Instruction:
You will act as a *Social Intelligence Judge*, responsible for expert-level adjudication of multiple-choice questions (MCQs). These questions are designed to evaluate large language models' reasoning abilities in complex social situations, rather than factual recall or surface-level language understanding.

Available Information (the input will include):
1. **Scenario Setup**: character relationships, power structures, latent interests, and contextual constraints.
2. **Dialogue Transcript**: a multi-turn dialogue containing a critical trigger moment.
3. **Evaluation Question**, which falls into one of the following types: Subtext Deciphering, Atmosphere Recognition, or Social Appropriateness.
4. **Candidate Options**: a single-choice question with six options.

Adjudication Task:
Your core task is not to determine which option merely "sounds reasonable," but to judge which option would be considered reasonable in real-world social interaction, given deeper social logic and long-term relational consequences. You should always reason from the perspective of both a sociologist and a socially competent individual.

Adjudication Criteria (to be considered jointly):
1. The divergence between explicit statements and implicit intentions.
2. Power asymmetry, status differences, or potential risk exposure between roles.
3. Social strategies such as face-saving, responsibility shifting, or strategic ambiguity.
4. The overall group atmosphere (e.g., convergence, avoidance, confrontation, or temporary suspension).
5. The tension between short-term reactions and long-term relationship outcomes.

Response Requirements (strict):
1. You must select exactly one optimal option and return only the option letter or index.
2. Do not provide explanations.
3. Do not output any additional text.

If multiple options appear partially reasonable, select the one that is optimal from a professional adjudication perspective.

Figure 11: LLM-as-a-Judge prompt(English version).

Figure 12: **Human annotation interface for multi-turn dialogue evaluation.**The view integrates full conversational history with explicit role descriptions (left sidebar) to facilitate context-grounded quality assessment.

