Consensus-Driven Multi-Agent Cognitive Reasoning for Enhancing the Emotional Intelligence of Large Language Models

Geng Tu[1*], **Dingming Li**[3*], **Jun Huang**[3†], **Ruifeng Xu**[1,2,4†]
[1]Harbin Institute of Technology, Shenzhen
[2]Peng Cheng Laboratory
[3]University of Electronic Science and Technology of China
[4]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
22b951011@stu.hit.edu.cn, xuruifeng@hit.edu.cn

Abstract

Large Language Models (LLMs) have demonstrated strong performance in various NLP tasks but remain limited in emotional intelligence (EI). Benchmarks such as EmoBench attribute this gap to deficiencies in cognitively demanding tasks that require inferring others' latent mental states, intentions, and emotions in nuanced social contexts. To address this, we propose MACRo, a Multi-Agent Cognitive Reasoning framework that generates a structured Cognitive Chain of Thought comprising Situation, Clue, Thought, Action, and Emotion. Each component is generated by a specialized agent, enabling modular, interpretable multi-step reasoning. To ensure coherence and mitigate hallucinations, a coordinator agent verifies outputs, and a consensus game mechanism enforces alignment across reasoning steps. Extensive Experiments on EmoBench show that MACRo significantly enhances both emotional understanding and application across LLMs. Further evaluations confirm its generalizability to real-world social applications such as emotional support conversations.

Introduction

While Large Language Models (LLMs) perform well on tasks such as text generation, summarization, and question answering (Zhou et al. 2024a; Zhong et al. 2024), they often fall short in understanding and reasoning about human affective and cognitive states (Sabour et al. 2024).

In contrast to traditional emotion recognition tasks (Tu et al. 2024; Liu et al. 2025; Tu et al. 2025) that primarily focus on identifying surface-level affective cues, emotional intelligence (EI) involves higher-order abilities such as inferring latent mental states, appraising social context, and managing interpersonal dynamics (Salovey and Mayer 1990). These advanced competencies are critical for applications requiring nuanced social reasoning and human-centered interaction (Zhou et al. 2024b).

Recent evaluations, especially from EmoBench (Sabour et al. 2024), highlight significant limitations in the EI of LLMs. While LLMs exhibit some competence in recognizing explicit emotional cues, they still struggle with cognitively demanding tasks such as perspective taking, which involves inferring others' latent mental states, intentions, and emotions within nuanced social contexts. EmoBench conceptualizes EI into two key dimensions, emotional understanding (EU) and emotional application (EA), and finds that LLMs tend to perform better on tasks requiring straightforward EA but underperform on those demanding deeper reasoning about others' emotions and perspectives. This pattern aligns with observed weaknesses in Theory of Mind (ToM) related capacities (Strachan et al. 2024; Kosinski 2024), suggesting the EI gap stems from a lack of fundamental cognitive modeling rather than prompting or data alone.

Despite its importance, ToM is often dismissed as a heterogeneous psychometric construct (Wang et al. 2023), leaving its potential as a theoretical foundation for enhancing the EI of LLMs largely underexplored. A potential solution is to inject ToM-inspired reasoning into LLMs via structured cognitive chains, generated by a trained cognitive language model (COLM) (Wu et al. 2024). However, our experiments show that COLM struggles to generalize in EI tasks, often generating garbled outputs. As a result, the low reliability of these chains limits their utility in enhancing the EI of LLMs.

Similarly, LLMs without dedicated training often produce hallucinated or incoherent reasoning chains. As illustrated in Figure 1, the generated `thought` speculates about Jamie's intention rather than expressing Alex's emotional response, breaking the emotional continuity and leading to a disconnect from the preceding `clue`. Such breakdowns reflect a deeper limitation: the absence of mechanisms that enforce stepwise coherence in generated reasoning. We refer to this coherence as consensus, a verifiable alignment between successive elements in the cognitive chain. Without constraints to preserve this internal consistency, cognitive outputs tend to be fragmented and cognitively unreliable.

Motivated by the need for coherent and cognitively grounded reasoning, we propose a Multi-Agent Cognitive Reasoning framework (MACRo) for generating Cognitive Chain of Thought (CogCoT), a structured sequence of `Situation`, `Clue`, `Thought`, `Action`, and `Emotion` guided by ToM. Each specialized agent is responsible for generating a specific component in the chain, enabling interpretable and modular reasoning. Unlike standard CoT prompting, CogCoT enforces stepwise cognitive dependen-

*These authors contributed equally.
†Corresponding authors.

