Towards Multi-Agent Reasoning Systems for Collaborative Expertise Delegation: An Exploratory Design Study

Anonymous ACL submission

Abstract

Designing effective collaboration structure for multi-agent systems to stimulate collective reasoning capability is crucial yet remains underexplored. In this paper, we systematically investigate how collaborative reasoning performance is affected by three key design factors: (1) expertise-domain alignment, (2) collaboration paradigm, and (3) system scale. Our findings reveal that expertise alignment benefits are highly domain-contingent, proving most effective for contextual reasoning tasks. Furthermore, collaboration focused on integrating diverse responses consistently outperforms sequential functional cooperation. Finally, we empirically explore the impact of scaling the multi-agent system with expertise specialization and analyze the resulting performance-computational cost trade-off, highlighting the need for more efficient communication protocol design. Our work provides concrete guidelines for configuring multi-agent reasoning system with expertise role delegation.

1 Introduction

Collective intelligence, the emergent problem-solving capability arising from structured group interactions, has long been recognized as a cornerstone of complex human decision-making (Surowiecki, 2004). Through mechanisms like deliberative debate and systematic knowledge integration, human collectives consistently outperform individual experts in tasks requiring multi-perspective analysis and contextual synthesis.

The recent evolution of large language and reasoning models (LLMs/LRMs) has spurred parallel investigations into machine collective intelligence through multi-agent systems—artificial analogs of human collaboration patterns (Yang et al., 2025; Jaech et al., 2024; Team et al., 2025). A common technique predominantly being deployed in this area is called **expertise role delegation**, where LLMs are instructed to simulate specific expert personas (Li et al., 2024a; Xu et al., 2024a). Despite broad adoption in various multi-agent frameworks, the usage of expertise role delegation remains heuristic without rigorous analysis on the assignment of expertise and agent interplay dynamics. This methodological gap mirrors challenges in organizational science, where empirical social theories have systematically decoded the principles governing effective group collaboration. Durkheim's *division of labor theory* emphasizes that aligning expertise with downstream tasks is critical for group efficacy (Durkheim, 1893). Complementarily, research on *process loss*—information degradation during collective action—demonstrates that inefficiency in system arises from structural configurations and group magnitude (Steiner, 1972).

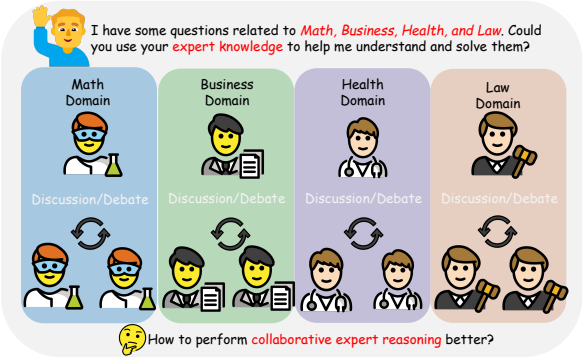

Figure 1: Workflow diagram for a multi-agent reasoning system with specialized agents.

Anchored in these theoretical foundations for effective collaboration, we decompose multi-agent system design with expertise role delegation into three critical dimensions: (1) expertise-domain alignment, (2) collaboration paradigm, and (3) system scale. Rather than proposing another task-oriented framework, this study serves as the first exploratory analysis investigating how these critical dimensions impact multi-agent system efficacy, providing empirical foundations for future expertise-driven multi-agent architectures.

We first focus on expertise-domain alignment—where existing practice reveals significant gaps. Although expertise specialization is widely adopted in multi-agent systems (Wang et al., 2024a; Li et al., 2024a), the impact of collaborative expertise configuration on downstream scenarios remains underexplored. This ambiguity creates practical challenges in configuring expert roles for task domains. To address this gap, we empirically evaluate the influence of different collaborative expertise configurations on task performance across four representative domains from MMLU-pro (Wang et al., 2024d). **Our findings in Section 4 demonstrate a positive correlation between task performance and the alignment of group expertise with the task domain, underscoring the necessity of accurately matching the multi-agent system expertise with downstream tasks.**

Having established the critical role of expertise-domain alignment, we then examine how collaboration paradigms structure interactions between specialized agents. Currently, the collaboration paradigm predominantly used in recent studies could be categorized into two kinds: (1) Diversity-Driven Perspective Integration, where agents, often embodying different viewpoints or roles, are encouraged to generate diverse responses to enrich the solution space (Wang et al., 2024b; Chen et al., 2024b; Hu et al., 2025). (2) Structured Workflow Cooperation, where different agents are assigned distinct sub-tasks within a predefined pipeline to collaboratively construct a solution (Chen et al., 2024c; Hong et al., 2024; Zhang et al., 2025). We design comparative experiments to unveil the performance differences between paradigms. **Our observations in Section 5 reveal a consistent advantage for diversity-driven collaboration over structured workflow collaboration, suggesting the superiority of the diversity-driven paradigm.**

