Identifying Collective Intelligence Factor in LLM Agent Groups for Generalizable Multi-Agent System Design

Zhilun Zhou[1], Zihan Liu[1], Jiahe Liu[1], Yihan Wang[1], Qingyu Shao[1], Fengli Xu[1]*, Depeng Jin[1], Yong Li[1]

[1]Department of Electronic Engineering, BNRist, Tsinghua University

Abstract

Large language model (LLM)-based multi-agent systems (MASs) have shown impressive performance in solving a wide range of complex problems. However, previous studies mainly focus on designing customized MAS for specific tasks, while a critical research problem remains unclear: Do LLM agent groups exhibit a form of "general intelligence" that reflects their general ability across various tasks? Researchers have found a Collective Intelligence (CI) factor in human groups that captures their general capability. Inspired by this, in this study, we aim to investigate whether an analogous CI factor also exists in LLM agent groups, which is crucial for building generalizable MAS. Motivated by human cognitive psychology experiments, we construct 108 LLM agent groups with diverse group sizes, LLM compositions, and communication topologies. We systematically evaluate these groups across a wide range of tasks and analyze their performances. Our results demonstrate that an Artificial Collective Intelligence (ACI) factor can be extracted from LLM agent groups to predict the generalization performance on new tasks. Inspired by this, we train a model to predict the ACI based on the features of MAS, and show that it can be used as a plug-in to enhance the generalization ability of MAS optimization methods. Our code is open-source at https://github.com/tsinghua-fib-lab/LLM_Collective_Intelligence.

1 Introduction

The rapid development of large language models (LLMs) has given rise to LLM-based multi-agent systems (MAS), which have shown remarkable capabilities in many domains. Prior studies reveal that different MAS may excel in different tasks (Zhang et al., 2025b), and thus researchers have proposed a variety of methods to design MAS optimized for specific applications, such as coding (Qian et al., 2024) and game playing (Chen et al., 2023). However, a fundamental question remains unclear: do LLM-based MAS exhibit a form of "general intelligence" that goes beyond task-specific performance and reflects a group's overall ability across diverse tasks?

In human cognitive psychology research, the quest for a "general intelligence" measure has a long history (Spearman, 1904), with the most popular test known as the "IQ test". This line of research seeks to derive a single statistical factor that measures the generalizable mental capabilities of individuals across various cognitive tasks. More recently, studies have shown that the cognitive performance of human groups can also be predicted to a large extent by a single statistical factor, which is referred to as the "collective intelligence" (CI) factor (Woolley et al., 2010; Riedl et al., 2021). This factor captures the task-independent capability of groups across a wide range of domains. Since LLMs have shown many human-like behaviors (Chen et al., 2025), a natural question is whether a similar CI factor also exists in LLM agent groups. If so, it not only indicates that LLM agent groups share similar general intelligence with human groups, but also would provide critical insights for designing more effective and generalizable LLM agent groups.

In this work, we conduct systematic experiments to investigate the existence and properties of the CI factor in LLM agent groups. We aim to answer three research questions: (1) Does a general CI factor exist in LLM agent groups? (2) What features of an LLM agent group affect its CI? (3) Can insights from CI be used to guide the design of LLM agent groups? To answer these questions, we construct 108 LLM agent groups spanning 8 different LLMs, while varying group size, communication topology, and model composition. These dimensions are

*Corresponding author: fenglixu@tsinghua.edu.cn

Findings of the Association for Computational Linguistics: ACL 2026, pages 12827–12842
July 2-7, 2026 ©2026 Association for Computational Linguistics

chosen based on human experiments (Riedl et al., 2021), which ensure the diversity and robustness of our experiments. We then evaluate the groups on a broad spectrum of cognitive tasks, including commonsense reasoning, mathematics, game playing, coding, and writing. Our findings can be summarized as follows. First, we demonstrate that a general CI factor, which we term **Artificial Collective Intelligence (ACI)**, can be extracted from LLM agent groups to capture their general capability across different tasks. Second, ACI in LLM agent groups shows similar patterns with CI in human groups, where the collaboration process is the most important determinant of ACI. Third, we find that the features of LLM agent groups can be used to predict the performance for new groups and on new tasks. Based on these findings, we further train an ACI prediction model and show that it can serve as a plug-in to enhance the generalization ability of MAS optimization algorithms. Experiments show that our method improves the performance of MAS optimization algorithms on unseen tasks by an average of 4.6%.

The main contributions of this work are threefold:

- We demonstrate the existence of a general ACI factor in LLM agent groups, which accounts for 66.3% of the variance in group performance and generalizes well across tasks.
- We analyze the indicators of LLM agent groups that affect the ACI and find similar patterns with human groups. Specifically, the collaboration process has the greatest impact on ACI, followed by individual intelligence, with group size having a relatively smaller effect. Moreover, we show that these indicators can be used to predict the performance of LLM agent groups.
- Inspired by these findings, we propose an ACI prediction model, which can be used as a plug-in to improve the generalization ability of existing MAS optimization algorithms.

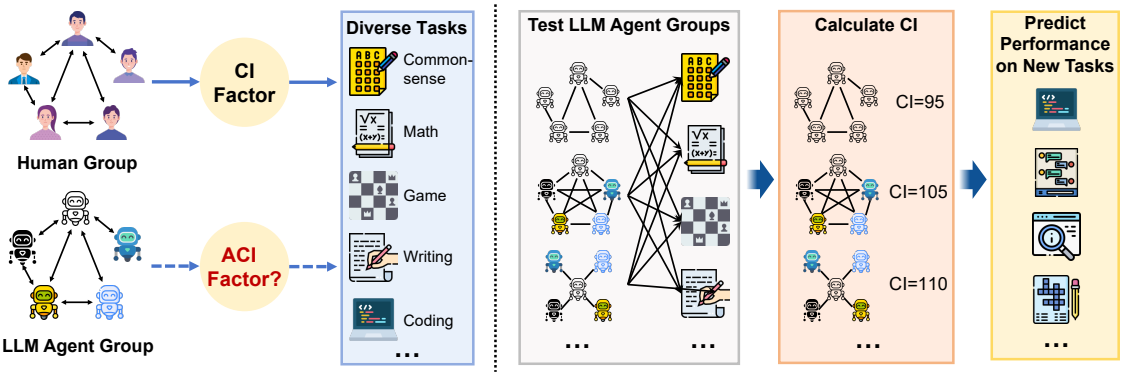
2 Related Work

2.1 Collective Intelligence of Human

Individual intelligence of humans is commonly conceptualized as a statistical factor, which predicts performance across various tasks (Spearman, 1904). Similarly, CI describes a group's ability to perform a range of tasks, also captured by a single statistical factor. Woolley et al. demonstrated the existence of CI factor in human groups, which accounts for over 40% of the variance in group performance (Woolley et al., 2010). They also found that CI is correlated not only with the individual intelligence of group members but also with their average social sensitivity and the proportion of females in the group. Riedl et al. conducted large-scale experiments and further verified the existence of CI (Riedl et al., 2021). They found that the group collaboration process is more important in predicting CI than individual intelligence. These studies on CI in human groups provide a valuable framework for investigating the CI in LLM-based multi-agent systems. LLMs have demonstrated many human-like behaviors, and it has been pointed out that individual LLMs show interrelated cognitive-like capabilities like humans (Ilić and Gignac, 2024). However, it remains unclear whether groups of LLM agents also have a general CI factor.