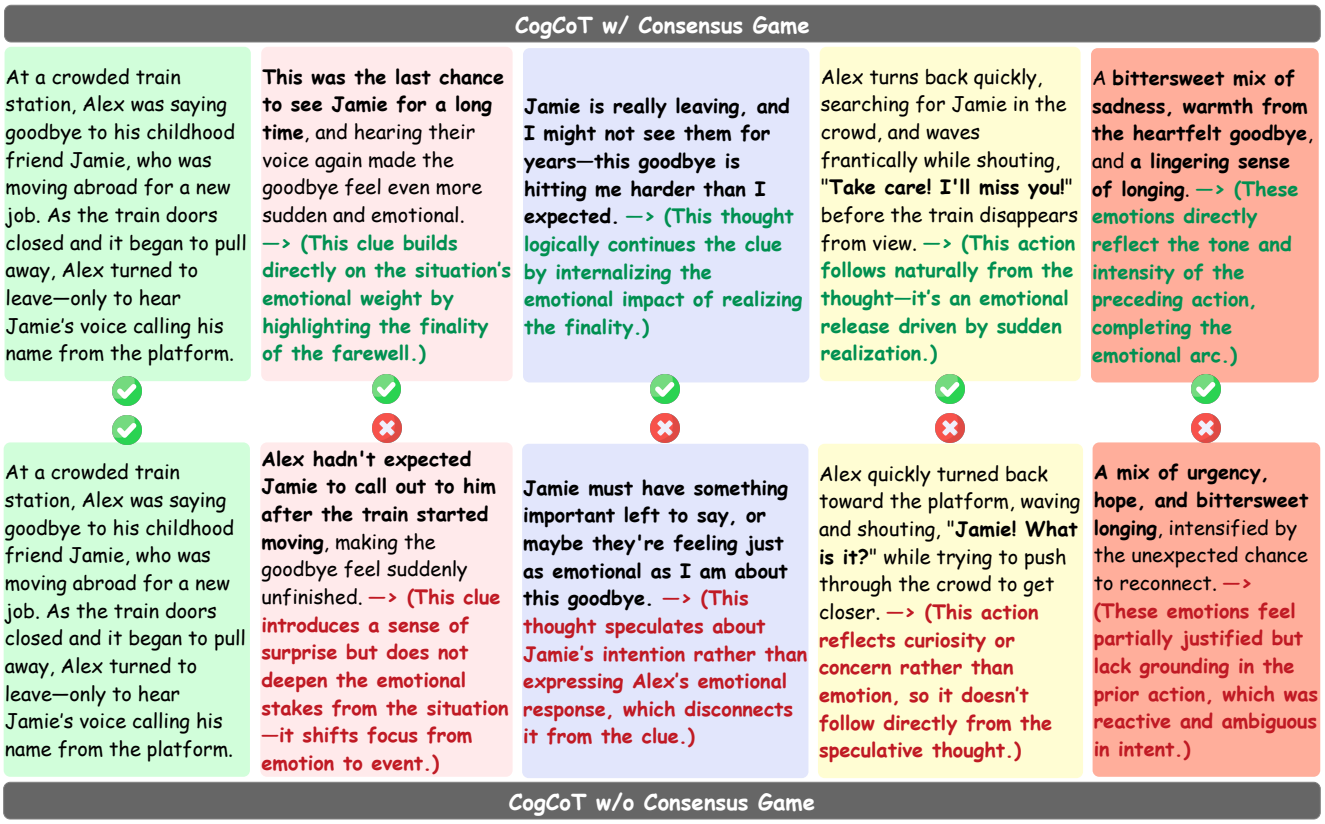

Figure 1: Illustration of two CogCoT examples with and without a consensus game. Each component is color-coded for clarity. Coherent and incoherent reasoning explanations are highlighted in green and red text, respectively.

cies and explicit structure that mirror human-like inference. A dedicated coordinator agent verifies output formats and produces the final summary, mitigating hallucinations and improving output consistency.

To further ensure internal coherence, we formulate CogCoT generation as a signaling game under imperfect information, where agents interact to align the semantic consistency of successive reasoning segments. Building on no-regret dynamics (Fudenberg and Levine 1995), we introduce a consensus game mechanism that iteratively refines both agents' policies to reach a Nash equilibrium, encouraging semantic alignment across reasoning steps. This method requires no additional model training and can be flexibly integrated with various LLMs.

In summary, our contributions are as follows:

- We propose the MACRo that generates a structured CogCoT to enhance the EI of LLMs.
- We propose a consensus game mechanism that leverages iterative policy refinement to achieve semantic coherence across reasoning steps via equilibrium convergence.
- We conduct comprehensive evaluations on EmoBench, demonstrating the effectiveness of our MACRo in enhancing the EI of LLMs and proving its generalizability to other social applications, such as the downstream Emotional Support Conversation (ESC) task.

Related Work

Emotional Intelligence of LLMs

Recent studies on the EI of LLMs have followed two main directions. The first focuses on incorporating psychological theories or standardized scales to construct public benchmarks for evaluating emotional understanding (Wang et al. 2023; Paech 2023; Huang et al. 2023), such as EmoBench (Sabour et al. 2024), which have been introduced to more comprehensively assess EI across multiple dimensions. The second involves fine-tuning LLMs on EI-related downstream tasks, mainly classification and regression, to improve their task-specific performance (Zhang et al. 2023; Lei et al. 2023; Liu et al. 2024; Li et al. 2024). However, existing approaches often rely on superficial emotional cues or statistical patterns, lacking deep cognitive reasoning that models the underlying mental states essential for robust EI.

Cognitive Knowledge for LMs

Cognitive knowledge is essential for intuitive reasoning in language models and is commonly represented as text-based knowledge graphs encoding event and object relations. Notable examples include ConceptNet (Speer, Chin, and Havasi 2017) for taxonomic facts; ATOMIC (Sap et al. 2019) and its extensions (Hwang et al. 2021; West et al. 2022; Kim et al. 2023) focusing on social commonsense; and NOVATOMIC (West et al. 2023), which uses natural

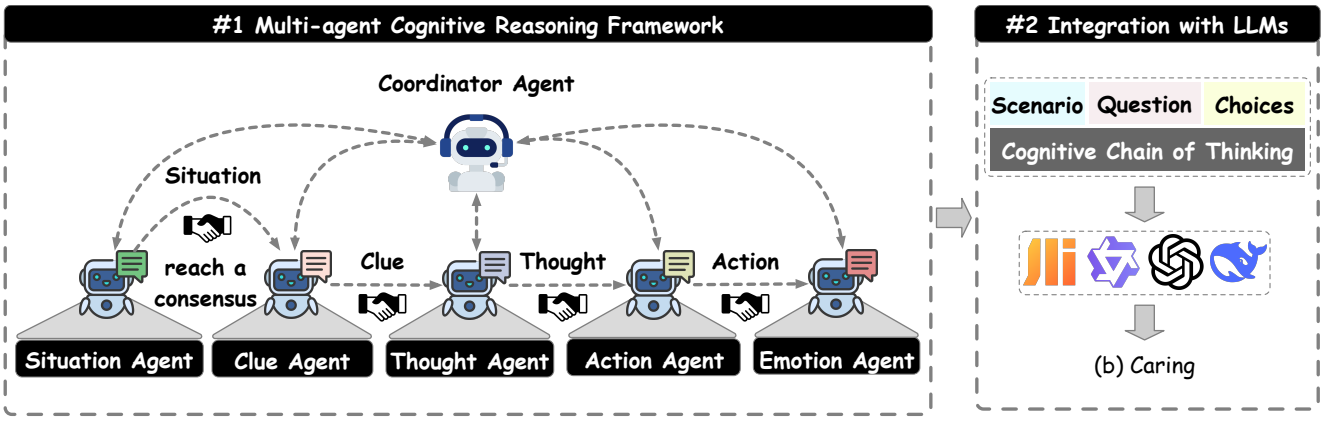

Figure 2: Overview of the MACRo for generating CogCoT. Specialized agents independently generate each component under coordinator control, with a consensus game ensuring coherence through equilibrium-based ranking.