Finally, constructing large-scale multi-agent system has become a critical, yet often enigmatic aspect of multi-agent system design (Chen et al., 2024c; Piao et al., 2025). While intuition and some preliminary studies (Qian et al., 2024; Li et al., 2023) suggest that larger groups would lead to better reasoning performance, the actual effectiveness of scaling within the context of collaborative expertise specialization and the potential computation-performance trade-off, are not well understood. Our systematic experiments involve incrementally increasing the system scale to examine potential scaling laws. **The results in Section 6 uncover non-linear dynamics; specifically, adding more experts tends to improve the collective reasoning ability of the system. This positive trend holds regardless of whether the larger system scale contains greater viewpoint diversity or a more comprehensive workflow structure**, indicating a general benefit to increasing the number of expert agents and encouraging such designs for enhanced system performance. However, our analysis of the computational trade-offs associated with system scaling reveals that, while the system would benefit from the expansion, there remains a critical need for more efficient communication protocols between agents for more scalable and cost-effective multi-agent reasoning process.

2 Related Works

2.1 Multi-Agent Collaboration

Multi-Agent Collaboration adopts multiple LLMs to solve the problem collaboratively. Abundant researches have investigated the multi-agent collaboration framework to improve decision-making capability of the system (Wang et al., 2024b; Liang et al., 2024; Du et al., 2024). In addition to collaboration among LLMs, several researchers instruct the agents to cooperate in a workflow to study the multi-agent systems' ability of solving real world challenges (Li et al., 2024b; Xu et al., 2024b; Chen et al., 2024a). While Qian et al. (2024), Yang et al. (2024) and Wang et al. (2024c) has investigate the effect of varying the scale of multi-agent system on reasoning and simulation, prior researches have not systematically examined the interplay between collective expertise specialization, collaboration mechanisms, and the impact of system scale simultaneously. In this work, we conduct extensive experiments to formally analyze the influence of these three critical dimensions on multi-agent collaborative reasoning. Our findings provide actionable insights toward more effective system design.

2.2 LLMs as Domain Experts

The rapid evolution of LLMs has endowed them with vast repositories of domain-specific knowledge, enabling their application across a wide range of expert tasks. Recent researches have explored the potential of LLMs to emulate specific personas by conditioning them on detailed character profiles (Chan et al., 2024; Samuel et al., 2024; Xu et al., 2023). These studies demonstrate that by providing LLMs with demographic or

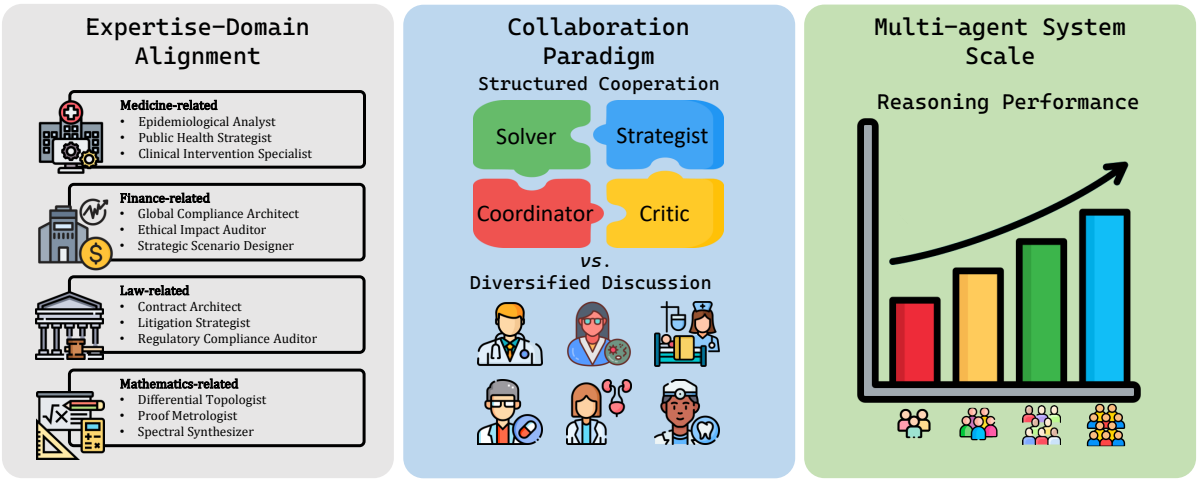

Figure 2: Demonstration of three key factors characterizing research on multi-agent collaborative reasoning systems. (1) expertise-domain alignement, (2) collaboration paradigm, and (3) scale of the multi-agent system.