2.2 LLM Multi-agent Collaboration

In recent years, there have been extensive studies on multi-agent collaboration (Xiao et al., 2023; Li et al., 2023; Hong et al., 2024; Qian et al., 2024; Chen et al., 2023), which can be categorized into three types. The first line of studies aims to design multi-agent collaboration methods for specific tasks. These methods typically follow human collaboration mechanisms such as debate (Du et al., 2024) and standardized operating procedures (SOP) (Hong et al., 2024). Another line of studies further proposes to automatically design and optimize the collaboration strategy. For instance, Agentverse lets LLM generate and adjust the agent composition based on the status of the task (Chen et al., 2023). G-designer proposes to optimize the communication network of agents through a variational graph auto-encoder (Zhang et al., 2025b). GPTSwarm represents multi-agent systems as composite graphs and optimizes node-level prompts as well as edges between agents (Zhuge et al., 2024). Moreover, a third line of studies focuses on the underlying mechanism of multi-agent collaboration, such as the impact of agents' traits (Zhang et al., 2024) and hyperparameters (Smit et al., 2024), and the scaling law of multi-agent systems (Qian et al., 2025). However, existing studies mainly focus on task-specific scores and overlook the general ability of LLM agent groups across diverse tasks.



(a) (b)

Figure 1: (a) We aim to investigate whether LLM agent groups also exhibit an ACI factor similar to that observed in human groups. (b) The overall framework of our experiments.

3 Experiment Framework

In this study, we investigate the CI of LLM agent groups from the following aspects:

1. Does an ACI factor exist in LLM agent groups? We conduct factor analysis to extract the latent factor from the performance of different LLM agent groups across a wide range of tasks, which shows that there exists a factor accounting for 66.3% of the variance. (Section 4)

2. What features of an LLM agent group affect its ACI? We analyze the characteristics of LLM agent groups that affect their ACIs, and find that the collaboration process plays the most important role. (Section 5.1)

3. Can insights from ACI be used to guide the design of LLM agent groups? We demonstrate that the features of LLM agent groups can be used to predict ACI for unseen groups and on new tasks, which could help estimate the group performance without testing on specific tasks. We also show that the predicted ACI can be used to enhance the generalization ability of MAS optimization methods. (Section 5.2, 5.3 and 6)

We first introduce our experiment framework as follows.

3.1 Multi-agent Collaboration Framework

We leverage a widely used LLM multi-agent collaboration framework (Du et al., 2024; Wang et al., 2025; Yu et al., 2025). Specifically, the LLM agents can be modeled as a graph $\mathbf{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ is the set of nodes, each node is an LLM agent, and $\mathcal{E}$ is the set of edges. We also refer to the graph $\mathbf{G}$ as the *communication topology* of LLM agent groups. Given a query q, each agent $v_i \in \mathcal{V}$ independently generates an initial response $r_i^{(1)} = v_i(q)$. Then in round $t (t \geq 2)$, each agent observes the previous answers of neighboring agents, and updates its own answer:

$$r_i^{(t+1)} = v_i(\{r_j^{(t)} | j \in \mathcal{N}(v_i)\}), \quad (1)$$

where $\mathcal{N}(v_i)$ denotes the neighboring nodes of v_i. After T rounds, the final answer is obtained by aggregating the responses of all agents

$$r^{(T)} = Aggregate(r_1^{(T)}, r_2^{(T)}, \ldots, r_N^{(T)}). \quad (2)$$

3.2 Composition of LLM Agent groups

We choose 8 different LLMs from various providers and with different sizes to ensure diversity, including OpenAI (gpt-3.5-turbo-0125, gpt-4o-mini-2024-07-18), Qwen (Qwen2.5-7B-Instruct, Qwen2.5-32B-Instruct, Qwen2.5-72B-Instruct), GLM (glm-4-9b-chat), InternLM (internlm2_5-20b-chat), and Google (gemma-2-27b-it). Using these models, we construct LLM agent groups with varying group sizes, number of rounds, communication topologies, and LLM compositions. Specifically, the group sizes range from {3,5,8}, the number of rounds is set to {2, 3}, and the communication topologies include {decentralized network, centralized network, random network}. Additionally, each group is composed of either homogeneous (same LLM) or heterogeneous (different LLMs) agents, resulting in a total of 108 groups. Their details are shown in Appendix A.1.

3.3 Evaluation Tasks

We evaluate the performance of LLM agent groups on five tasks: commonsense reasoning (Wang et al., 2024), mathematics (Hendrycks et al.), games (Srivastava et al., 2023), coding (Chen et al., 2021), and writing (Madaan et al., 2024). The reasons for selecting these tasks are twofold. First, the task selection covers widely adopted benchmarks

in multi-agent system research (Zhuge et al., 2024; Zhang et al., 2025b; Zhou et al., 2025), providing a diverse and representative set of tasks that effectively assess the collective intelligence of LLM agent groups. Second, the tasks exhibit a moderate level of difficulty for our selected LLMs, with average single-model performance ranging from 0.40 to 0.81 across tasks, which avoids too trivial or hard settings and thus allows for a clear comparison of different groups. More implementation details are presented in Appendix A.1.

4 Identifying ACI Factor in LLM Agent Groups

4.1 Empirical Analysis

We first demonstrate that a general ACI factor exists in LLM agent groups. First, the performances of LLM agent groups across different tasks show a strong positive correlation, as shown in Figure 2. The average correlation coefficient is $r = 0.55$, notably higher than the $r = 0.28$ observed in human groups (Riedl et al., 2021). This strong cross-task correlation suggests the presence of a shared underlying capability—analogous to the general CI factor found in human groups—that influences group performance across different tasks.

To further examine this possibility,we perform exploratory factor analysis (EFA) to assess whether a single latent factor can account for performance variation across tasks. The analysis reveals a dominant factor that explains 66.3% of the total variance, substantially more than the 43% reported in human groups, while the second factor accounts for only 18.7%. We then conduct confirmatory factor analysis (CFA) by fitting a single-factor structural model. The resulting fit indices ($\chi^2 = 30.6$, $p < 0.001$, CFI = 0.967) indicate a good model fit, further supporting the presence of a general ACI factor. Taken together, these findings demonstrate that LLM agent groups, much like human groups, exhibit a form of collective intelligence that reflects a generalizable capability across tasks.