language queries for general reasoning. These primarily capture event-centered commonsense, like temporal and causal relations, but offer limited explicit modeling of ToM, crucial for understanding human mental states. Recent work, such as COKE models ToM via a chained structure (Situation, Clue, Thought, Action/Emotion) to better mirror human social inference (Wu et al. 2024). However, existing work views ToM mainly as a simplistic, heterogeneous psychometric task unsuitable for standardized EI evaluation (Wang et al. 2023), thereby largely overlooking its potential as a foundational theory for improving the EI of LLMs.

Preliminaries

What is EI?

EI, originally defined by Salovey and Mayer (1990), refers to the ability to monitor one's own and others' feelings, differentiate between them, and use this information to guide thoughts and actions. Subsequent works have expanded this notion from different perspectives: Goleman (1996) and Bar-On (1997) emphasized emotional self-awareness, self-motivation, and social competence, while Schuller and Schuller (2018) highlighted emotion recognition, adaptation, and goal-oriented emotional reasoning. Despite these variations, EI fundamentally involves two key capabilities: understanding emotions and applying them effectively in context (Sabour et al. 2024).

Methodology

Task Definition

Following Sabour et al. (2024), EI is conceptualized into two dimensions: EU and EA, each defined through specific multiple-choice tasks. Specifically, the EU examines the model's ability to infer emotional states, underlying causes, intentions, and beliefs. It encompasses four task types: complex emotions (CE), personal beliefs and experiences (PBE), perspective taking (PT), and emotional cues (EC). EA assesses the model's capacity to apply emotional understanding by selecting contextually appropriate responses in various relational scenarios, including personal-self, personal-others, social-self, and social-others.

Multi-Agent Cognitive Reasoning Framework

In Figure 2, the MACRo employs a hierarchical architecture comprising a Coordinator and five role-specialized Executors. Each Executor Agent is responsible for generating one distinct segment of the CogCoT, collectively simulating the stepwise structure of human emotional reasoning. These roles include contextual grounding (**Situation**), trigger identification (**Clue**), cognitive appraisal (**Thought**), behavioral projection (**Action**), and affective synthesis (**Emotion**). This modular design enhances both the interpretability and controllability of the reasoning process.

Following Yi et al. (2025), the **Coordinator Agent** continuously monitors and verifies the intermediate outputs of Executors to reduce hallucinations and enforce format consistency. If an Executor's output meets quality standards, the process proceeds; otherwise, the Coordinator issues revision feedback. After all components of the CogCoT are generated, the Coordinator summarizes and refines the final reasoning chain to ensure global consistency across steps.

Consensus Game

To ensure the internal alignment within the CogCoT, the MACRo incorporates a game-theoretic consensus mechanism among Executor agents. This interaction is inspired by the piKL framework (Jacob et al. 2022), which coordinates symmetric policy distributions of a single model through KL-regularized no-regret updates over a shared output space (Jacob et al. 2024). In contrast to piKL's homogeneous and symmetric setting, the MACRo generalizes this to heterogeneous agents with distinct roles and disjoint output spaces, enforcing directional, stepwise coherence rather than symmetric consensus.

After the n-th agent generates its CogCoT segment, it passes the result as prior knowledge to the $(n+1)$-th agent. This agent then generates its segment either using or ignoring received prior knowledge, represented by latent decision $\mathcal{Z} \in \{\texttt{used}, \texttt{unused}\}$. Crucially, v is observed only by the

$(n+1)$-th agent (as the **Generator**), while the n-th agent (as the **Discriminator**) does not, and infer whether z=`used` from the generated response y. This aligns with the core principle of signaling games (Lewis 2008), where the Generator sends a signal y conditioned on hidden state z, and the Discriminator attempts to decode z based on y.

The Generator's initial policy $\pi_G^{(1)}(y \mid z)$ is computed from the model's generation probabilities conditioned on z, then normalized. The Discriminator's initial policy $\pi_D^{(1)}(z \mid y)$ is obtained by scoring each candidate y using a coherence-evaluation prompt relative to the preceding CogCoT segment. Both agents aim to maximize the shared utility:

$$u_G(\pi_G, \pi_D) = \frac{1}{2} \sum\nolimits_{z \in \mathcal{Z}} \sum\nolimits_{y \in \mathcal{Y}} \pi_G(y \mid z) \cdot \pi_D(z \mid y) \quad (1)$$

$$u_D(\pi_G, \pi_D) = \frac{1}{2} \sum\nolimits_{z \in \mathcal{Z}} \sum\nolimits_{y \in \mathcal{Y}} \pi_G(y \mid z) \cdot \pi_D(z \mid y) \quad (2)$$

where $\mathcal{Y}$ is the set of candidate responses. A Nash equilibrium (π_G^*, π_D^*) occurs when neither agent can improve its expected utility by changing its strategy alone (Monderer and Shapley 1996). This ensures both agents agree on how messages y map to latent decisions z, thereby promoting coherent generation. To avoid degenerate equilibria, regularization terms are added to penalize divergence from the initial policies $\pi_G^{(1)}$ and $\pi_D^{(1)}$, anchoring policies toward language model priors. At each iteration t, the Generator and Discriminator update their policies using feedback from the opponent's past behavior.

