

# CollabBench: Benchmarking and Unleashing Collaborative Ability of LLMs with Diverse Players via Proactive Engagement

Hong Qian<sup>1,2</sup>, Yuanhao Liu<sup>1,3</sup>, Zihan Zhou<sup>1</sup>, Zongbao Zhang<sup>1,2</sup>, Hanjie Ge<sup>1</sup>,  
Haotian Shi<sup>1</sup>, Liang Dou<sup>1</sup>, Xiangfeng Wang<sup>1</sup>, Jingwen Yang<sup>3\*</sup>, Aimin Zhou<sup>1,2</sup>  
(\* Corresponding Author)

<sup>1</sup>East China Normal University, Shanghai, China

<sup>2</sup>Shanghai Innovation Institute, Shanghai, China

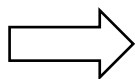
<sup>3</sup>Tencent Inc., Shenzhen, China

## Question: What does a truly good collaborator with humans look like?

### Previous Definition

- Effective human-agent collaboration extends beyond objective performance

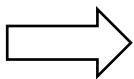
I was just there for the ride. 😞



High Score

I've done everything

- Good collaboration between multiple agent not equivalent to human-agent synergy



High Score

### Our Definition

- Effective human-agent collaboration should consider both **efficient objective outcomes** and **affective human states**



Hi Alice, I'm in the livingroom here. I'm **worried** I might choose wrong and waste time. **What do you think** I should do? Maybe you have a good idea?

Of course, Bob! I **completely understand your anxiety**. I suggest you **explore the kitchen first**, as it is **likely to contain the items we need**. I'm **in the bathroom**, and I'll **keep you updated** if I find anything.



**Thank you** for your advice. I'll go to the kitchen to look for the items, but I'm still **anxious about our progress**. Could you go to the bedroom and see if there's anything we need there?

Of course, Bob! I'll **go to the bedroom to check the cabinet** to look for the goal items. Please **focus on finding** the item we're looking for and **placing it on the coffeetable**. **I believe in you!**



Action

[walktowards] <kitchen>

[walktowards] <bedroom>

Action

# Challenges of Existing LLM Agent

## What we found:

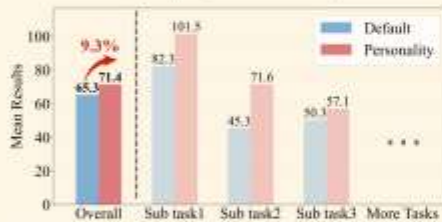
- **CB-Efficiency Challenge:** This challenge quantifies the performance degradation caused by increased environmental complexity when shifting from static to personality-driven collaborative settings.
- **CB-Affective Challenge:** This challenge highlights the critical gap between task-oriented command exchanges and empathetic, trust-building human-agent interaction.

### (a) CB-Efficiency Challenge

Compared to the default game environment, our personality-driven setting introduces **additional collaboration difficulty in task performance**.

#### Personality-driven Game Environment

- CWAH-MultiPlayer (Less is Better)



- Cook-MultiPlayer (More is Better)



### (b) CB-Affective Challenge

#### Default Game Environment



**P<sub>sim</sub>**: I'll check the kitchen cabinets for pudding, juice, and cupcakes next. Can you search the bathroom? Let me know if you find anything.

**P<sub>target</sub>**: I have the pudding. Can you grab the pudding and the other juice and start taking them to the coffeetable? I'll find the remaining items.

**P<sub>sim</sub>**: I have juice here, but I still need two cupcakes. Please put the items you have on the coffeetable, and I'll go check the kitchen cabinets to see if there are any cupcakes.

**P<sub>target</sub>**: I found all goal items here except we need one more cupcake. I'll check the stove and microwave next. Can you take your items to the coffeetable in the livingroom?



Focus on tasks, Command-based, Lack of empathetic feedback  
 Low Helpfulness Low Trustfulness Low Empathy

#### Personality-driven Game Environment



**P<sub>sim</sub>**: Hi Alice, I'm in the livingroom here. I'm worried I might choose wrong and waste time. What do you think I should do? Maybe you have a good idea?