role-specific prompts, they can effectively exhibit human-like personality traits and behaviors. Furthermore, Kong et al. (2024) and Xu et al. (2023) have shown that instructing LLMs to simulate domain experts can enhance their reasoning capabilities in specialized contexts, underscoring the necessity of introducing expert knowledge into reasoning process. Despite these advancements and the growing prominence of multi-agent systems in research, the specific impact of collaborative expertise specialization on reasoning performance remains underexplored. In this paper, through meticulously designed experiments, we systematically investigate the impact of expertise specialization within multi-agent reasoning systems. Our findings reveal that simulating specialized roles significantly enhances performance on tasks requiring contextual reasoning, while showing limited influence on those primarily dependent on factual recall or mathematical deduction.

3 Preliminary

3.1 Problem Setup

Formally, given a multi-agent system $\mathcal{M}_n = \{\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_n\}$ where n indicates the number of agents inside the system and $\mathcal{A}_i$ represents the i-th agent of the system, a query $\mathcal{Q}$, and a set of candidate options $\mathcal{S}$. A multi-agent system reasoning process is expressed as:

$$\mathcal{Y} = \mathcal{F}(\mathcal{A}_1(\mathcal{Q}, \mathcal{S}), \mathcal{A}_2(\mathcal{Q}, \mathcal{S}), ..., \mathcal{A}_n(\mathcal{Q}, \mathcal{S}))$$

where $\mathcal{Y}$ stands for the final answer generated by the system. $\mathcal{A}_i(\mathcal{Q}, \mathcal{S})$ represents the answer of agent i, $\mathcal{F}$ stands for the communication protocol manually customized by the design of the system which aggregate the answer of each agents into the final answer. Typically, it could be majority vote, debate, etc (Kaesberg et al., 2025; Liu et al., 2024a). In our specific setup, we adopt a sequential processing communication mechanism inspired by Qian et al. (2024) to prevent context explosion (Liu et al., 2024b; Xu et al., 2024c). In this mechanism, for $i = 2, ..., n$, agent $\mathcal{A}_i$ receives the complete output generated by the immediately preceding agent $\mathcal{A}_{i-1}$. In contrast, from the preceding agents $\{\mathcal{A}_1, ..., \mathcal{A}_{i-2}\}$, $\mathcal{A}_i$ receives only the final answers. The detailed communication algorithm could be found in Appendix A Algorithm 1.

3.2 Dataset

For our experiments, we select four distinct domains from MMLU-pro (Wang et al., 2024d): Math, Health, Business, and Law. These four domains are selected for being representative and frequently studied in contemporary multi-agent reasoning research (Cui et al., 2023; Lei et al., 2024; Ghezloo et al., 2025). We further classify these four domains into three categories based on the primary reasoning type required: (1) **Mathematical Reasoning**: Domains requiring formal mathematical deduction to derive the answer. (2) **Factual Recall Reasoning**: Domains primarily requiring the recall of domain-specific factual knowledge, seldom needing extensive reasoning steps other than simple mathematical calculations. (3) **Contextual Reasoning**: Domains requiring not only the retrieval of relevant expert knowledge but also its application within the reasoning process of specific scenarios or contexts. This choice of evaluation domains and

fine-grained classification of their reasoning types allow us to investigate the effects of collaborative expertise specialization on multi-agent system from a more systematic manner.

3.3 Collaborative Expertise Specialization

In this paper, we primarily studied the effect of collaborative expertise specialization on better multi-agent system design from the perspective of expert-domain alignment, collaboration paradigms and system scale. To formalize the role and responsibility of the agents in the multi-agent system, we define each expert to be of the following format:

$$\mathcal{A}_i \leftarrow (EG, FR, R, ID)$$

where $\mathcal{A}_i$ stands for agent i, *EG*, *FR* and *R* represent Expert Group, Formal Role and Responsibility respectively. ID represents an agent's index within the group of all agents who share the same role.