4.2 Measuring ACI Factor

Based on previous analysis, we define the ACI of LLM agent groups following the definition of CI in human groups (Woolley et al., 2010; Riedl et al., 2021). Specifically, we first standardize the performance scores on each dataset because the scales of scores may vary across datasets. Let s_{ij} be the standardized score of group j on dataset

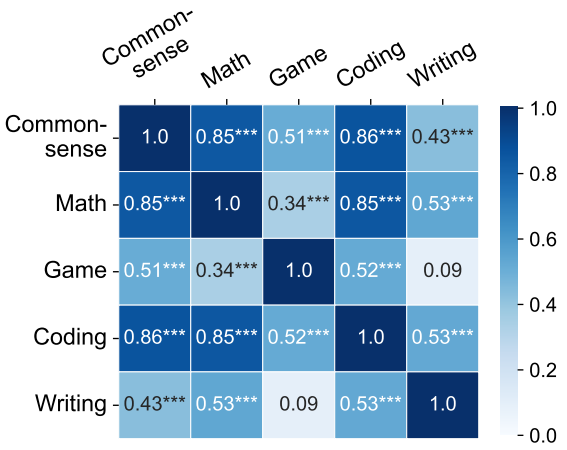


Figure 2: Correlations between tasks. ***$p < 0.001$,**$p < 0.01$,*$p < 0.05$.

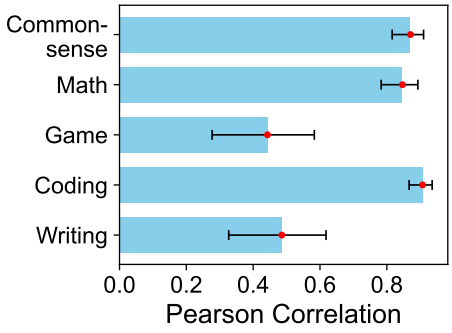


Figure 3: Correlation of leave-one-out ACI with criterion task.

i. Using the aforementioned factor analysis, we obtain a factor loading w_i for each dataset i (all $p < 0.001$), which reflects how strongly each observed variable (i.e., the performance on each dataset) is associated with the underlying ACI factor. Then the ACI factor of group j is computed as the weighted score across all datasets

$$\text{ACI}_{j,raw} = \sum_{i=1}^{5} w_i s_{ij} / \sum_{i=1}^{5} w_i. \quad (3)$$

Following conventions in intelligence testing, we standardize these raw ACI scores by scaling them to have a mean of 100 and a standard deviation of 15:

$$\text{ACI}_j = \frac{\text{ACI}_{j,raw} - \text{mean}(\text{ACI}_{raw})}{\text{std}(\text{ACI}_{raw})} \times 15 + 100. \quad (4)$$

The resulting ACI scores for all LLM agent groups are reported in Appendix A.1.

To verify the generalizability of the ACI factor, we perform leave-one-out experiments where we

use one of the five datasets as the held-out criterion task and compute the ACI factor using the remaining four datasets. We then assess how well these leave-one-out ACI scores predict group performance on the held-out task. As shown in Figure 3, the correlations exceed 0.8 on three of the tasks, and reach around 0.5 on the rest tasks, all statistically significant with $p < 0.001$. These results indicate that the ACI factor derived from any subset of four tasks generalizes well to unseen tasks, supporting its robustness as a measure of general group capability.

5 Examining ACI in LLM Agent Groups

5.1 Predicting ACI with Group Indicators

We have demonstrated that LLM agent groups have an ACI factor similar to human groups. An emerging question is what characteristics of a group affect its ACI? Existing studies have shown that the CI of a human group is affected by indicators like group size, individual intelligence, and collaboration process (Woolley et al., 2010; Riedl et al., 2021). Following these findings, we construct a set of indicators for LLM agent groups with three categories as follows.

- **Group Size**: These indicators measure the size of a group, including the number of agents in a group (N) and its square (N^2).

- **Individual Intelligence**: These indicators characterize the ability of agents in a group. It has been demonstrated that individual LLM exhibits a general intelligence factor (Ilić and Gignac, 2024). Here we adopt the same method as calculating ACI (Section 4.2) to obtain an individual intelligence score g for each LLM agent. We use the average g and maximum g of all agents in a group as indicators.

- **Collaboration Process**: These indicators describe how agents collaborate to solve the tasks (Hackman, 1978; Riedl et al., 2021). (1)*Variance of degree* is calculated as the variance of degrees of each node. It corresponds to the inequality of speaking turns in human groups, which has been demonstrated to be negatively correlated with CI (Woolley et al., 2010). (2)*Effort* is calculated as the total amount of activity that all agents perform during the task completion process. In our collaboration process, the activity refers to the communication between agents. Therefore, we define *Effort* as the number of rounds times the number of edges in the graph, i.e., *Effort*$= T \times |E|$ (3) *Skill congruence* measures the extent to which agents contribute efforts in proportion to their ability. In other words, a group where agents with higher capabilities put in more effort would have a high congruence. We define this indicator as the Pearson correlation between agents' individual intelligence and their node degrees. Experiments in human groups show that skill congruence is a strong positive predictor of CI.

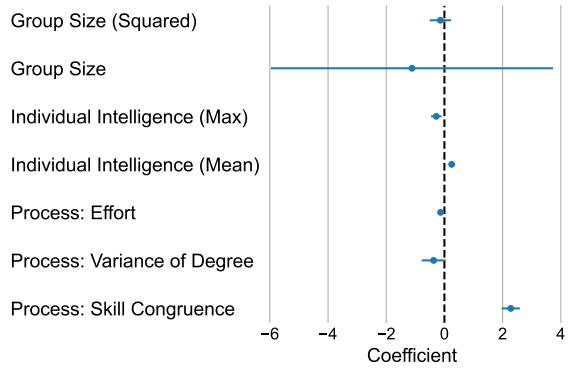


Figure 4: Regression coefficients of indicators predicting ACI.

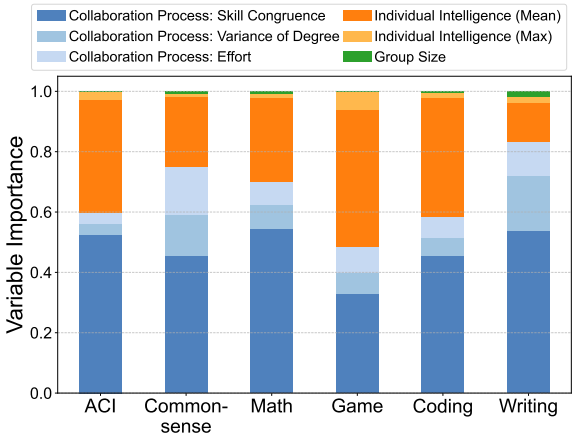


Figure 5: Importance of different indicators predicting ACI and task performances.

We first fit a linear model to predict the ACI with these indicators, and present the standardized regression coefficient in Figure 4. Consistent with human experiments, skill congruence and average individual intelligence are both significant positive predictors of ACI, while group size and effort are not strong predictors.