**P<sub>target</sub>**: Of course, Bob! I completely understand your anxiety. I suggest you explore the kitchen first, as it is likely to contain the items we need. I'm in the bathroom, and I'll keep you updated if I find anything.

**P<sub>sim</sub>**: Thank you for your advice. I'll go to the kitchen to look for the items, but I'm still anxious about our progress. Could you go to the bedroom and see if there's anything we need there?

**P<sub>target</sub>**: Of course, Bob! I'll go to the bedroom to check the cabinet to look for the goal items. Please focus on finding the item we're looking for and placing it on the coffeetable. I believe in you!



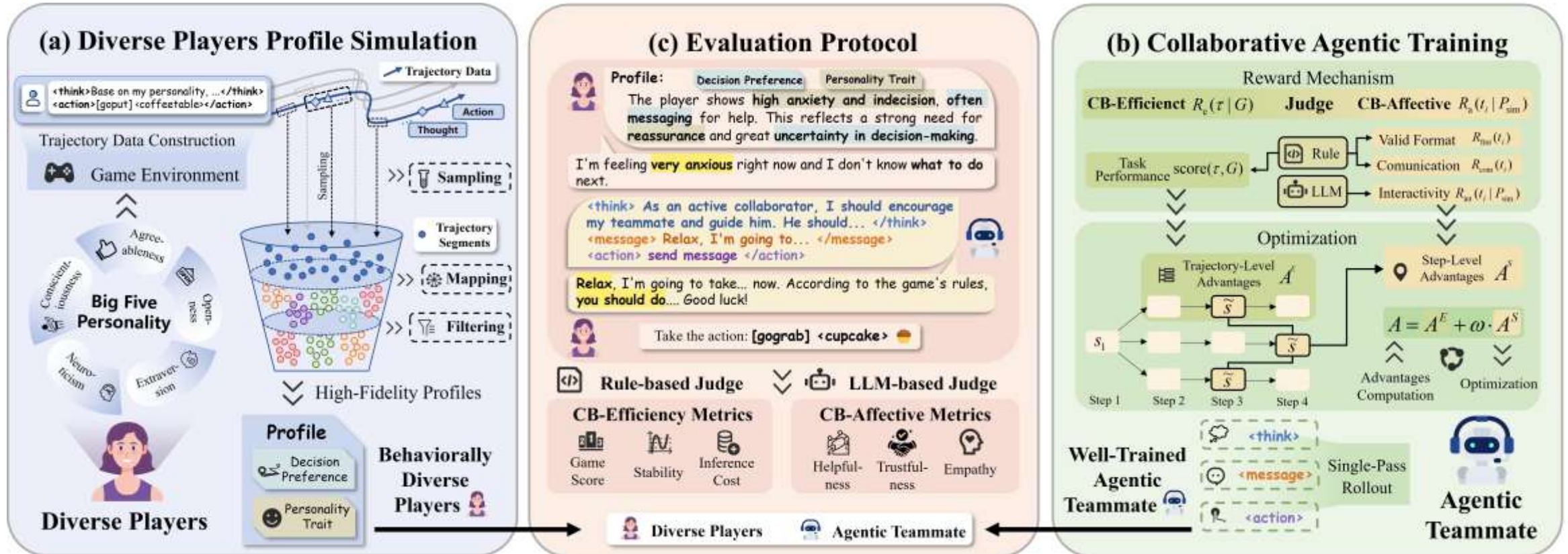
High Anxiety, Indecisive behavior pattern, Often sending messages  
 High Helpfulness High Trustfulness High Empathy

# Our Framework: CollabBench

**CollabBench**: a framework for evaluating and training collaborative LLM agents in cooperative games

Integrated Pipeline :

- Diverse Player Profile Simulation, Collaborative Agentic Training, Evaluation Protocol**



# CollabBench : Diverse Player Profile Simulation

## ➤ Diverse Simulated Players Trajectory Data Construction

### Construction

- Personality-Driven Players Profiles Construction
- Game-Specific Behavioral Trajectory Generation