$$Q_G^{(t)}(z) = \frac{1}{2t} \sum\nolimits_{\tau=1}^{t} \pi_D^{(\tau)}(z = \texttt{used} \mid y) \quad (3)$$

$$Q_D^{(t)}(y) = \frac{1}{2t} \sum\nolimits_{\tau=1}^{t} \pi_G^{(\tau)}(y \mid z) \quad (4)$$

where τ indexes previous iterations. $Q_G^{(t)}(y \mid z)$ represents the Generator's estimate of the Discriminator's posterior probability of latent decision z given response y, averaged over past rounds. Similarly, $Q_D^{(t)}(z \mid y)$ denotes the Discriminator's estimate of the Generator's conditional probability of producing response y given z. Each agent then updates its policy via a KL-regularized exponential weighting rule that balances payoff feedback with adherence to initial policies:

$$\pi_G^{(t+1)}(y \mid z) \propto \exp\left(\frac{Q_G^{(t)}(y) + \lambda_G \log \pi_G^{(1)}(y \mid z)}{\frac{1}{\eta_G t} + \lambda_G} \right) \quad (5)$$

$$\pi_D^{(t+1)}(z \mid y) \propto \exp\left(\frac{Q_D^{(t)}(z) + \lambda_D \log \pi_D^{(1)}(z \mid y)}{\frac{1}{\eta_D t} + \lambda_D} \right) \quad (6)$$

where λ_G and λ_D are regularization coefficients that control how strongly the agents should adhere to their initial strategies. At the same time, η_G and η_D determine the rate at which the agents incorporate feedback over iterations.

Integration with LLMs

The generated CogCoT serves as a structured Chain-of-Thought prompt that can be seamlessly integrated into various LLMs to enhance EI by injecting step-by-step cognitive reasoning into the input. Furthermore, it generalizes to other social applications, such as the downstream ESC task.

Experiments

Datasets

We conduct our experiments on EmoBench, a comprehensive benchmark designed to evaluate the EI of LLMs. It includes 400 carefully crafted bilingual (Chinese-English) multiple-choice questions across two EI dimensions: EU and EA. The dataset demonstrates high reliability, with a Fleiss' Kappa of 0.852 from multi-annotator agreement.

Baselines

In our experiments, following Sabour et al. (2024), we evaluate a range of recent LLMs with strong benchmark performance. For **API-based LLMs** accessible via APIs, we include OpenAI's `GPT-3.5` (Brockman et al. 2015), `Deepseek-v3` (Guo et al. 2025), and `Baichuan2-53B` (Yang et al. 2023). For **locally deployable**, we include `LLaMA2-Chat-7B and 13B` (Touvron et al. 2023), `Baichuan2-Chat-7B and 13B`, `Qwen-Chat-7B and 14B` (Bai et al. 2023), as well as `ChatGLM3-6B and Yi-Chat-6B` (Young et al. 2024). Given the computational overhead of multi-turn reasoning in MACRo, we adopt `DeepSeek-V3` as a practical substitute for GPT-4 (Achiam et al. 2023), based on its demonstrated performance on EmoBench, which is comparable to that of GPT-4 (Sabour et al. 2024).

Experimental Settings

Setting of Baselines. For LLaMA-based models, we adopt default decoding settings with top-p sampling ($p = 0.9$) and temperature $= 0.6$. For other models, we use their default inference configurations via either public APIs or the HuggingFace Transformers interface.

Setting of CogCoT Generation. All CogCoT are generated in MACRo, using the base model itself without any additional training. In the consensus game, we set all game dynamics parameters to $\eta_D = \lambda_D = \eta_G = \lambda_G = 0.1$ across experiments for consistency, though further tuning may improve performance. The consensus game mechanism runs for 5000 iterations to select coherent reasoning steps. This is a numerical consensus optimization, not repeated LLM inference. In `Deepseek-v3`, 5,000 iterations take 3.3s, while 500 iterations take 0.05s with under 0.3% overall drop.

For CogCoT generated by COLM, since COLM checkpoints are unavailable, we reimplement and retrain COLM following the original paper's settings. Notably, COLM is trained on a dataset derived from a cognitive knowledge graph (Wu et al. 2024) rather than EmoBench. COLM generates each CogCoT component separately, which we then concatenate to form the complete reasoning chain.

Setting of Baselines with CogCoT. The final CogCoT is concatenated with the original prompt and fed back into the same base model for task completion.

Setting of ESC Models. We train both standard BlenderBot (Roller et al. 2021) and BlenderBot enhanced with CogCoT, following the experimental setup in (Wu et al. 2024) for fair comparison. CogCoT is generated by either COLM (based on Llama2) or a similarly sized `Llama2-Chat-7B`.