To further assess the relative importance of each indicator, we fit a random forest regression model,

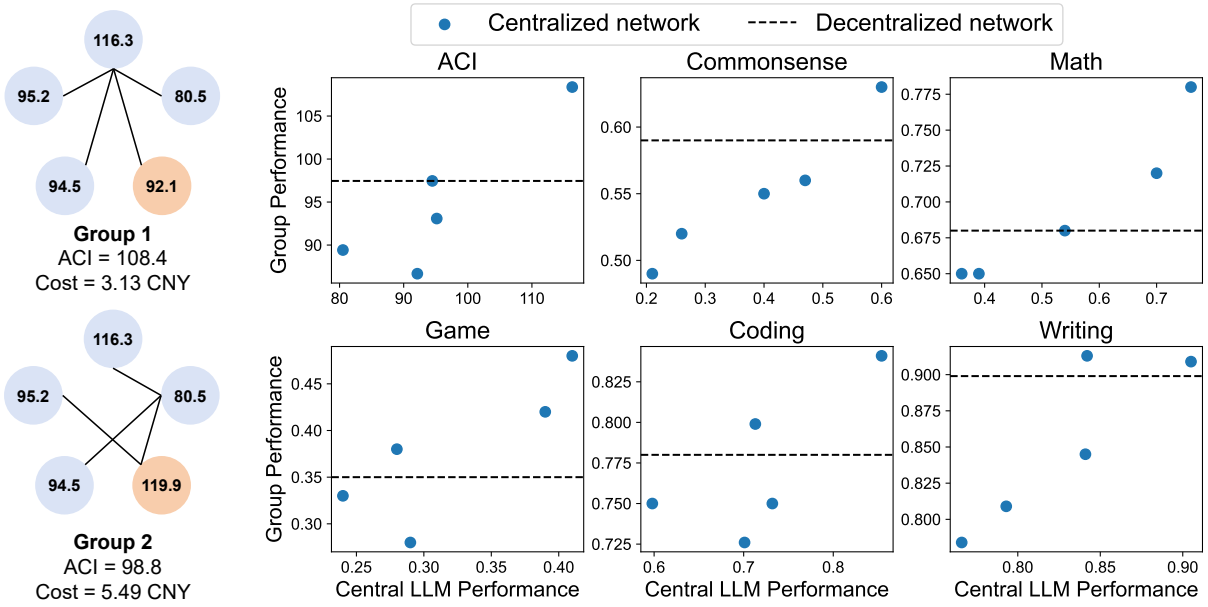


Figure 6: (a) Comparison of two LLM agent groups, where the second one has stronger LLMs and a higher cost but a lower ACI. The number in each circle represents the individual intelligence score g of LLM. The cost is the total API cost for five tasks. (b) Comparison of decentralized networks and centralized networks with different LLMs serves as the central node. The dashed line shows the ACI/performance of the decentralized network. The blue dots show the relationship between the ACI/performance of the group and the ACI/performance of the central agent in centralized networks.

which can capture nonlinear and more complex relationships between the indicators and ACI, and calculate the importance of each variable. As shown in Figure 5, the collaboration process plays the most significant role in predicting ACI, even more important than individual intelligence. We also fit a model to predict group performance on each of the datasets, yielding similar results. This finding aligns with prior research on human groups (Riedl et al., 2021). Specifically, the skill congruence indicator accounts for more than 50% of the total importance, and the average individual intelligence accounts for 37%. In comparison, the maximum individual intelligence, group size, and effort account for less than 5%. Such results suggest that the way agents interact with each other has a greater impact than their individual abilities.

These findings provide several insights for MAS design. First, simply increasing the ability of individual agents, such as employing stronger LLMs, does not necessarily lead to better outcomes. We present a case in Figure 6(a), where the second group has a stronger LLM (Qwen2.5-72B, $g = 119.9$) than the first group (internlm2.5-20b, $g = 92.1$). Consequently, the second group also incurs a cost 75% higher than the first group. However, the ACI of the first group is 9.7% higher than the second one, highlighting the critical role of communication topology.

Second, compared with adding more communication links between agents, it would be better to let each agent do what matches their capabilities. In our collaboration framework, this means that stronger agents should be placed on nodes with higher degrees. We further verify this by comparing the performance of decentralized networks with centralized networks. Specifically, we construct six groups with five different LLMs , including a decentralized network where all agents are fully connected, and five centralized networks where each agent is selected in turn to serve as the central node. The ACI and performance of these groups on all datasets are presented in Figure 6(b). It can be observed that in centralized networks, the task performances and ACI are positively correlated with those of the central agents, which is consistent with previous findings. Moreover, when the strongest LLM serves as the central node, the group performance not only achieves the best among centralized groups in most cases but also surpasses the decen-

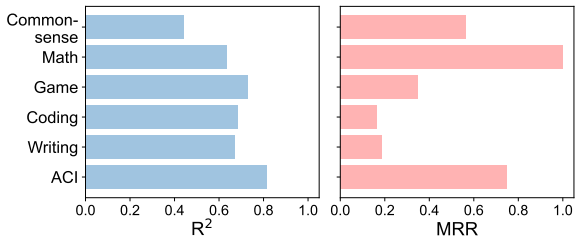


Figure 7: Results of predicting ACI and task performances using group indicators, evaluated by R^2 and mean reciprocal rank (MRR).

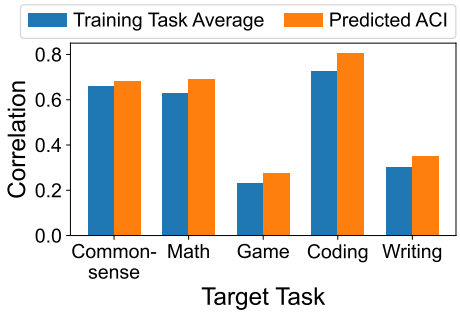


Figure 8: We choose each of the five tasks as the target task, and use data from the remaining four tasks to predict the performance of LLM agent groups on the target task. We compare the performance of predicted ACI and a baseline that simply averages the group's performance on the four training tasks. The performance is measured by the Pearson correlation coefficient.

tralized network. Additionally, since the edges in a centralized network are a subset of the edges in a decentralized network, the centralized network has a lower time and token cost than the decentralized network. Such results suggest that with proper design of communication topology, the agent group can achieve better performance with lower cost.

5.2 Generalization to New Groups

We further examine whether the previously defined group indicators that predict ACI can generalize to unseen groups. Specifically, we conduct a 2-fold cross-validation experiment, using half of the groups to fit a random forest regression model and predict the ACI or task performance of the rest groups based on their indicators. As shown in Figure 7, the R^2 achieves over 0.8 on ACI prediction, and over 0.6 on most of the tasks, suggesting a good generalization ability. Note that the indicators for LLM agent groups are solely dependent on the configuration of multi-agent collaboration, and there is no need to test the group on the target tasks. As a result, these indicators offer a promising way to predict the performance of new groups without incurring time or token costs.

Moreover, in some cases, such as when designing a multi-agent system, the goal is to identify the best-performing group instead of predicting exact performance. Therefore, we also present the mean reciprocal rank (MRR) metric for predicting the best group in Figure 7. The results indicate that the MRRs for ACI, Commonsense, and Game exceed 0.35, meaning the best group is typically within the top-3 predicted groups. For the coding and writing tasks, the best group can be found within the top-6 predicted groups. On the Math dataset, the model can even achieve 100% accuracy in identifying the best LLM agent group. These findings highlight the potential of using these indicators to optimize the design of LLM multi-agent systems.