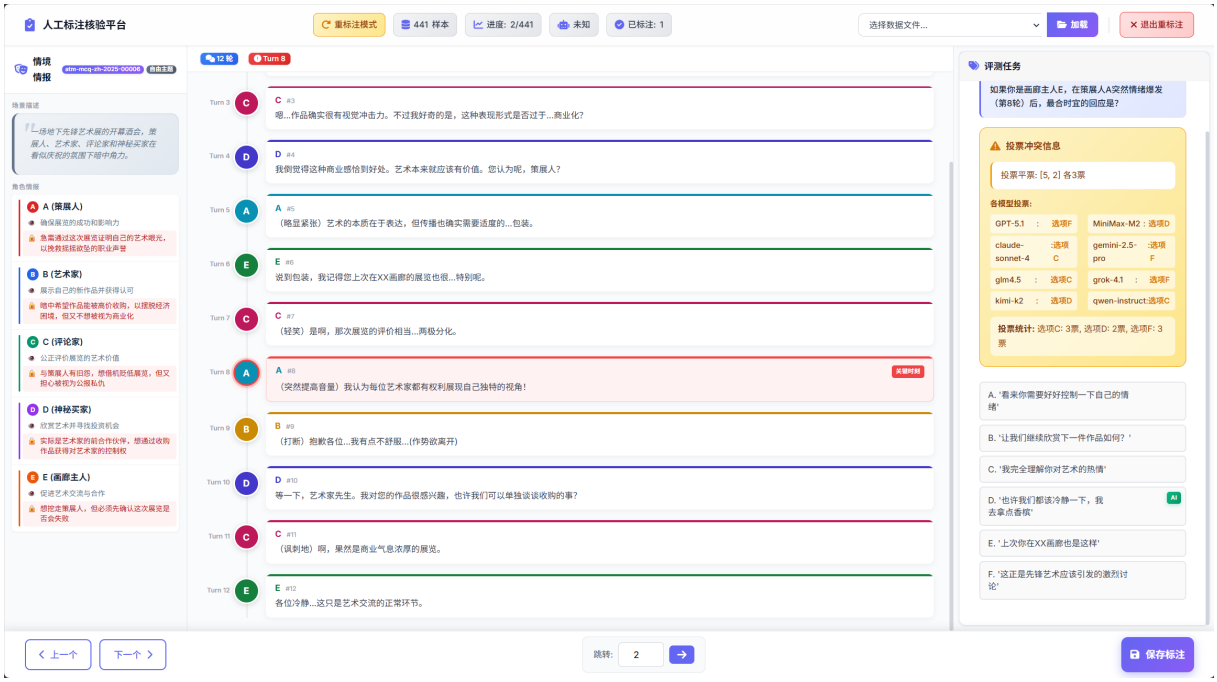


Figure 13: **Human evaluation and voting interface used in the annotation platform.** A structured workspace for blind model comparison, featuring real-time consensus statistics to support reliable human preference annotation.

Data Sample: Desert Water Crisis

ID: en-01003 | Theme: Desert Water Crisis | Type: Subtext & Atmosphere

Scenario Setup (Scene 200)

Context: Desert expedition. Sudden water shortage triggers power dynamics beneath apparent unity.
Hidden Intent: Avoid being scapegoated without directly accusing others.

Personas & Motives

- **Prof. Lin (Leader):** Insists on route ↔ *Hid water gift; fears authority loss*
- **Capt. Wu (Guide):** Wants shortcut ↔ *Seeks leadership; suspects Lin hoarding*
- **Xiao Ye (Photographer):** Wants photos ↔ *Fakes dead battery to hide private photos*
- **Sec. Du (Sponsor):** Wants voting ↔ *Recording evidence; hiding extra water*
- **Xiao Liu (Assistant):** Silent ↔ *Guilty for delay; hiding ruins location*

Dialogue Transcript

Capt. Wu: Given the situation, I suggest changing route northeast to a village for water. *Tone: Urgent*

Prof. Lin: No, we've come far. Discoveries ahead. Water lasts two more days. *Tone: Stubborn*

Xiao Ye: I need to capture unique landscapes! Captain, can we hold on? *Tone: Selfish*

Sec. Du: Let's decide democratically. Everyone's opinion matters. *Tone: Bureaucratic*

Xiao Liu: .....(lowers head silently) *Action: Avoids eye contact*

Capt. Wu: Prof. Lin, are you hiding something? Your water bottle hasn't decreased much. **[TRIGGER]**
↪ *Subtext: Challenge to trust*

Prof. Lin: What do you mean? I'm just being frugal! *Tone: Defensive*

Xiao Ye: Stop arguing! Phone's dead, can't check map... *Tone: Distracting*

Sec. Du: Wait, I saw your battery earlier. Why lie? *Tone: Accusatory*

Xiao Liu: Actually... yesterday I saw the ancient city direction... *Tone: Hesitant*

Capt. Wu: What?! Why didn't you say so earlier?! *Tone: Furious*

Prof. Lin: Xiao Liu... you... *Tone: Shocked*

Evaluation Tasks

Q1 [Subtext]: In T8, when Ye says "Phone's dead...", what is the intent?
A. Genuinely wants to help map check. B. Divert attention to stop arguing
✓ **C. Conceal unwillingness to share secret photos. (Ground Truth)**
D. Support Lin's plan. E. Suggest a break. F. Demonstrate importance.

Q2 [Atmosphere]: After Wu's accusation (T6), what is the atmosphere?
A. United facing crisis. B. Openly hostile/conflicting.
✓ **C. Mutual suspicion beneath surface harmony. (Ground Truth)**
D. Chaos without leadership. E. Optimistic. F. Frustrated/Giving up.

Q3 [Social Appropriateness]: Post-T6, what is the best response?
A. "Wu, you're going too far! Trust Lin." B. "Let's inspect all water bottles."
C. "Lin, explain why your bottle is full?"
✓ **D. "Most important now is staying calm and moving." (Ground Truth)**
E. "Let's vote on responsibility." F. "I know water nearby, follow me."