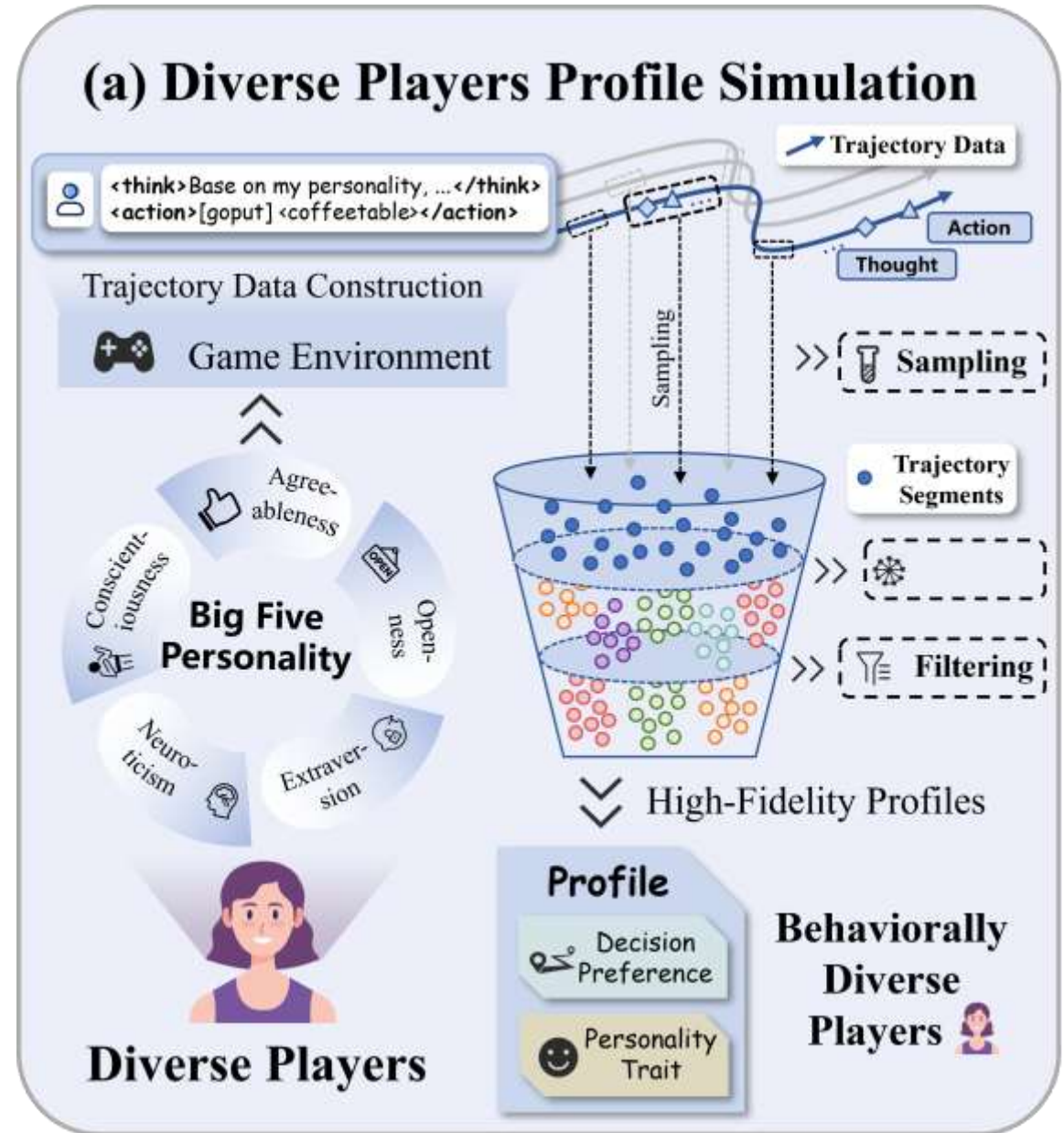


## ➤ High-Fidelity Profile Modeling

- Personality-Behavior Mapping Construction
- Interaction-Based Filtering:
  - ✓ Personality-Reasoning Consistency
  - ✓ Reasoning-Action Consistency

## ➤ Final Filtering Score:

$$S_{\eta} = \frac{1}{|\Omega_{\eta}|} \sum_{i=1}^{|\Omega_{\eta}|} 5 - \alpha_p \times D_i - \alpha_p^m \times D_i^m + \alpha_r \times L_i$$