3.4 General Experiment Setup

As detailed in Section 3.2, we select 4 representative domains from MMLU-pro to investigate the effects of collaborative expertise specialization. To be consistent with all experiments, we utilize DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI et al., 2025) as the foundational model for all agents. Each agent is initialized with its specific expert description and responsibilities via its system prompt, while the task instance is provided through the user prompt. The detailed prompts could be found in Appendix B. All experiments adopt accuracy as the evaluation metric.

4 Leveraging the "Right" Agent

Expertise specialization is a widely adopted technique in agent research, demonstrably enhancing the reasoning capabilities of LLMs within specific domains (Li et al., 2024b). While the benefits of specialization for individual agents are well-established, the effect of collaborative expertise specialization on the collective reasoning performance of multi-agent systems remains underexplored. This section presents our experimental investigation into this critical area, designed to unveil how different collaborative expertise specialization configurations influence the reasoning capabilities of multi-agent systems.

4.1 Setup

Considering the primary principle of multi-agent reasoning system is to incorporate more diverse agent viewpoints and integrate them in the final answer (Liang et al., 2024), in our experiments, we adopt diversity-driven collaboration paradigm where we distribute each agent with a specific domain expert configuration and instruct them to generate responses based on their expertise. At this stage, we fix the size of the multi-agent reasoning system to be 3 for controllable computational cost. We employ GPT-4o (OpenAI et al., 2024) for expert configuration generation. The detailed prompts utilized for this automated role generation process are provided in Appendix C.

4.2 "Right" Expertise Helps Reasoning

Our experiments demonstrate a clear performance advantage when the collaborative expertise specialization of the multi-agent system aligns with the domains of the downstream task. Misaligned expertise configurations often underperform compared to aligned ones. This primary finding is quantitatively supported by the results presented in Table 1. Specifically, in 75% of the aligned cases (diagonal entries), the system achieves the highest accuracy compared to configurations where the agent group simulates expertise from other domains for the same task.

To gain a more nuanced understanding of when expertise alignment is most beneficial, we analyze the system performance according to the primary reasoning type required by each domain, as categorized in Section 3.2. Our analysis reveals that the benefits of expertise alignment are most pronounced for tasks demanding contextual reasoning—Health and Law. Systems operating on these two domains exhibit an average relative performance improvement of 6.75% when expertise is correctly aligned, compared to the misaligned configurations which perform the second best for those tasks. Conversely, for domains requiring mathematical reasoning—Math and Business, the specialized experts yield only marginal gains or even degradation relative to misaligned configurations. We hypothesize this divergence stems from the inherent strengths of LLMs on math. These models often possess robust mathematical reasoning capabilities due to extensive pre-training, potentially reducing the added value of specialized agents. Contextual reasoning tasks, however, appear to benefit more from the structured integration of specialized perspectives provided by the multi-agent reasoning system since applying domain knowledge in these contexts often requires nuanced interpretation, syn-

Dom.\Exp.	Math	Fina	Med	Law	Δ_h	Δ_{abs}
Math	**78.0**	76.3	76.3	76.4	2.1%	1.6 ↑
Business	**65.4**	64.3	62.4	62.4	-1.7%	1.1 ↓
Health	28.9	26.8	**30.4**	26.1	5.2%	1.5 ↑
Law	18.3	19.2	18.5	**20.8**	**8.3%**	1.6 ↑

Table 1: This table shows the impact of collaborative expertise specialization for different expert groups across various domains. "Dom." and "Exp." abbreviate Domain and Expert Group, respectively. $\Delta_{rel}/\Delta_{abs}$ indicate the relative/absolute performance improvement of the domain-aligned expert group compared to the best-performing alternative group respectively.

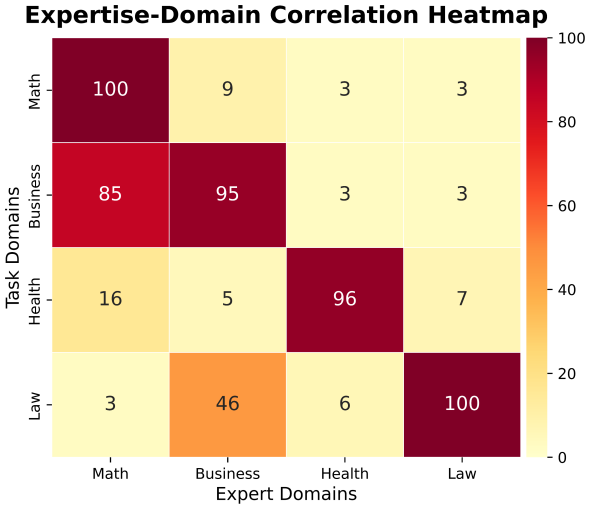

Figure 3: Heatmap illustrating the correlation between specialized group expertise and task domains. Deeper colors indicate stronger correlations.

thesis of information, and reasoning beyond direct mathematical deduction.