Emotional Understanding Ability	CE		PBE		PT		EC		Overall	
LLM	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH
Yi-Chat-6B (Base)	16.33	20.41	12.95	20.54	7.84	13.43	17.86	24.11	12.75	18.62
Yi-Chat-6B (CogCoT)	19.90	23.98	15.18	14.73	10.82	18.28	28.87	20.54	16.75	19.00
ChatGLM3-6B (Base)	24.49	30.61	19.64	14.73	13.43	11.19	30.36	37.50	20.25	20.62
ChatGLM3-6B (CogCoT)	29.08	29.59	21.43	15.18	17.54	15.67	39.29	48.21	24.50	23.50
Llama2-Chat-7B (Base)	13.27	13.27	9.37	9.37	13.06	4.85	10.71	5.36	11.75	8.25
Llama2-Chat-7B (CogCoT)	18.88	13.78	13.84	6.25	8.58	10.07	14.29	17.86	13.88	11.00
Baichuan2-Chat-7B (Base)	30.10	25.00	20.98	12.50	16.04	13.06	26.79	36.61	22.38	19.12
Baichuan2-Chat-7B (CogCoT)	17.86	28.06	13.39	12.95	12.69	17.54	16.07	29.46	14.62	20.50
Qwen-Chat-7B (Base)	28.06	26.02	21.88	16.96	16.42	15.30	28.57	31.25	22.50	20.62
Qwen-Chat-7B (CogCoT)	30.61	33.67	22.77	13.39	14.18	16.79	35.71	37.50	23.62	22.88
Llama2-Chat-13B (Base)	22.45	11.22	17.41	9.38	12.69	4.85	19.64	8.04	17.38	8.12
Llama2-Chat-13B (CogCoT)	17.35	13.78	16.52	4.46	10.82	7.46	25.00	11.61	16.00	8.75
Baichuan2-Chat-13B (Base)	34.69	37.24	24.55	19.64	18.66	20.15	33.04	37.50	26.25	26.62
Baichuan2-Chat-13B (CogCoT)	29.08	35.20	20.98	23.66	13.81	20.52	34.82	33.04	22.50	26.75
Qwen-Chat-14B (Base)	46.94	43.37	35.27	30.36	26.12	19.40	38.39	41.96	35.50	31.50
Qwen-Chat-14B (CogCoT)	48.98	44.90	32.59	25.89	27.61	25.00	43.75	40.18	36.50	32.25
Baichuan2-53B (Base)	43.88	46.43	31.25	25.00	25.37	25.37	49.11	50.89	34.88	34.00
Baichuan2-53B (CogCoT)	67.35	58.67	37.95	38.43	34.33	38.43	64.29	59.82	47.62	45.25
GPT 3.5 (Base)	41.84	30.61	33.48	18.30	21.64	22.01	44.64	45.54	33.12	26.38
GPT 3.5 (CogCoT)	46.43	36.22	29.02	20.98	24.63	26.49	46.43	38.39	34.25	29.00
Deepseek-v3 (Base)	79.59	70.92	50.89	40.18	44.40	45.52	63.39	70.54	57.50	53.75
Deepseek-v3 (CogCoT)	72.96	71.94	50.89	40.62	49.63	44.78	70.54	73.21	58.62	54.25

Table 1: Performance of various LLMs on emotional understanding ability in English and Chinese. CE, PBE, PT, and EC indicate Complex Emotions, Personal Beliefs and Experience, Perspective Taking, and Emotional Cues, respectively. Highlighted cells represent the best performance across three different model sizes.

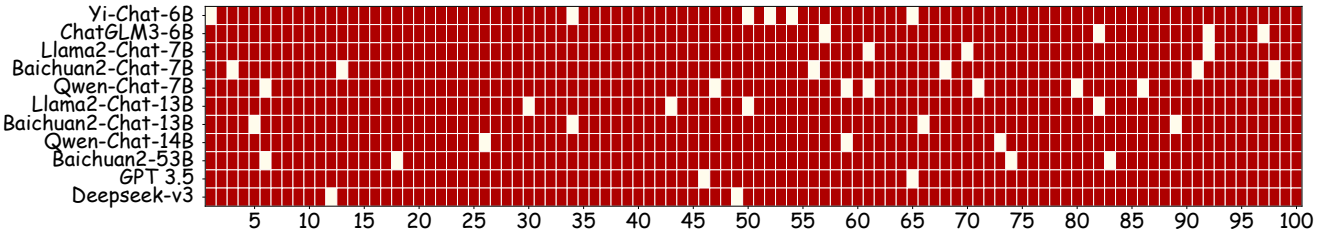

Figure 3: Consensus success rates on 100 sampled tasks (50 in English and 50 in Chinese, balanced across EU and EA). A cell is highlighted in red only when all adjacent agents achieve consensus during CogCoT generation.

All experiments followed the setup detailed in Wu et al. (2024) to ensure a fair comparison.

Evaluation Metrics. Following Sabour et al. (2024), we prompt each LLM five times per multiple-choice question and determine the final answer via majority voting. To mitigate choice-order bias (Zheng et al. 2024), answer choices are randomly permuted four times. Accuracy, averaged over all permutations, serves as the primary evaluation metric. Model outputs are parsed using heuristic rules.

For ESC, we report automatic evaluation metrics including perplexity (PPL), BLEU-2 (Papineni et al. 2002), ROUGE-L (Lin 2004), and the BOW Embedding-based (Liu et al. 2016) Extrema matching score. Additionally, we conduct human evaluation on 30 sampled validation dialogues. Two expert native-speaking annotators rate each response on Fluency, Identification, Comforting, and Suggestion using 3-point Likert scales. Inter-annotator agreement achieves Cohen's $\kappa = 0.73$, indicating substantial consistency.

Results and Discussion

As shown in Tables 1 and 2, `DeepSeek-V3` achieves the best performance across both EU and EA tasks. While most models perform slightly better in English, likely due to training data differences, overall trends remain consistent across languages, as illustrated in Figure 4. Larger models generally perform better, aligning with previous observations on LLM scaling laws (Brown et al. 2020).

EU proves substantially more challenging than EA, as it requires identifying both emotions and their underlying causes, and its scenarios are constructed to be abstract and nuanced to prevent reliance on superficial cues. In contrast, EA samples are more concrete and behavior-focused.