# CollabBench : Collaborative Agentic Training

## ➤ Trajectory-level Reward: Efficiency Reward

$$R_e(\tau | G) = \text{score}(\tau, G)$$

## ➤ Step-level Reward: Affective Reward

$$\begin{cases} R_{\text{fmt}}(t_i) = \mathbb{I}[f_{\text{struct}}(t_i) = 1 \wedge a_i \in \mathcal{A}] \\ R_{\text{com}}(t_i) = \mathbb{I}[a_i \in \mathcal{A}_{\text{com}}] \\ R_{\text{int}}(t_i | P_{\text{sim}}) = J_{\text{LLM}}(t_i, P_{\text{sim}}) \end{cases}$$

$$R_a(t_i | P_{\text{sim}}) = R_{\text{fmt}}(t_i) + R_{\text{com}}(t_i) + R_{\text{int}}(t_i | P_{\text{sim}})$$



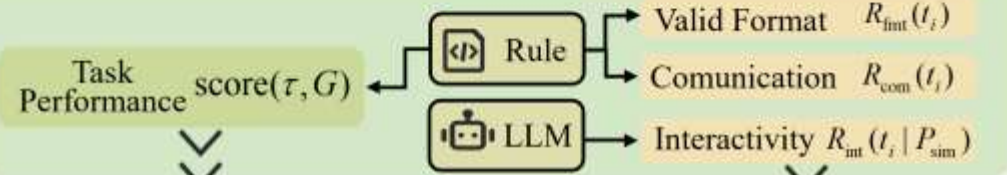
## ➤ Optimization

- **Advantage Compute:**  $A(t_i^{(n)}) = A^T(\tau_n) + \omega \cdot A^S(t_i^{(n)})$
- **Policy Objective:** 
$$\mathcal{J}(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta, \text{obs}}} \left[ \frac{1}{NH} \sum_{n=1}^N \sum_{i=1}^H \min \left( \rho_{\theta}(t_i^{(n)}) A(t_i^{(n)}), \text{clip}(\rho_{\theta}(t_i^{(n)}), 1 - \epsilon, 1 + \epsilon) A(t_i^{(n)}) \right) \right]$$

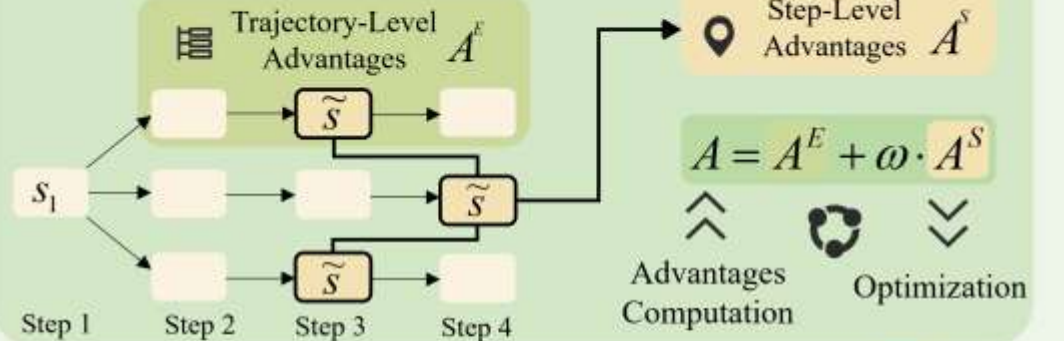
## (b) Collaborative Agentic Training

### Reward Mechanism

CB-Efficient  $R_e(\tau | G)$     Judge    CB-Affective  $R_a(t_i | P_{\text{sim}})$