5.3 Generalization to New Tasks

Since ACI captures the general capability of LLM agent groups, we further examine whether it can be used to predict their performance on unseen tasks. Specifically, we adopt a leave-one-out evaluation setting, in which each of the five benchmarks is selected in turn as the target task, while the remaining four benchmarks are used as training tasks. Using only the performance on the four training tasks, we train a random forest regression model to predict the leave-one-out ACI based on group indicators. We then evaluate how well the predicted ACI reflects performance on the target task by computing the Pearson correlation coefficient between them. Additionally, we compare with a simple baseline by averaging the group's performance over the training tasks.

As shown in Figure 8, the correlations are all positive and exceed 0.6 on the commonsense, math, and coding tasks, indicating that the predicted ACI can effectively estimate the performance of LLM agent groups on unseen tasks. Moreover, the predicted ACI consistently outperforms the average-performance baseline, suggesting that ACI captures general group ability beyond a naive aggregation of task-specific performances.

6 Generalizable MAS Design with ACI

There have been extensive studies on optimizing the communication topology of MASs to improve their performance on certain tasks (Zhuge et al., 2024; Zhang et al., 2025b). However, most existing methods are task-specific and require retraining for each task, which limits their ability to generalize to

Table 1: Comparison of MAS optimization with and without predicted ACI across different tasks.

Method	Commonsense	Math	Game	Coding	Writing
GPTSwarm	0.75	0.81	0.50	0.87	0.88
+*Predicted ACI*	0.77	0.91	0.53	0.89	0.89
Improvement	2.7%	12.3%	6.0%	2.3%	1.1%
G-Designer	0.73	0.86	0.51	0.84	0.88
+*Predicted ACI*	0.77	0.89	0.55	0.86	0.90
Improvement	5.5%	3.5%	7.8%	2.4%	2.3%

unseen tasks. In this section, we demonstrate that ACI can serve as a plug-in regularization term that effectively enhances the generalization ability of existing MAS optimization methods.

6.1 Experiment Settings

Existing MAS optimization methods typically aim to optimize a task-specific utility function:

$$\max_{\mathbf{G}} u(\mathbf{G}), \tag{5}$$

where $\mathbf{G}$ denotes the communication topology of the MAS, and the utility function $u(\cdot)$ is usually defined as the performance (e.g., accuracy) on a particular benchmark. To improve generalization across tasks, we augment the optimization objective by incorporating ACI as a regularization term:

$$\max_{\mathbf{G}} u(\mathbf{G}) + \lambda \cdot \hat{ACI}(\mathbf{G}), \tag{6}$$

where λ is a hyperparameter, and $\hat{ACI}(\mathbf{G})$ denotes the predicted ACI of the MAS. Here we use predicted ACI as it is time-consuming to calculate the true ACI for unseen MAS, and we use a random forest model to predict the ACI as mentioned in Section 5.3.

We evaluate our approach on two representative MAS optimization methods: GPTSwarm (Zhuge et al., 2024) and G-Designer (Zhang et al., 2025b). GPTSwarm optimizes the probabilistic distribution of edges in the MAS, while G-Designer trains a variational graph auto-encoder to generate communication topologies. We construct an MAS consisting of eight agents instantiated with our selected eight LLMs, and apply these methods to optimize the communication topology between them.

To assess generalization performance, we adopt a leave-one-out setting. Specifically, four tasks are used to train the MAS optimization model, while the remaining task is held out for evaluation. The ACI prediction model is also trained exclusively on the same four training tasks. For each target task, we compare the performance of the vanilla MAS optimization method with its ACI-augmented variant.

6.2 Results and Analysis

As shown in Table 1, incorporating ACI consistently improves performance on unseen tasks for both GPTSwarm and G-Designer by 1.1% to 12.3%. Such results demonstrate that ACI can serve as a regularization term to enhance the generalization ability of MAS optimization methods.

The improvement can be attributed to the fact that different characteristics of MAS contribute differently across tasks. Some structural features yield strong performance across a wide range of tasks, reflecting general collaborative capability. For example, we have shown that MAS with high skill congruence may exhibit strong general ability (Section 5.1). In the meantime, some other features of MAS are highly specialized and only suitable for particular tasks. Existing methods optimize solely for task-specific utility, which tends to overfit these task-dependent features, resulting in limited transferability to new tasks. By contrast, ACI captures properties that reflect more general capabilities of an MAS. As a result, when used as a regularization term, ACI biases the optimization process toward topologies that preserve such generalizable characteristics, thereby improving performance on unseen tasks.

7 Conclusion

In this study, we investigated the presence of an ACI factor in LLM agent groups, examining their general abilities across diverse tasks. Our extensive experiments revealed that LLM agent groups exhibit a generalizable ACI factor, accounting for 66.3% of the variance in performance, which can well predict the performance on new tasks. Furthermore, our analysis identified collaboration processes as the most critical determinant of ACI, mirroring patterns observed in human groups. This insight underscores the importance of designing

effective collaboration strategies to enhance MAS performance and provide guidelines for MAS design. Finally, we demonstrated that key indicators of LLM agent groups can be leveraged to predict the performance of unseen groups and on new tasks, based on which we propose a plug-in to improve the generalization ability of MAS optimization algorithms. Overall, our findings contribute to a deeper understanding of collective intelligence in LLM agent groups and pave the way for more efficient and generalizable MASs.

Limitations

While this study provides an initial exploration of the ACI in LLM agent groups, several limitations must be noted. First, our findings are primarily based on empirical analysis rather than theoretical frameworks, which have limits on the understanding of the underlying mechanism of ACI. Second, regarding the multi-agent collaboration method, we focus on one typical multi-agent collaboration framework (Du et al., 2024). Other collaboration strategies may be further considered to offer more insights for ACI. Finally, following the settings in human experiments (Riedl et al., 2021), we only consider groups with fewer than 10 agents. Although this scale is consistent with most of the existing LLM multi-agent collaboration frameworks (Qian et al., 2024; Chen et al., 2023; Hong et al., 2024; Li et al., 2023; Du et al., 2024), the scalability of ACI in larger LLM agent groups remains an open question. Future exploration is needed to understand the pattern of ACI with larger group sizes.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFC3307605; in part by the National Natural Science Foundation of China under Grant 62472241.

References

Lin Chen, Yunke Zhang, Jie Feng, Haoye Chai, Honglin Zhang, Bingbing Fan, Yibo Ma, Shiyuan Zhang, Nian Li, Tianhui Liu, and 1 others. 2025. Ai agent behavioral science. *arXiv preprint arXiv:2506.06366*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, and 1 others. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Paul Erdős and Alfréd Rényi. 1960. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–60.

Andrew Estornell and Yang Liu. Multi-llm debate: Framework, principals, and interventions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

J Richard Hackman. 1978. The design of work in the 1980s. *Organizational Dynamics*, 7(1):3–17.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. Metagpt: Meta programming for A multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

David Ilić and Gilles E Gignac. 2024. Evidence of interrelated cognitive-like capabilities in large language models: Indications of artificial general intelligence or achievement? *Intelligence*, 106:101858.

junyou li, Qin Zhang, Yangbin Yu, QIANG FU, and Deheng Ye. 2024. More agents is all you need. *Transactions on Machine Learning Research*.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2024. Self-refine: Iterative refinement

with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, and 1 others. 2024. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186.

Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2025. Scaling large language model-based multi-agent collaboration. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Christoph Riedl, Young Ji Kim, Pranav Gupta, Thomas W Malone, and Anita Williams Woolley. 2021. Quantifying collective intelligence in human groups. *Proceedings of the National Academy of Sciences*, 118(21):e2005737118.

Andries P. Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. 2024. Should we be going mad? A look at multi-agent debate strategies for llms. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Charles Spearman. 1904. "general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–293.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Samuel PL Veissière, Axel Constant, Maxwell JD Ramstead, Karl J Friston, and Laurence J Kirmayer. 2020. Thinking through other minds: A variational approach to cognition and culture. *Behavioral and brain sciences*, 43:e90.

Junlin Wang, Jue WANG, Ben Athiwaratkun, Ce Zhang, and James Zou. 2025. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442.

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

Ziyang Xiao, Dongxiang Zhang, Yangjun Wu, Lilin Xu, Yuan Jessica Wang, Xiongwei Han, Xiaojin Fu, Tao Zhong, Jia Zeng, Mingli Song, and 1 others. 2023. Chain-of-experts: When llms meet complex operations research problems. In *The Twelfth International Conference on Learning Representations*.

Miao Yu, Shilong Wang, Guibin Zhang, Junyuan Mao, Chenlong Yin, Qijiong Liu, Kun Wang, Qingsong Wen, and Yang Wang. 2025. Netsafe: Exploring the topological safety of multi-agent system. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 2905–2938. Association for Computational Linguistics.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. 2025a. Cut the crap: An economical communication pipeline for llm-based multi-agent systems. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. 2025b. G-designer: Architecting multi-agent communication topologies via graph neural networks. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14544–14607. Association for Computational Linguistics.

Han Zhou, Xingchen Wan, Ruoxi Sun, Hamid Palangi, Shariq Iqbal, Ivan Vulić, Anna Korhonen, and Sercan Ö Arık. 2025. Multi-agent design: Optimizing agents with better prompts and topologies. *arXiv preprint arXiv:2502.02533*.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. 2024. Gptswarm: Language agents as optimizable graphs. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

A Appendix

A.1 Implementation Details

A.1.1 Details of LLM Agent Groups

The details of different communication topologies are as follows.

- **Decentralized Network**: It is defined as a fully connected graph in which every pair of nodes is connected by a unique edge, i.e., each agent can receive the answers from all other agents.
- **Centralized Network**: It corresponds to a star graph structure where a central node is connected to all other nodes.
- **Random Network**: We generate random graphs using the Erdős–Rényi (ER) model (Erdős and Rényi, 1960) and Watts–Strogatz (WS) model (Watts and Strogatz, 1998). In the ER model, each pair of vertices is independently connected with a certain probability p. The WS model generates small-world networks by starting with a regular lattice and randomly rewiring edges with a certain probability.

We present the communication topologies and LLMs of all LLM agent groups here, including centralized networks (Figure 9), decentralized networks (Figure 10), and random networks (Figure 11). For each topology, there are two LLM agent groups with 2 rounds and 3 rounds. We also present the ACI of each LLM agent group in the figures.

A.1.2 Datasets and Metrics

The details of the five benchmarks we use are as follows.

- **Commonsense**: We choose the MMLU-Pro (Wang et al., 2024) benchmark, which is a more challenging version of MMLU (Hendrycks et al.) dataset containing multiple-choice questions with four to ten options. It contains problems from various disciplines, serving as a benchmark to test the general knowledge and commonsense reasoning ability of LLMs. The performance of LLM is measured by accuracy.
- **Math**: We use the MATH (Hendrycks et al.) benchmark, which contains math problems to test the mathematical reasoning ability of LLMs. The performance is measured by accuracy.

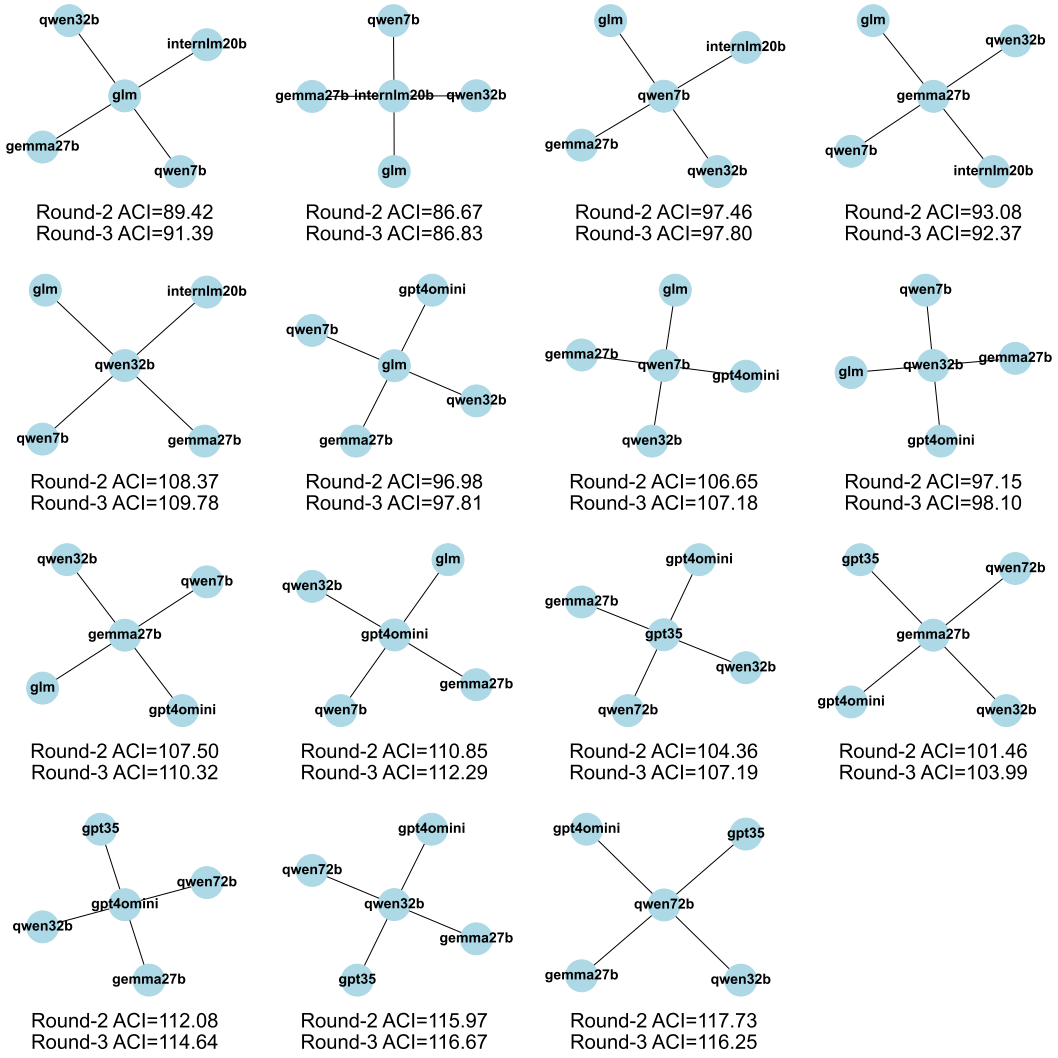


Figure 9: Topologies and ACIs of LLM agent groups with centralized network structure.