Emotional Application Ability	Personal-Self		Personal-Others		Social-Self		Social-Others		Overall	
LLM	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH
Yi-Chat-6B (Base)	50.50	49.50	40.00	39.50	50.50	37.50	48.00	35.50	47.25	40.50
Yi-Chat-6B (CogCoT)	59.00	53.00	41.00	46.50	47.00	49.50	54.50	41.50	50.38	47.46
ChatGLM3-6B (Base)	62.00	48.00	55.00	47.50	51.50	47.00	54.00	44.50	55.62	46.75
ChatGLM3-6B (CogCoT)	61.00	60.50	55.00	56.00	54.00	47.50	53.50	55.50	55.88	54.88
Llama2-Chat-7B (Base)	58.50	44.50	55.50	36.00	45.00	34.00	41.50	42.50	50.12	39.25
Llama2-Chat-7B (CogCot)	60.00	44.50	54.00	48.50	55.00	48.50	55.50	37.50	56.12	44.75
Baichuan2-Chat-7B (Base)	59.50	48.50	52.00	38.00	48.50	47.50	50.00	44.00	52.50	44.50
Baichuan2-Chat-7B (CogCoT)	65.50	53.00	59.00	45.50	54.50	57.00	52.50	47.00	57.88	50.62
Qwen-Chat-7B (Base)	62.50	44.00	50.50	49.00	55.50	51.50	50.00	42.00	54.62	46.62
Qwen-Chat-7B (CogCoT)	61.00	52.00	53.50	54.50	58.50	56.50	55.00	56.50	57.00	54.88
Llama2-Chat-13B (Base)	59.50	43.50	54.00	37.50	48.50	44.00	46.50	36.00	52.12	40.25
Llama2-Chat-13B (CogCoT)	55.50	46.50	46.00	38.00	54.00	46.50	51.00	40.00	51.62	42.75
Baichuan2-Chat-13B (Base)	52.00	51.50	52.00	51.50	52.00	58.00	58.50	58.00	53.62	54.75
Baichuan2-Chat-13B (CogCoT)	49.50	55.00	56.50	54.50	56.50	57.50	57.00	51.50	54.88	54.62
Qwen-Chat-14B (Base)	74.00	69.00	54.00	56.50	60.50	56.50	53.50	50.50	60.50	58.12
Qwen-Chat-14B (CogCoT)	68.00	70.50	54.50	53.00	60.00	62.50	52.50	60.00	58.75	61.50
Baichuan2-53B (Base)	75.50	63.00	62.50	59.00	62.00	68.00	61.50	58.00	65.38	62.00
Baichuan2-53B (CogCoT)	83.50	79.00	73.50	71.50	72.00	71.00	68.00	69.00	74.25	72.62
GPT 3.5 (Base)	64.50	57.00	61.00	57.00	60.50	53.00	59.50	56.00	61.38	55.75
GPT 3.5 (CogCoT)	71.00	70.50	60.50	63.50	62.00	60.00	63.00	61.50	64.12	63.88
Deepseek-v3 (Base)	80.00	75.50	72.50	61.50	73.00	73.50	76.00	74.50	75.38	71.25
Deepseek-v3 (CogCoT)	83.00	75.50	74.50	66.00	73.50	71.50	76.50	75.00	76.88	72.00

Table 2: Accuracy of various LLMs on emotional application ability in English and Chinese.

Among EU scenarios, perspective-taking is the most difficult, consistent with prior findings on ToM limitations in LLMs (Ullman 2023). CogCoT significantly improves performance on PT and other difficult EU categories. For EA, performance varies by relationship type and problem category, with "Personal-Others" scenarios being the most difficult. Again, CogCoT offers consistent gains.

Figure 4 shows that the commonly used step-by-step reasoning prompt provides limited benefits and can degrade performance. In contrast, CogCoT consistently enhances performance, especially for larger models. Table 3 compares CogCoT with COLM, a supervised model trained to generate CogCoT components. Despite its explicit supervision, COLM performs worse across both tasks, highlighting the superior generalization ability of our zero-shot approach.

Ablation Study

In this section, we conduct ablation experiments to evaluate the contributions of key components in the proposed MACRo framework. We select three representative models with the best performance across different model scales: `ChatGLM-6B`, `Qwen-Chat-14B`, and `DeepSeek-V3`. These models are tested on both EU and EA tasks. As shown in Table 4, each module yields significant performance gains. This is further supported by the statistical analysis, where the p-value $\ll 0.05$ for the paired t-test.

Model Variant	Overall (EU)		Overall (EA)	
	EN	ZH	EN	ZH
Llama2-Chat-7B (Base)	11.75	8.25	50.12	39.25
w/ Llama2-Chat-7B (COLM)	12.38	8.00	50.25	31.75
w/ Llama2-Chat-7B (CogCoT)	13.88	11.00	56.12	44.75

Table 3: Accuracy comparison between LLaMA2-Chat-7B enhanced with COLM and with CogCoT.

Analysis of Multi-agent Architecture. Removing the multi-agent architecture causes the largest performance drop, with average EU declines of 3.1% in English and 2.1% in Chinese, and EA decreases of 3.0% and 2.1%, respectively. The smaller drop in Chinese may stem from linguistic and cultural factors influencing reasoning complexity. These findings underscore the importance of role specialization and modular design. Without them, reasoning chains become less structured and more error-prone, whereas the agent-based approach decomposes tasks into interpretable subtasks, enhancing reliability throughout the CogCoT.

Analysis of Consensus Game. Removing the consensus game also results in significant performance drops. On aver-

Model Variant	Overall (EU)		Overall (EA)	
	EN	ZH	EN	ZH
ChatGLM3-6B (CogCoT)	24.50	23.50	55.88	54.88
w/o Consensus Game	22.75	22.25	53.38	54.00
w/o Multi-agent	20.38	20.62	51.50	52.50
Qwen-Chat-14B (CogCoT)	36.50	32.25	58.75	61.50
w/o Consensus Game	35.62	31.75	56.50	59.38
w/o Multi-agent	34.00	30.50	55.50	58.88
Deepseek-v3 (CogCoT)	58.62	54.25	76.88	72.00
w/o Consensus Game	57.75	53.38	76.00	71.38
w/o Multi-agent	55.88	52.50	75.38	70.75

Table 4: Ablation study of MACRo. The "w/o Multi-agent" variant removes role specialization, thereby rendering the consensus game mechanism inapplicable.