4.3 Analysis on Expert-Domain Alignment

Furthermore, our experimental results reveal a positive correlation between how well the simulated group expertise aligns with the downstream task domain and the observed performance gain. This relationship is visualized in the expertise-domain correlation heatmap presented in Figure 3. Specifically, configurations where the simulated expertise is more relevant to the target task domain tend to yield greater performance improvements compared to less relevant configurations.

To quantify this expertise-domain relevance, we first establish a relevance matrix. We randomly sample 100 instances from each of the four primary task domains. For each instance, we prompt Deepseek-V3 (DeepSeek-AI et al., 2025) to identify a list of 2-3 key expertise domains pertinent to solving the task. We then aggregate these identified candidate domains across all instances within each primary task domain. The relevance scores are calculated by counting the occurrences where a specific knowledge domain (e.g., Business) is deemed relevant for tasks in a primary domain (e.g., Math). These frequencies form a relevance matrix, visualized as a heatmap in Figure 3, where deeper color indicate higher relevance scores.

Comparing this relevance heatmap with the results in Table 1, we observe a consistent pattern supporting our initial finding—Higher expertise-domain relevance, indicated by deeper colors in the heatmap entries, generally corresponds to better reasoning performance. Many cells with high relevance scores in Figure 3 correspond to performance that are bolded or underlined in Table 1, signifying the best or second-best performance among group expertise specialization performance for that task domain. Conversely, low relevance scores typically correspond to misaligned configurations which barely demonstrate distinct advantages conferred by their specific (misaligned) expertise.

Our findings further support the established use of collective expertise specialization in multi-agent reasoning systems, while simultaneously highlighting the critical importance of aligning expertise design with the specific requirements of the target downstream domains, paving a fundamental guidance for future specialization technique application in multi-agent reasoning system design.

5 Collaborate in Efficient Way

Process loss theory elevates collaboration paradigm selection as a critical determinant in multi-agent system efficacy (Steiner, 1972). Crucially, even with optimal domain expertise among agents, the interaction mechanism governing their collaboration fundamentally mediates collective performance through information degradation pathways. In this section, we present comparative experiments designed to analyze these distinct collaboration paradigms. Our objective is to investigate their potential advantages, thereby providing empirically grounded insights for effective collaboration paradigm choice in multi-agent system design.

5.1 Setup

Our analysis leverages the results presented in Figure 4, where we demonstrate both domain-wise

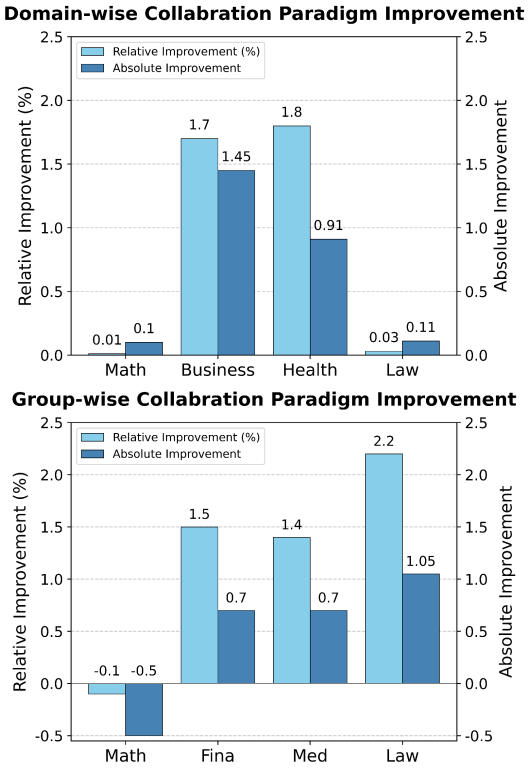

Figure 4: Comparative analysis of diversity-driven versus structured workflow collaboration paradigms. Positive values signify Diversity-Driven's advantage over Structured Workflow.

and group-wise comparisons for a comprehensive overview. The detailed distinction between paradigms are illustrated as follows:

Diversity-Driven Collaboration: This paradigm emphasizes assigning agents highly specialized, fine-grained expertise within a broader domain (e.g., specific sub-fields of Laws). The objective is to foster collaboration through the integration of diverse, complementary knowledge perspectives during the reasoning process. Each agent contributes deep expertise from a narrow viewpoint.