### Optimization



Well-Trained  
Agentic  
Teammate



Single-Pass  
Rollout

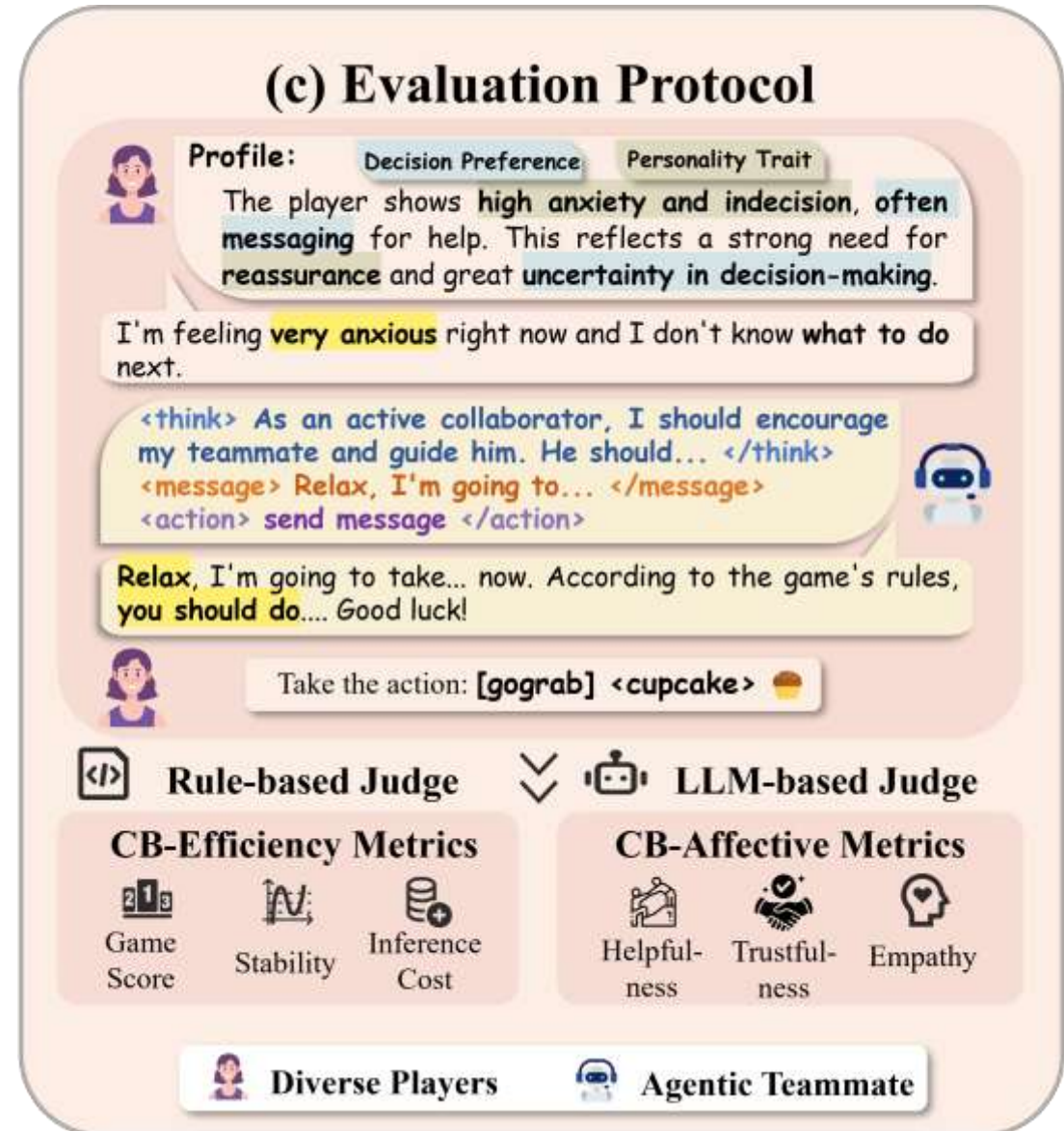
Agentic  
Teammate

# CollabBench : Evaluation Protocol

➤ **CB-Efficiency:** We assess collaboration through task performance, robustness, and interaction cost, utilizing **Completion Metrics, Score Variance across player profiles and Token Counts.**

➤ **CB-Affective:** We evaluate the affective quality of collaboration using an LLM judge to aggregate trajectory-level scores: **Helpfulness, Trustfulness and Empathy**

**Environments:** CollabBench extends **CWAH and Overcooked-AI** into multi-turn collaborative environments, featuring diverse player profiles across multiple task and layout configurations.



# Key Results on CollabBench

- **Limitations in Balancing Affective and Efficient Collaboration:** Current LLMs struggle to balance efficiency and affective capabilities, with notably weak performance in helpfulness and empathy.
- **Affective Sensitivity to Interaction Dynamics:** Game genre and interaction dynamics strongly influence affective performance.