Metric	Vanilla	+ COLM	+ CogCoT
PPL ↓	16.96	15.98	**15.80**
BLEU-2 ↑	6.93	6.42	**7.01**
ROUGE-L ↑	15.01	15.57	**16.21**
Extrema ↑	50.28	50.49	**50.68**
Fluency ↑	2.30	2.50	**2.70**
Identification↑	1.95	2.10	**2.30**
Comforting ↑	2.10	2.35	**2.55**
Suggestion ↑	1.35	1.75	**2.10**

Table 5: Automatic and human evaluation on ESC.

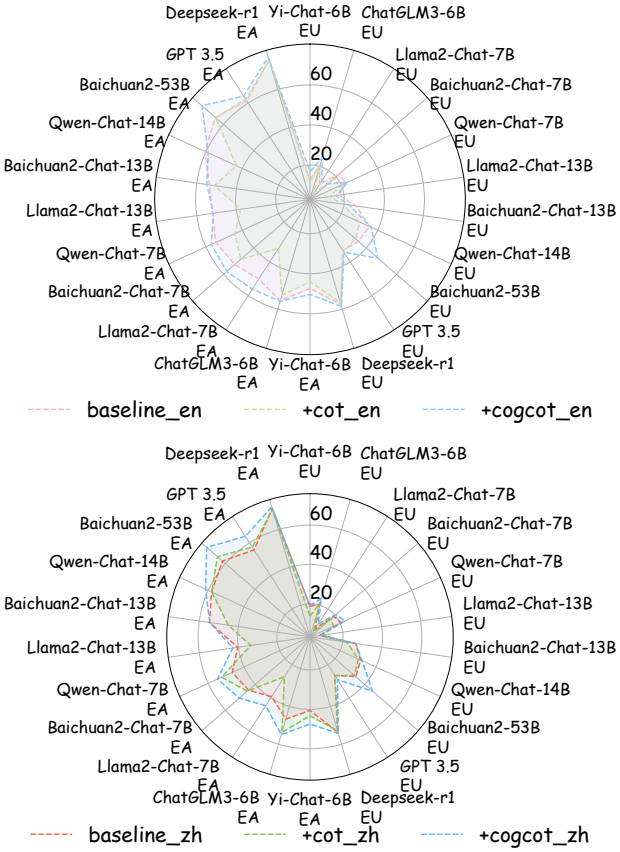

Figure 4: Overall accuracy comparison across baseline LLMs, CoT-enhanced models (step-by-step prompting), and CogCoT-enhanced models. Top: English; Bottom: Chinese.

age, the EU decreases by about 1.2% in English and 0.9% in Chinese, while EA drops by approximately 1.9% and 1.2%, respectively. This indicates that the consensus game further refines the outputs by fostering inter-agent alignment and coherence in the reasoning chain. As illustrated in Figure 3, the consensus game effectively promotes agreement among agents, enhancing the coherence of the CogCoT. Notably, larger LLMs exhibit higher success rates, likely due to their stronger reasoning abilities and more stable generation.

Error Analysis

We analyze cases where CogCoT leads to incorrect predictions compared to the base DeepSeek-V3 model. Most errors arise in the Chinese setting, particularly within perspective-taking and social-self tasks. These failures may stem from high linguistic and cultural complexity that challenges nuanced emotional reasoning. While the structured CogCoT enhances reasoning coherence and interpretability, it can sometimes over-constrain inference, limiting flexibility to capture implicit social cues, unlike the base model's looser but sometimes more adaptable reasoning.

Empowering Social Applications

We evaluate the generalizability of our MACRo on the ESC task (Liu et al. 2021), which requires generating empathetic and context-aware responses for users in distress. As a socially grounded dialogue task, ESC demands ToM capabilities, including the inference of latent emotions, user intentions, and dynamic conversational cues. We compare three model variants: the base BlenderBot, BlenderBot augmented with COLM-generated recognitive claim, and BlenderBot with CogCoT. As Table 3 shows, CogCoT achieves notable improvements across both automatic metrics and human evaluations. These results demonstrate that CogCoT significantly enhances the quality and emotional effectiveness of responses in real-world social AI applications.

Conclusion

In this paper, we present MACRo, a novel Multi-Agent Cognitive Reasoning framework designed to enhance the EI of LLMs through structured CogCoT. By decomposing recognitive reasoning into modular, specialized components coordinated via a game-theoretic consensus mechanism, MACRo effectively addresses coherence challenges and mitigates hallucination in reasoning processes. Extensive experiments on the EmoBench demonstrate that MACRo significantly improves both EU and EA across diverse LLMs in English and Chinese. Furthermore, evaluations on the Emotional support conversation task validate the framework's strong generalizability in affective and social reasoning.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China 62576120, the Major Key Project of PCL2025A09 and Key Laboratory of Computing Power Network and Information Security, Ministry of Education under Grant No.2024ZD020.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bar-On, R. 1997. *BarOn Emotional Quotient Inventory: A measure of emotional intelligence*. Multi-health systems.

Brockman, G.; Sutskever, I.; Team, O. A.; et al. 2015. Introducing OpenAI. *Zugang: https://blog. openai. com/introducing-openai/(15.12. 2019) and the future Digital Single Market. Zugang: https://edps. europa. eu/sites/edp/files/publication/18-04-24_giovanni_buttarelli_keynote_speech_telecoms_forum_en. pdf (15.12. 2019)*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Fudenberg, D.; and Levine, D. K. 1995. Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 19(5-7): 1065–1089.

Goleman, D. 1996. Emotional intelligence. Why it can matter more than IQ. *Learning*, 24(6): 49–50.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Huang, J.-t.; Lam, M. H.; Li, E. J.; Ren, S.; Wang, W.; Jiao, W.; Tu, Z.; and Lyu, M. R. 2023. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. *arXiv preprint arXiv:2308.03656*.