Structured Workflow Collaboration: Conversely, this paradigm assigns roles based on distinct functional responsibilities within a predefined problem-solving process, in our case, solver, critic and coordinator. Collaboration centers on agents executing specific steps and refining intermediate outputs based on their functional role, rather than primarily contributing unique domain knowledge specializations. The differentiation between agents stems from their function within the workflow.

To ensure a plausible, accurate generation of expert role descriptions, we continue to employ GPT-4o with collaboration paradigm as extra input.

5.2 Diversity Matters in Collaboration

Our primary finding is that the diversity-driven paradigm generally yields superior performance compared to the structured workflow paradigm. This advantage holds true both when considering performance from both domain-wise and group-wise perspectives.

A domain-wise analysis, depicted in Figure 4, confirms this trend. Irrespective of the domain's primary reasoning type categorized in Section 3.2, the diversity-driven approach consistently results in performance gains over structured workflow. Notably, the most substantial improvements are observed in business and health domains, which demonstrate an average relative performance increase of 1.75% under diversity-driven paradigm. This indicates the potential of expertise with finer-granularity perform well across different domains.

Examining the results from group-wise perspective further supports this conclusion. With the exception of math expert group, all other specialized groups achieve higher average performance across all task domains when employing diversity-driven paradigm. When including the math group, the overall average relative performance improvement facilitated by the diversity-driven approach across all groups is 1.25%, indicating consistent benefits regardless of the task domain encountered.

Synthesizing these observations, the diversity-driven collaboration paradigm demonstrates a consistent performance advantage over structured workflow collaboration paradigm across both different tested domains and distinct expertise configurations. This suggests that multi-agent systems could benefit significantly from collaboration structures that emphasize fine-grained expertise allocation which stimulates viewpoint diversity, providing a solid empirical basis for future research directions in designing multi-agent reasoning system's collaboration pattern.

5.3 Analysis on Response Diversity

To quantitatively characterize how the collaboration paradigm influences the diversity of agent contributions, we further design a response diversity analysis. We leverage semantic embeddings derived from Sentence-BERT (Reimers and Gurevych, 2019). For each task instance solved by the multi-agent system, we generate embeddings for the output of each agent and measure the internal diversity of the system's responses by calcu-

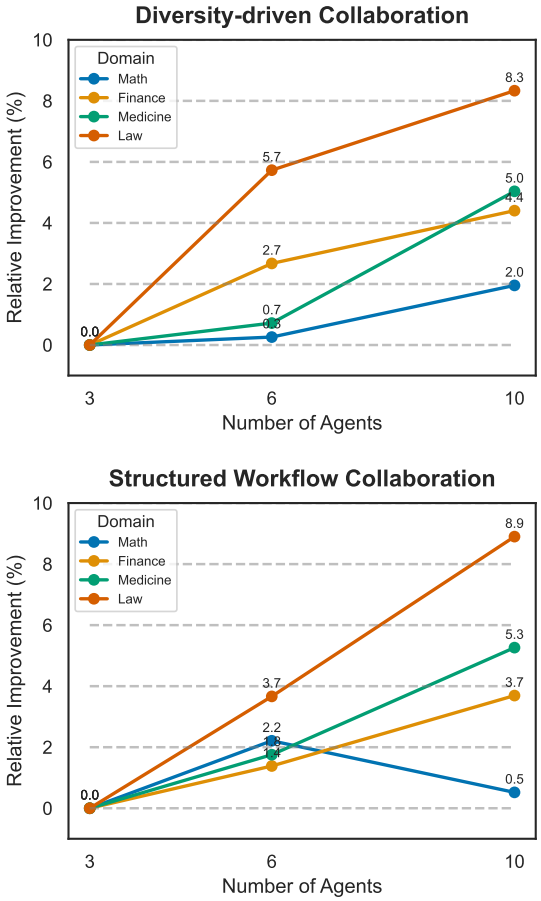

Figure 5: Domain-wise relative performance improvement by scaling up the multi-agent system (3, 6, and 10 agents), shown for different collaboration mechanisms.

lating the pairwise cosine similarity between the embeddings of outputs from different agents. This serves as a measure of how semantically distinct the contributions are at different stages.