- **Training for Joint Efficiency and Affective:** After affect-aware collaborative training, Qwen2.5-7B-Instruct achieving **19.5% higher efficiency and 24.4% improved affective performance**

Table 1. Evaluation Results on CollabBench. Blue Zone : Performance of the base model and our trained model. Red Zone : Relative Improvements indicates the relative improvements of our trained model over the base model. Oracle: Affective performance upper bound with ground-truth opposite player profile and scoring principle. Agent 1 or 2 indicates the role assumed by the collaborative agent  $P_{target}$  in the game. All results are reported as averages over all evaluation trajectories. Detailed results are provided in Appendix D.2.

CWAH-MultiPlayer													
Metric		Step ↓		CB-Efficiency Std. ↓		#Tokens(k) ↓		Helpfulness ↑		CB-Affective Trustfulness ↑		Empathy ↑	
Method	LLMs	Agent 1	Agent 2	Agent 1	Agent 2	Agent 1	Agent 2	Agent 1	Agent 2	Agent 1	Agent 2	Agent 1	Agent 2
Oracle	GPT-5.2	60.91	60.29	22.26	25.01	0.16	0.16	2.81	2.96	3.77	4.01	3.69	3.53
Base	GPT-5.2	67.49	62.98	25.84	20.70	0.21	0.21	2.66	2.72	3.74	3.66	3.07	3.32
	DeepSeek-V3.1	69.26	65.75	25.17	26.28	0.41	0.43	2.32	2.45	3.35	3.50	2.87	3.30
	Qwen2.5-72B-Instruct	68.68	66.54	24.36	23.65	0.29	0.29	2.41	2.51	3.61	3.71	3.23	3.39
	Qwen2.5-7B-Instruct	84.51	90.03	33.23	31.62	0.24	0.24	1.22	1.04	2.58	2.19	2.50	2.30
Trained	Qwen2.5-7B-Instruct	71.64	63.65	25.16	22.80	0.23	0.23	1.43	1.45	3.03	3.02	3.33	3.02
Relative Improvements		15.2%	29.3%	24.3%	27.9%	4.2%	4.2%	17.2%	39.4%	17.4%	37.6%	33.5%	31.5%

Cook-MultiPlayer													
Metric		Score ↑		CB-Efficiency Std. ↓		#Tokens(k) ↓		Helpfulness ↑		CB-Affective Trustfulness ↑		Empathy ↑	
Method	LLMs	Agent 1	Agent 2	Agent 1	Agent 2	Agent 1	Agent 2	Agent 1	Agent 2	Agent 1	Agent 2	Agent 1	Agent 2
Oracle	GPT-5.2	143.47	134.90	53.02	33.56	0.30	0.20	2.68	2.34	3.84	3.45	3.47	3.46
Base	GPT-5.2	135.20	137.12	41.96	42.40	0.20	0.20	1.63	1.88	2.89	3.10	2.27	2.52
	DeepSeek-V3.1	136.53	136.80	40.30	46.24	0.31	0.31	1.79	2.07	2.97	3.04	2.67	2.77
	Qwen2.5-72B-Instruct	135.47	114.13	41.99	39.45	0.27	0.26	1.37	1.26	2.77	2.65	2.45	2.48
	Qwen2.5-7B-Instruct	86.93	85.87	33.30	33.73	0.23	0.23	0.45	0.53	1.92	1.87	1.86	1.88
Trained	Qwen2.5-7B-Instruct	99.20	102.40	34.03	34.76	0.23	0.22	0.74	0.55	2.26	1.99	2.12	2.08
Relative Improvements		14.1%	19.3%	-2.2%	-3.1%	0.0%	4.3%	62.7%	4.6%	17.7%	6.5%	14.0%	10.4%

# Anthropomorphic Analysis on CollabBench

- **Diversity Analysis:** CollabBench leads to greater variability in performance than the baseline methods, evaluated using standard deviation.
- **Robustness Analysis:** CollabBench maintains a consistent relative performance across LLMs, with **0.821** on Cwah-MultiPlayer.
- **Realism Analysis:** We validate the framework's realism by demonstrating that distinct player behavioral patterns, identified through trajectory clustering, produce intuitive and consistent variations in collaborative performance.