- **Game**: We use the Chess move validity tasks from BIG-Bench Benchmark (Srivastava et al., 2023), where the LLM agent is asked to provide a valid move of a piece given the history of chess moves. The performance is also measured by accuracy.

- **Coding**: We choose HumanEval (Chen et al., 2021), a widely used benchmark to measure the ability of function-level code generation. We use the *pass@1* metric to measure the correctness of generated functions on test cases.

- **Writing**: We use the Commongen-Hard (Madaan et al., 2024) benchmark. Each problem in this dataset consists of 20-30 concepts, and the agent is asked to generate coherent sentences that include all these concepts, which measures its reasoning and text generation ability. The performance is measured by the percentage of covered concepts (Chen et al., 2023).

We use the prompts from the datasets' original papers for all tasks and adopt a zero-shot setting. To ensure the diversity of the agents' output, we set the temperature parameter to 1.0 for all experiments (Zhang et al., 2025b). We employ majority voting to aggregate the answers of all agents in a group. Specifically, for closed-ended questions (Commonsense, Math, Game), we calculate the most frequent answer. For open-ended questions (Coding, Writing), we follow a previous work (junyou li et al., 2024) and find the answer that is most similar to others, i.e.,

$$r^{(t)} = \arg\max_{r_i} \sum_{j=1, j\neq i}^{N} sim(r_i^{(t)}, r_j^{(t)}), \quad (7)$$

where $r_i^{(t)}$ is the response of agent v_i at round t, and the similarity is calculated as BLEU score (Papineni et al., 2002).

A.1.3 Computer Resources

All experiments are conducted on Windows 10 OS. The Python version is 3.10.12. We use LLM API provided by Azure OpenAI [1] (for OpenAI models) and SiliconFlow [2] (for non-OpenAI models). The factor analysis is implemented using Python package `factor_analyzer` [3]. The code and original data to calculate ACI and reproduce figures in this paper are released at https://github.com/tsinghua-fib-lab/LLM_Collective_Intelligence.

A.2 Further Discussion

A.2.1 Guidelines for LLM Agent Group Design

There have been studies on optimizing the configurations of LLM agent groups, such as prompt and topology, to improve their performance on certain tasks (Zhuge et al., 2024; Zhang et al., 2025b,a). While these works are based on the assumption that the optimal group structure varies across different tasks, our study indicates that an LLM agent group has a general factor that characterizes its ability across tasks. This might seem contradictory at first glance, but the relationships between our study and these studies can be explained as follows. The ACI we find actually captures the *capability* (or *potential*) of a group instead of its *performance* on certain tasks. According to previous analysis, ACI captures both the individual ability and the alignment of individual abilities during the collaboration process, which is facilitated by group members' capacity to understand and interpret the intentions and goals of others (Veissière et al., 2020). This capability can predict the general task performance to some degree, while the performance is also affected by the characteristics of the specific task. This could somehow be demonstrated by the difference in the importance of collaboration process and individual intelligence (Figure 5). For example, on the writing task that requires divergent thinking and aggregation of ideas from different agents, the collaboration process contributes more to the performance. On the contrary, performance on the game task with closed-ended questions is more affected by individual intelligence.

On the other hand, our findings can serve as a general principle to guide task-specific group structure optimization algorithms. For example, we find that it is generally better to place strong agents on nodes with higher degrees. However, this principle does not specify the exact topology of the group, as the optimal structure may still depend on the specific task, which can be found by optimization algorithms. Moreover, we demonstrate in Section 5.2 that we can predict the performance of groups based on some indicators as well as rank the best groups, which could be integrated into group optimization algorithms to make them more economical.

Overall, based on previous findings, we summarize the following guidelines for designing LLM agent groups:

- First, select high-performing LLMs. This is intuitive, and experimental results show that the individual intelligence of group members is a strong predictor of ACI.
- Second, align agents' efforts with their capabilities. This is supported by the finding that skill congruence is the most important predictor of ACI. In other words, assigning more capable agents to nodes with higher degrees will maximize their influence on the group, leading to better performance.
- Third, simply increasing the group size or effort does not yield significant benefits. Both the number of agents and the number of rounds have a minimal effect on ACI. Furthermore, creating a fully connected network among agents, as some previous studies suggest (Du et al., 2024; Estornell and Liu), is not necessary.
- Finally, it is possible to predict group performance and identify optimal configurations without conducting extensive experiments, thus reducing the cost of optimization algorithms.

A.2.2 Code of Ethics

All datasets used in this study are publicly available, which involves no problem regarding privacy and copyright. No personally identifiable information was collected or used. We cite the resources in Section 3.3.

[1] https://azure.microsoft.com/en-us/products/ai-services/openai-service

[2] https://siliconflow.cn/

[3] https://github.com/EducationalTestingService/factor_analyzer

A.2.3 Broader Impacts

The implications of our findings are particularly significant for the development of Artificial General Intelligence (AGI). The emergence of a generalizable, task-independent ACI factor in LLM agents suggests that LLM agent groups possess an inherent mechanism that influences performance across various tasks. This mechanism could be related to factors such as agents' mutual understanding, shared cognitive processes, and the way they integrate their individual capabilities into a cohesive group effort. The presence of the ACI factor exhibits a form of *general intelligence* among the agents, which transcends specific tasks and contributes to their overall adaptability and effectiveness. Moreover, our findings point to the critical importance of collaboration in LLM agent groups. ACI in LLM agent groups demonstrates that, beyond individual capabilities, the way in which agents interact and collaborate can significantly affect their collective problem-solving abilities. This insight is foundational for advancing AGI, as it suggests that achieving human-like intelligence in artificial systems may depend less on replicating individual cognitive capabilities and more on fostering efficient collaboration within multi-agent frameworks. Finally, the ability of LLM agents to exhibit a general intelligence factor, akin to human groups, also implies that scaling and optimizing these systems for increasingly complex tasks could follow a similar trajectory to human cognitive development, further accelerating the path toward AGI.