Hwang, J. D.; Bhagavatula, C.; Le Bras, R.; Da, J.; Sakaguchi, K.; Bosselut, A.; and Choi, Y. 2021. On symbolic and neural commonsense knowledge graphs.

Jacob, A. P.; Shen, Y.; Farina, G.; and Andreas, J. 2024. The Consensus Game: Language Model Generation via Equilibrium Search. In *The Twelfth International Conference on Learning Representations*.

Jacob, A. P.; Wu, D. J.; Farina, G.; Lerer, A.; Hu, H.; Bakhtin, A.; Andreas, J.; and Brown, N. 2022. Modeling strong and human-like gameplay with KL-regularized search. In *International Conference on Machine Learning*, 9695–9728. PMLR.

Kim, H.; Hessel, J.; Jiang, L.; West, P.; Lu, X.; Yu, Y.; Zhou, P.; Bras, R.; Alikhani, M.; Kim, G.; et al. 2023. SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12930–12949.

Kosinski, M. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45): e2405460121.

Lei, S.; Dong, G.; Wang, X.; Wang, K.; and Wang, S. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *CoRR*.

Lewis, D. 2008. *Convention: A philosophical study*. John Wiley & Sons.

Li, Z.; Chen, G.; Shao, R.; Xie, Y.; Jiang, D.; and Nie, L. 2024. Enhancing emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, C.-W.; Lowe, R.; Serban, I. V.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2122–2132.

Liu, H.; Wei, R.; Tu, G.; Lin, J.; Jiang, D.; and Cambria, E. 2025. Knowing What and Why: Causal emotion entailment for emotion recognition in conversations. *Expert Systems With Applications*, 274: 126924.

Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; and Huang, M. 2021. Towards Emotional Support Dialog Systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3469–3483.

Liu, Z.; Yang, K.; Xie, Q.; Zhang, T.; and Ananiadou, S. 2024. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5487–5496.

Monderer, D.; and Shapley, L. S. 1996. Potential games. *Games and economic behavior*, 14(1): 124–143.

Paech, S. J. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E. M.; Boureau, Y.-L.; et al. 2021. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 300–325.

Sabour, S.; Liu, S.; Zhang, Z.; Liu, J.; Zhou, J.; Sunaryo, A.; Lee, T.; Mihalcea, R.; and Huang, M. 2024. EmoBench: Evaluating the Emotional Intelligence of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5986–6004.

Salovey, P.; and Mayer, J. D. 1990. Emotional intelligence. *Imagination, cognition and personality*, 9(3): 185–211.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3027–3035.

Schuller, D.; and Schuller, B. W. 2018. The age of artificial emotional intelligence. *Computer*, 51(9): 38–46.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Strachan, J. W.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; Manzi, G.; et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7): 1285–1295.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tu, G.; Wu, T.; Luo, X.; Zeng, X.; Li, W.; and Xu, R. 2025. Meta-Learning for Incomplete Multimodal Sentiment Analysis. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2911–2915.

Tu, G.; Xiong, F.; Liang, B.; and Xu, R. 2024. A persona-infused cross-task graph network for multimodal emotion recognition with emotion shift detection in conversations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2266–2270.

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

Wang, X.; Li, X.; Yin, Z.; Wu, Y.; and Liu, J. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17: 18344909231213958.

West, P.; Bhagavatula, C.; Hessel, J.; Hwang, J.; Jiang, L.; Le Bras, R.; Lu, X.; Welleck, S.; and Choi, Y. 2022. Symbolic Knowledge Distillation: from General Language Models to Commonsense Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4602–4625.

West, P.; Le Bras, R.; Sorensen, T.; Lin, B. Y.; Jiang, L.; Lu, X.; Chandu, K.; Hessel, J.; Baheti, A.; Bhagavatula, C.; et al. 2023. NovaCOMET: Open Commonsense Foundation Models with Symbolic Knowledge Distillation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Wu, J.; Chen, Z.; Deng, J.; Sabour, S.; Meng, H.; and Huang, M. 2024. COKE: A Cognitive Knowledge Graph for Machine Theory of Mind. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15984–16007.

Yang, A.; Xiao, B.; Wang, B.; Zhang, B.; Bian, C.; Yin, C.; Lv, C.; Pan, D.; Wang, D.; Yan, D.; et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Yi, X.; Zhou, Z.; Cao, C.; Niu, Q.; Liu, T.; and Han, B. 2025. From Debate to Equilibrium: Belief-Driven Multi-Agent LLM Reasoning via Bayesian Nash Equilibrium. In *Forty-second International Conference on Machine Learning*.

Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; Zhang, G.; Wang, G.; Li, H.; Zhu, J.; Chen, J.; et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Zhang, B.; Yang, H.; Zhou, T.; Ali Babar, M.; and Liu, X.-Y. 2023. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the fourth ACM international conference on AI in finance*, 349–356.

Zheng, C.; Zhou, H.; Meng, F.; Zhou, J.; and Huang, M. 2024. Large Language Models Are Not Robust Multiple Choice Selectors. In *The Twelfth International Conference on Learning Representations*.

Zhong, W.; Cui, R.; Guo, Y.; Liang, Y.; Lu, S.; Wang, Y.; Saied, A.; Chen, W.; and Duan, N. 2024. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. In *NAACL-HLT (Findings)*.

Zhou, A.; Wang, K.; Lu, Z.; Shi, W.; Luo, S.; Qin, Z.; Lu, S.; Jia, A.; Song, L.; Zhan, M.; et al. 2024a. Solving Challenging Math Word Problems Using GPT-4 Code Interpreter with Code-based Self-Verification. In *12th International Conference on Learning Representations (ICLR 2024)*. International Conference on Learning Representations, ICLR.

Zhou, X.; Zhu, H.; Mathur, L.; Zhang, R.; Yu, H.; Qi, Z.; Morency, L.-P.; Bisk, Y.; Fried, D.; Neubig, G.; et al. 2024b. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. In *The Twelfth International Conference on Learning Representations*.