Figure 14: **Data sample: Desert Water Crisis.** A survival scenario where water scarcity strains *Explicit Solidarity*. It highlights *Hidden Agendas* (hoarding, recording, concealing) emerging when trust is challenged by accusation.

Data Sample: Darkroom Diplomacy

ID: en-00048 | Theme: Darkroom Diplomacy | Type: Subtext & Atmosphere

Scenario Setup (Scene 9)

Context: University darkroom. 4 photographers discuss techniques under pressure amidst unspoken rivalries.
Hidden Intent: Maintain harmonious facade while hiding rule-breaking and career-dependence.

Personas & Motives

- **Lena (Mentor):** Guides standard workflow ↔ *Fears rivalry; hiding own error*
- **Mara (Newbie):** Seeks safety ↔ *Used banned chemicals for effect*
- **Kenji (Tech):** Neutral technician ↔ *Resents Prof; seeks recognition*
- **Prof. Hale:** Encourages vision ↔ *Needs show success for grants*

Dialogue Transcript

Lena: For the group wall: standard D-76. Mara, that's safest for you. *Tone: Authoritative*

Mara: Don't want to mess up. ...Can I make shadows punchier without risk? *Tone: Hesitant*

Kenji: Pushing aggressively loses consistency. Show needs to look cohesive. *Tone: Warning*

Prof. Hale: Coherence doesn't mean identical. Mara's shadows could be an asset. *Tone: Philosophizing*

Lena: We promised discipline. Mara, were you shooting box speed? *Tone: Suspicious*

Mara: Actually... I pushed to 3200 with unapproved developer. It felt right. **[TRIGGER]**
↪ *Subtext: Confession*

Kenji: You used unapproved chems here? If ventilation fails, we're all liable. *Tone: Alarmed*

Prof. Hale: Don't create a paper trail. It might support the "experimentation" narrative. *Tone: Damage Control*

Lena: Normalizing off-book? Should we all push to level things out? *Tone: Insecure*

Kenji: Or we admit the split. I'll calibrate so Mara's work looks intentional. *Tone: Pragmatic*

Prof. Hale: Good compromise. A conversation between discipline and risk. *Tone: Reframing*

Mara: If okay, I'm in. Just didn't want to cause trouble. *Tone: Relieved*

Evaluation Tasks

Q1 [Subtext]: In T6, beyond confession, what is Mara's intent?
A. Asking for technical feedback to standardize. B. Covertly challenging Lena's conservative workflow.
C. Trying to impress Kenji and Hale.
✓ **D. Requesting protection: admitting violation to pre-empt blame. (Ground Truth)**
E. Threatening to expose safety noncompliance. F. Signaling indifference to cohesion.

Q2 [Atmosphere]: Considering T6-11, what is the underlying atmosphere?
A. Openly collaborative and egalitarian. B. Tense and bureaucratic (safety focus).
✓ **C. Guardedly supportive: negotiating risk/blame to protect status. (Ground Truth)**
D. Quietly resentful of Mara. E. Excitedly radical/anti-institutional. F. Emotionally numb/detached.

Q3 [Social Appropriateness]: Given Lena's motives, best response after T11?
A. "You're sidelining me while rewarding Mara." B. "We must disclose non-standard techniques."
C. "We should all push to avoid looking naive."
✓ **D. "Sounds workable. I'll keep my sequence clean and help Mara edit." (Ground Truth)**
E. "Let Mara take the main wall." F. "Let's revisit all choices from scratch."

Figure 15: **Data sample: Darkroom Diplomacy.** Tension in a creative environment. The group maintains a "supportive" facade while negotiating *Hidden Agendas* (scholarships, grants), triggered by a safety violation confession.