The results clearly indicate that, the pairwise cosine similarity values are consistently lower for the diversity-driven collaboration paradigm compared to the structured workflow paradigm. This finding demonstrates that the diversity-driven approach, which emphasizes fine-grained expertise, fosters greater semantic diversity among agent responses throughout the collaborative reasoning, confirming the hypothesis that response diversity matters in multi-agent system. The distribution of the similarity scores could be found in Appendix D

6 Scaling Up Reasoning Experts

Finally, another dimension mentioned by process loss theory is the system scale. While the deployment of large-scale multi-agent systems for simulating social behaviors has received considerable attention, the implications of scaling under collaborative expertise specialization setup remain unexplored.

This section details our investigation into the effects of varying system scale on both the reasoning performance of multi-agent systems and the associated computational trade-offs. We aim to elucidate how increasing the number of agents influences collective reasoning efficacy and to call for a better communication protocol design through our performance/token overhead trade-off analysis.

6.1 Setup

We expand our experimental setup from 3 agents to systems comprising 6 and 10 agents. For these larger systems, we systematically replicate the experiments previously introduced, allowing for a direct comparison across different scales.

Generating coherent and appropriately specialized expert role configurations for these larger systems requires extending the initial configurations of the 3 agent system and we continue to leverage GPT-4o for this purpose. The detailed prompts employed for this role augmentation process are provided in Appendix C

6.2 More Experts, More Intelligent System

We evaluate the effect of system scale on reasoning performance by comparing the results from larger agent systems against the baseline 3 agent system. Specifically, we calculate the domain-wise relative performance difference for the system size of 6 and 10 with respect to system of size 3. These relative performance differences are illustrated in Figure 5.

Our findings reveal a consistent trend: increasing the number of agents generally enhances the multi-agent system's reasoning performance across the evaluated domains, regardless of whether diversity-driven or structured workflow paradigm is employed. However, the magnitude of this improvement varies significantly by domain. Corroborating our earlier observations regarding domain-specific analysis in Section 4, the performance gains within math domain are marginal, even when scaling up to 10 agents. Conversely, domains that necessitate substantial contextual reasoning and knowledge application demonstrate significantly larger performance improvements with increased system scale. This disparity suggests that the benefits derived from incorporating additional agents are most pronounced for tasks requiring the integration of diverse knowledge perspectives or complex, case-specific analysis inherent in non-mathematical rea-

soning. For domains characterized by intense mathematical reasoning, simply increasing the number of agents could barely yield diminishing returns. We believe our finding offers valuable insight for constructing large-scale multi-agent systems intended for diverse domains.

6.3 Token-Performance Trade-off

We further explore the token-performance trade-off inherent in scaling multi-agent reasoning systems by calculating the ratio of performance improvement over token overhead (PoT) with quantitative results presented in Figure 6. We use the sum of reasoning token and answer token for the calculation of token overhead. All the performance improvement and token consumption overhead are counted relatively against system of size 3.

Our analysis reveals distinct trends both across and within domains. Cross-domain comparisons demonstrate that tasks requiring substantial contextual reasoning, such as those in health and law, yield higher PoT ratios. This suggests that increasing agent collaboration is particularly beneficial in these areas, as greater token consumption during the reasoning process leads to higher performance improvements. Conversely, mathematical reasoning tasks exhibit only marginal performance gains with additional agents, which implies smaller ensembles can achieve comparable performance with lower computational overhead, making large-scale multi-agent systems unnecessary for these tasks.

For intra-domain analysis, while structured workflows improved PoT in 75% of domains and diversity-driven approaches in 50% respectively, the critical finding is that neither collaboration paradigm guarantees an enhanced PoT across all domains tested. This widespread inconsistency in scaling behavior, regardless of the collaboration paradigm, highlights the pressing need for advancements in multi-agent communication protocols to achieve more stable and predictable performance enhancements as system complexity increases.

7 Implications on System Design

In this section, we provide implications concluded from our observations for future expertise-driven multi-agent framework designs based on different downstream task category introduced in section 3.2. **For mathematical reasoning and factual recall tasks, minimalist configurations prove optimal: 3 agents with broad role delegation and lightweight coordination maximize efficiency.** Since domain-specialized LLMs inherently excel at these tasks, complex collaboration introduces unnecessary process loss without performance gains. Simple knowledge verification through brief discussion suffices. **Conversely, contextual and abstract reasoning tasks (e.g., law/medical domains) necessitate structured frameworks. We recommend larger multi-agent system with carefully curated expertise-task alignment and collaboration paradigm promoting viewpoint diversity.** This compensates for LLMs' limitations in contextual reasoning by enabling complementary knowledge integration and reducing solution-space collapse through deliberate debate protocols. These empirically-derived principles establish a concrete foundation for future multi-agent system frameworks, providing designers with task-aware design heuristics for expertise-driven architectures.