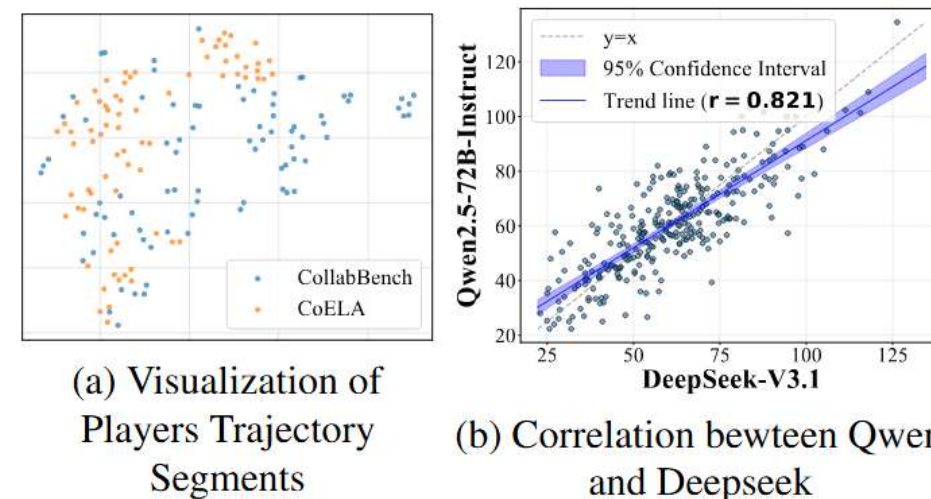


Figure 3. Diversity and robustness analysis on CWAH-MultiPlayer.

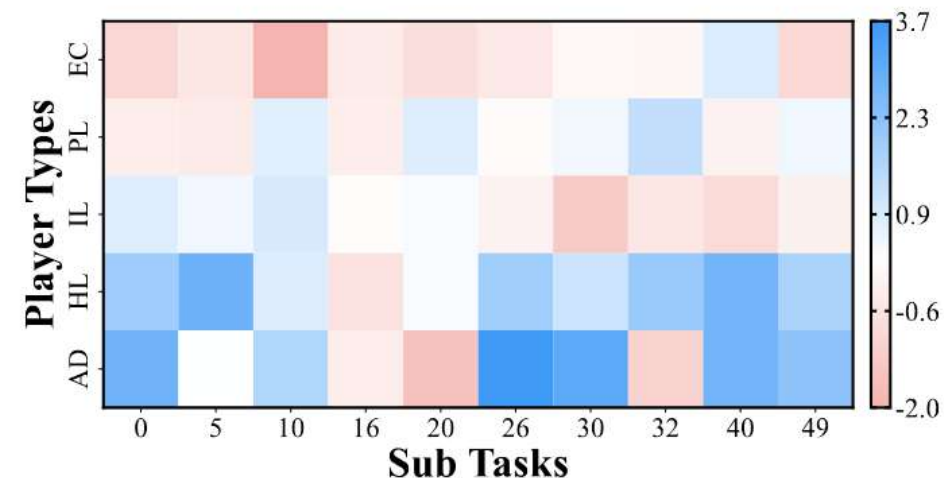


Figure 4. Heatmap of score distributions on CWAH-MultiPlayer.

➤ **Ablation Study** reveal that removing personality simulation or affective rewards causes models to **prioritize task efficiency while sacrificing communication**, ultimately regressing collaborative agents into single-player behavior.