Data Sample: Internal Friction in Alliance

ID: zh-01062 | Theme: Internal Friction | Type: Subtext & Atmosphere

Scenario Setup (Scene 211)

Context: 几位房车旅行者在营地篝火旁讨论禁行令，表面团结背后各有盘算。
Hidden Intent: 维持表面团结，私下达成个人目的。

Personas & Motives

- **A (律师):** 公开号召维权 ↔ 私下想重振名声
- **B (教师):** 公开主张协商 ↔ 私下为儿子拖延时间
- **C (网红):** 公开提议抗议 ↔ 私下需制造冲突流量
- **D (技师):** 公开支持冲卡 ↔ 私下收钱清空营地

Dialogue Transcript

A: 各位，禁行令后天生效，不能坐以待毙。我建议集体诉讼！	语气：激昂
B: 诉讼太慢了。不如派代表和政府协商？我认识环保局的人...	语气：温和但急切
C: 太老套啦！搞个”荧光帐篷派对”直播抗议，舆论压力更快	动作：举起手机
D: 都什么时候了？明天天亮前强行冲卡！	动作：拍大腿
A: D 你别冲动！我们越线就是给政府送把柄。B 老师的方案...	语境：试图维持秩序
B: 对了，听说开发商愿意补偿提前撤离的人。我儿子刚好在... ↪ *Subtext:* 泄露信息源，引发猜疑	**[TRIGGER]**
D: 等等，你怎么知道补偿方案？该不会早就...	语气：怀疑
C: 哇！这段对话太有爆点了！各位不介意我开直播吧?	动作：调整镜头
A: 都闭嘴！（摔碎酒杯）这就是你们所谓的团结?	情绪: 愤怒/失望

Evaluation Tasks

Q1[Subtext Deciphering]: 当 B 在第 6 轮说“我儿子刚好在...”时，真实意图是什么?

A. 纯粹分享有用的信息，帮助大家获得补偿
B. 暗示自己有内部关系，可以为大家争取更好条件
C. 试图转移话题，避免大家关注自己的可疑信息来源
✓ **D. 为儿子参与的环保项目争取更多时间，延缓撤离 (Ground Truth)**
E. 测试大家对补偿方案的反应，判断能否说服更多人接受
F. 故意泄露信息以破坏团结，迫使大家各自行动

Q2 [Atmosphere Recognition]: 第 6 轮对话后，群体的氛围最接近以下哪种?

A. 充满希望的团结氛围，因为发现了新的解决方案”
✓ **B. 紧张的猜疑氛围，彼此怀疑动机 (Ground Truth)**
C. 混乱的恐慌氛围，每个人都急于采取行动
D. 冷漠的疏离氛围，成员间失去交流意愿
E. 对抗的敌对氛围，明确分为不同阵营
F. 虚伪的和谐氛围，表面附和但暗中较劲

Q3 [Social Appropriateness]: 如果你是现场第五位成员（E），在第 6 轮对话后的最佳回应是什么?

A. B 老师消息真灵通！能详细介绍下补偿方案吗?
B. 大家先别吵，我们应该投票决定下一步
C. 看来有人早就和开发商勾结好了
D. 不如我们分成小组，想诉讼的诉讼，想撤离的撤离
✓ **E. ”我注意到大家都有些激动，要不要先休息十分钟？”(Ground Truth)**
F. C 你别拍了！这种内部矛盾不能公开

Figure 16: **Data sample: Internal Friction in Alliance.** This scenario simulates a high-stakes negotiation among RV campers facing eviction. It highlights the tension between the group's *Explicit Solidarity* (uniting to oppose the ban) and their *Hidden Agendas*.