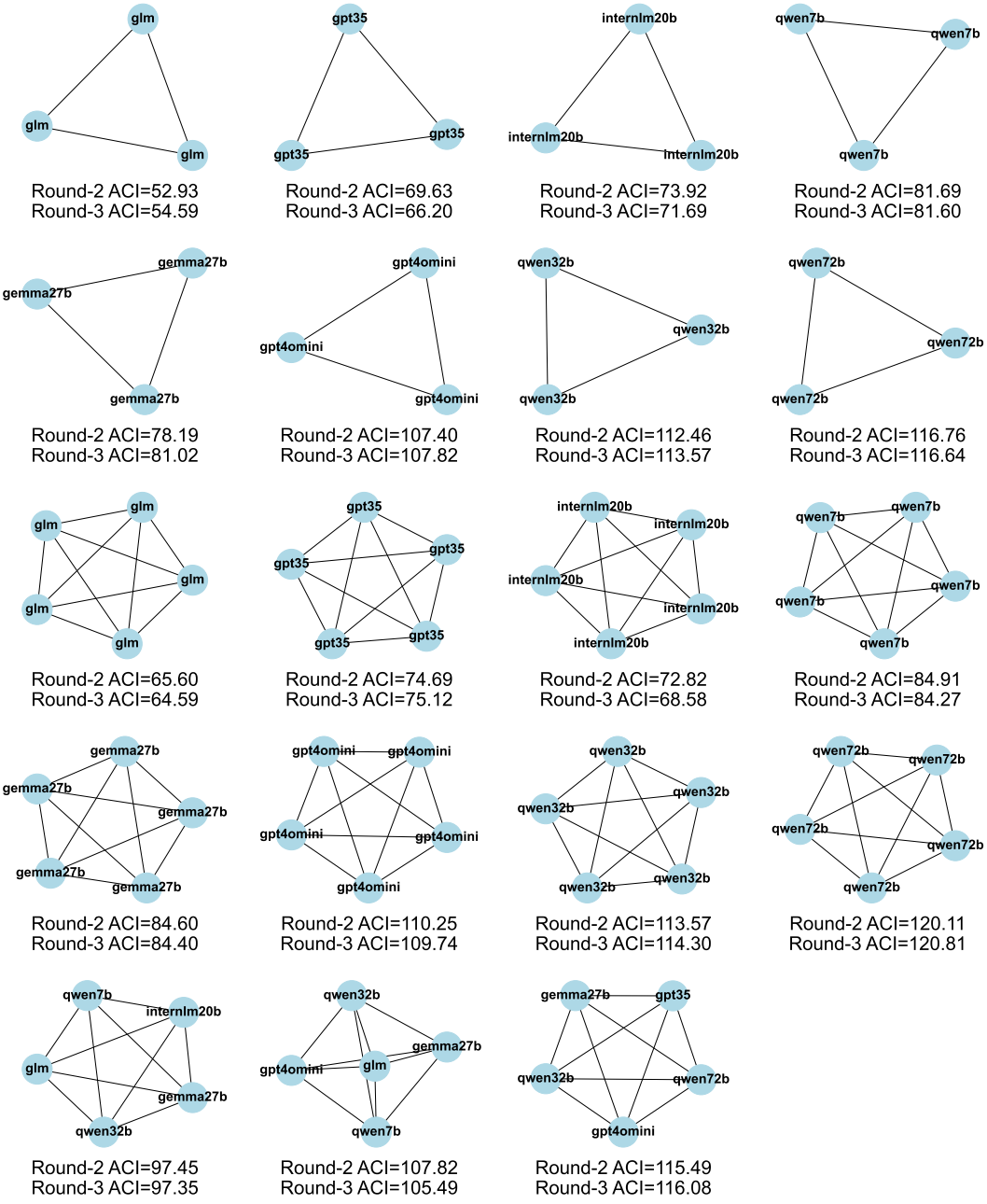


Figure 10: Topologies and ACIs of LLM agent groups with decentralized network structure.

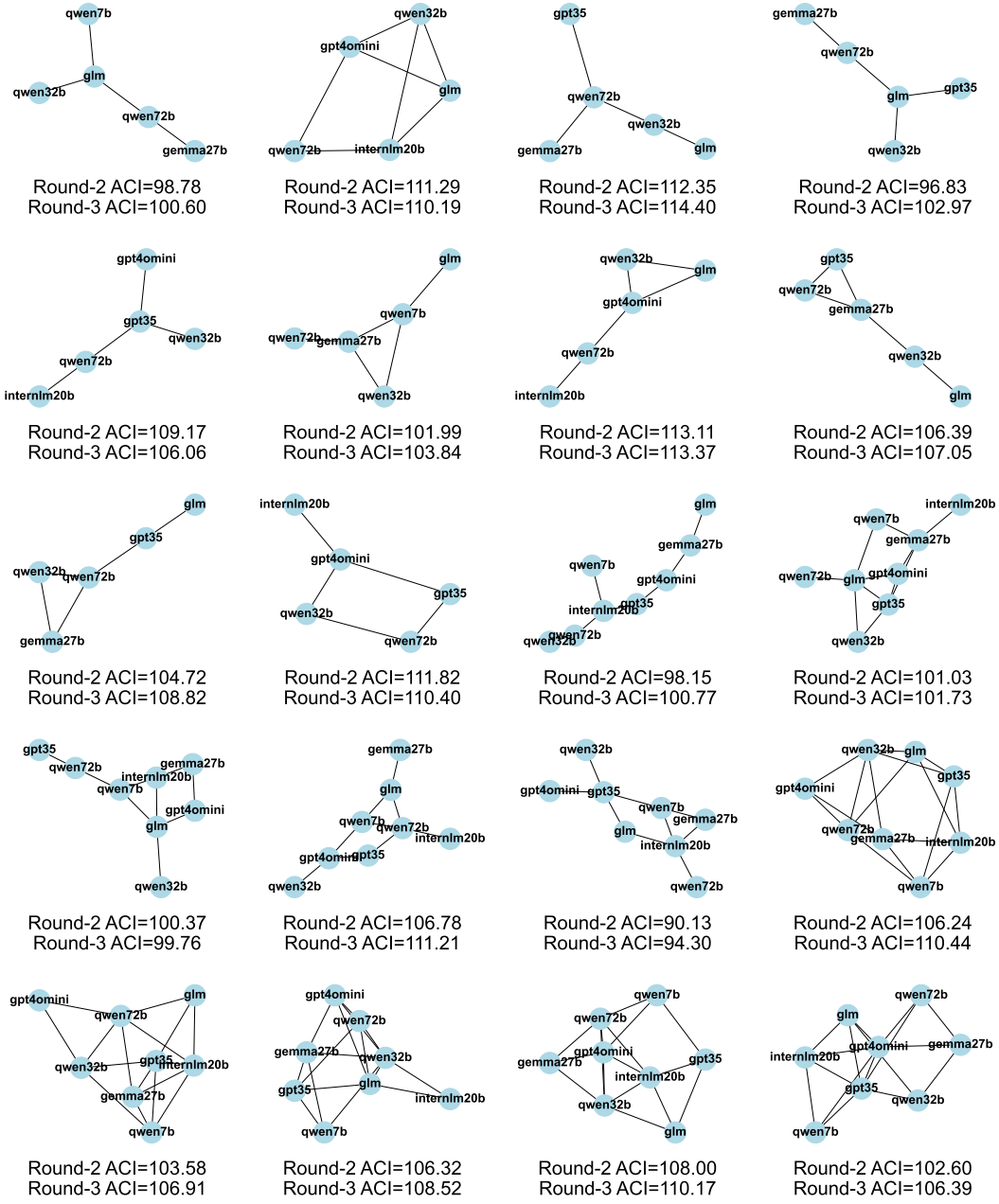


Figure 11: Topologies and ACIs of LLM agent groups with random network structure.