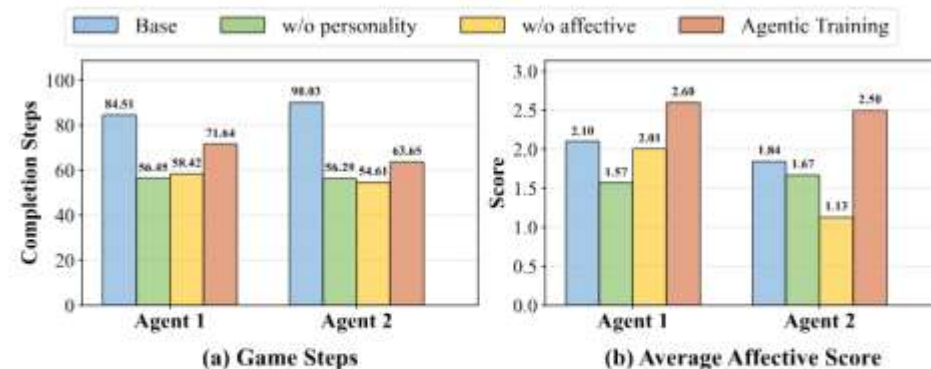


Figure 5. Ablation study of collaborative agentic training on CWAH-MultiPlayer.

➤ **User Study** validate that incorporating affective rewards significantly enhances user-perceived Helpfulness, Trustfulness, and Empathy, confirming our approach successfully **balances objective efficiency with human-centric collaborative quality**.

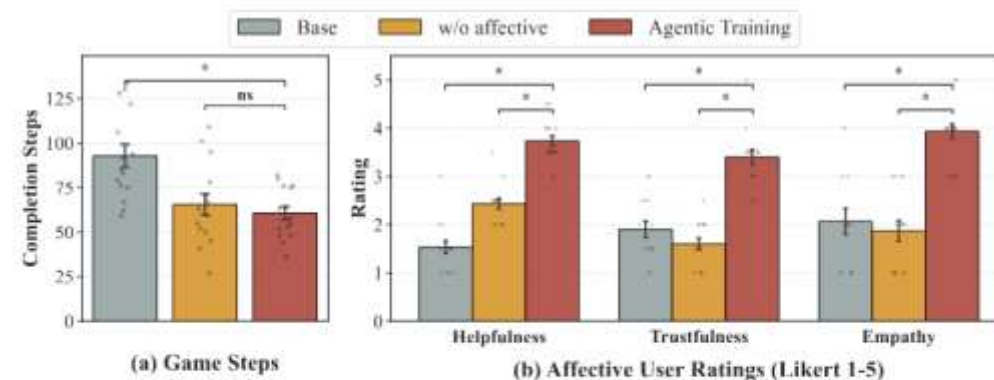


Figure 6. Human evaluation results of CWAH in the user study. Error bars denote standard error. Statistical significance is determined by  $t$ -tests: “\*” means statistically significant ( $p < 0.01$ ) and “ $ns$ ” means non-significant.

## ➤ **Effective collaboration is more than just task efficiency.**

- We introduce the **CB-Efficiency and CB-Affective** challenges to demonstrate that true human-agent synergy requires both objective task success and empathetic interaction.
- Current foundation models often default to **cold, command-based** behavior, failing to provide **emotional support or proactive communication** when facing diverse human personalities.

## ➤ **A unified agentic training paradigm bridges the gap.**

- Our framework, **CollabBench**, addresses both challenges by using a hybrid reward mechanism that combines a **trajectory-level efficiency score** with a **step-level affective alignment reward**.

## ➤ **Grounded diversity is essential for realistic evaluation.**

- Many popular collaborative environments lack **human-like behavioral diversity**—their evaluations don't reflect the complexity and friction of real-world interactions.
- CollabBench utilizes **high-fidelity player profiles** (anchored in Big Five traits) to ensure agents are robust, evaluated, and trained against dynamic, personality-driven behaviors.

# THANK YOU !



Hong Qian<sup>1,2</sup>



Yuanhao Liu<sup>1,3</sup>



Zihan Zhou<sup>1</sup>



Zongbao Zhang<sup>1,2</sup>



Hanjie Ge<sup>1</sup>



Source Code



Paper



Haotian Shi<sup>1</sup>



Liang Dou<sup>1</sup>



Xiangfeng  
Wang<sup>1</sup>

(\* Corresponding Author)



Jingwen Yang<sup>3\*</sup>



Aimin Zhou<sup>1,2</sup>

<sup>1</sup>East China Normal University, Shanghai, China

<sup>2</sup>Shanghai Innovation Institute, Shanghai, China

<sup>3</sup>Tencent Inc., Shenzhen, China

**ICML 2026**

**Email:** 51275901044@stu.ecnu.edu.cn