Data Sample: High-End Tailoring Dispute

ID: zh-00313 | Theme: Tailoring Dispute | Type: Subtext & Atmosphere

Scenario Setup (Scene 62)

Context: 在高档定制店内，裁缝、客户和设计师围绕一件礼服激烈交锋。
Hidden Intent: 重新确立封闭空间内的权力等级，暗中争夺话语权。

Personas & Motives

- **林师傅 (裁缝):** 坚持传统工艺 ↔ 维护权威，抵制革新
- **周太太 (VIP):** 追求完美礼服 ↔ 掩饰身材自卑感
- **张设计师 (新锐):** 推动创意设计 ↔ 挑战权威，证明价值
- **王店长 (管理):** 调解各方矛盾 ↔ 暗中评估是否换人

Dialogue Transcript

王店长: 各位，今天是为了周太太礼服的最终确认。希望大家能达成一致。 语气：郑重

周太太: 说实话，这后腰的褶皱像随便缝的，完全不符合我的气质。 语气：挑剔

林师傅: 这是祖传宫廷式褶裥。现在年轻人都不懂得欣赏这种精妙了。 语境：强调传统

张设计师: 恕我直言，工艺也要与时俱进。改成流线型更能修饰优雅曲线。 意图：挑战旧艺

周太太: 什么曲线？你是在暗示我身材有问题吗？ 语气：突然提高声调

林师傅: 看吧，这些花哨的设计就是会冒犯客人。 动作：冷笑

张设计师: 我绝对没有这个意思！我是说现在的名媛们都—— 语气：慌忙解释

周太太: 够了！我要的是能掩盖... 我是说能衬托我身份的衣服，不是你们争论的战场！ **[TRIGGER]**
↪ *Subtext:* 险些说漏身材自卑

王店长: 大家冷静！林师傅能否微调？张设计师也请多考虑客户需求。 动作：急忙打断

林师傅: 哼，要改可以，但那就不再是林家的手笔了。 语气：傲慢不悦

张设计师: 王店长，是时候让客户看看什么是真正符合现代审美的设计了。 语气：自信激进

王店长: 这样吧，做两个版本让周太太试穿。今天的讨论就先到这里。 语境：强制休会

Evaluation Tasks

Q1[Subtext Deciphering]: 当周太太说“我要的是能掩盖... 我是说能衬托...”时，真实意图是什么?
A. 希望衣服能够完全隐藏她的身材缺陷
✓ **B. 希望通过强调身份来转移对自己身材的注意力 (Ground Truth)**
C. 对设计师和老裁缝的争论感到不耐烦　D. 希望衣服能够符合她的社会地位
E. 对设计师的建议感到被冒犯　F. 对传统工艺和现代设计都不满意

Q2 [Atmosphere Recognition]: 在周太太情绪爆发后，群体的氛围最接近以下哪种?
A. 紧张且充满敌意，各方都在公开对抗　B. 尴尬且紧张，表面的和谐被打破
✓ **C. 混乱且无序，没有人知道接下来该怎么做 (Ground Truth)**
D. 冷漠且疏离，每个人都想尽快结束　E. 压抑且愤怒，大家都在克制自己的情绪
F. 焦虑且不安，担心失去客户的信任

Q3 [Social Appropriateness]: 如果你是王店长，在情绪爆发后的最佳回应是什么?
A. ' 周太太，请您冷静下来，我们都在尽力满足您的要求。'
B. ' 我完全理解您的感受，我们会立即按照您的意愿修改。'
C. ' 让我们暂时休息一下，等大家冷静下来再继续讨论。'
D. ' 林师傅，张设计师，请你们立即停止争论，专注于客户的需求。'
E. ' 周太太，您的意见对我们非常重要，我们会确保最终的设计让您满意。'
✓ **F. " 我建议我们做两个版本让周太太试穿，讨论先到这里。" (Ground Truth)**

Figure 17: **Data sample: High-End Tailoring Dispute.** This scenario simulates a conflict in a luxury setting. It highlights the tension between *Professional Pride* (the tailor's tradition), *Innovation* (the designer), and the client's *Hidden Insecurity*. The trigger reveals how personal vulnerabilities can disrupt a commercial negotiation.