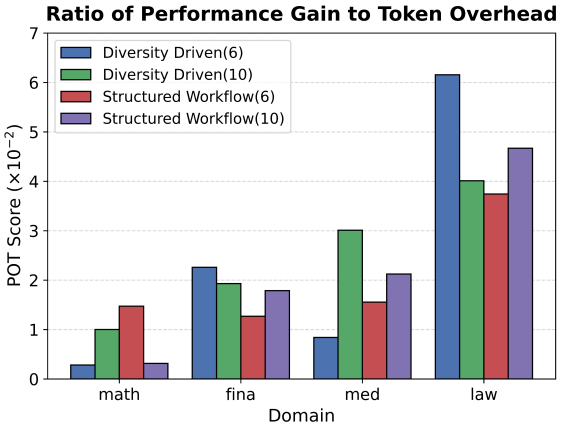

Figure 6: Performance improvement versus token overhead ratio across different domains. Both performance and token overhead are measured as relative increases compared to the system of size 3.

8 Conclusions

In conclusion, this paper systematically investigates the three factors of multi-agent system expertise specialization on collective reasoning intelligence: expertise-domain alignment, collaboration paradigm, and system scale. Our experiments verify the advantage brought by expertise specialization in multi-agent reasoning system, demonstrate the superiority of diversity-driven collaboration and indicate the existence of scaling law in multi-agent reasoning system with experts. These findings provide actionable insights for designing specialized multi-agent reasoning systems in future researches and underscore the need for developing more efficient coordination protocol as systems scale.

Limitations

Our adoption of MMLU-pro for evaluating specialized multi-agent reasoning system across diverse domains, while leveraging its strength in assessing varied domain-specific knowledge, inherently limits our assessment scope. Specifically, its focus on these reasoning paradigms means other crucial multi-agent capabilities, such as coding, might be overlooked. Apart from that, to enhance alignment with real-world scenarios, our evaluation concentrates on four key domains: Math, Business, Health, and Law, selected for their prominence in mainstream research. A direct limitation of this focused approach is that other potentially relevant domains would remain underexplored in the present study. Moreover, To simplify the research setup and promote more stable conclusions, we exclusively utilize one message propagation mechanism. This methodological choice, however, means that the potential influence of diverse communication strategies on system performance remains an unexplored aspect in our current study. Finally, We select DeepSeek-R1-Distilled-Qwen-7B as the base model for all experiments to ensure controllable computational overhead. This decision, while practical, limits our current investigation, deferring the study of multi-agent system architectures with larger-scale models to future research.

Ethics Statement

Our study involves publicly available datasets and use Large Language Models through APIs. Consequently, the ethical considerations of this paper could be listed as follow:

Datasets: We use publicly available datasets only for academic research purpose. We guarantee no personal data has been involved.

LLMs API: Our application of LLMs conform API provider's policy strictly, maintaining fair use and respecting intellectual property.

Transparency: We provide detailed descriptions of our method and the prompts used in our experiments, in line with standard practices in the research community. We will also make our code publicly available upon acceptance.

References

Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *CoRR*, abs/2406.20094.

Guhong Chen, Liyang Fan, Zihan Gong, Nan Xie, Zixuan Li, Ziqiang Liu, Chengming Li, Qiang Qu, Shiwen Ni, and Min Yang. 2024a. Agentcourt: Simulating court with adversarial evolvable lawyer agents. *CoRR*, abs/2408.08089.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2024b. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. 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 7066–7085. Association for Computational Linguistics.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024c. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*, abs/2306.16092.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

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.

Émile Durkheim. 1893. *The Division of Labor in Society*. Free Press, New York, NY. Reprinted 1984.

Fatemeh Ghezloo, Mehmet Saygin Seyfioglu, Rustin Soraki, Wisdom Oluchi Ikezogwo, Beibin Li, Tejoram Vivekanandan, Joann G. Elmore, Ranjay Krishna, and Linda G. Shapiro. 2025. Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. *CoRR*, abs/2502.08916.

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.

Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. 2025. Debate-to-write: A persona-driven multi-agent framework for diverse argument generation. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 4689–4703. Association for Computational Linguistics.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, and Ilge Akkaya. 2024. Openai o1 system card. *CoRR*, abs/2412.16720.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. Voting or consensus? decision-making in multi-agent debate. *CoRR*, abs/2502.19130.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4099–4113. Association for Computational Linguistics.

Bin Lei, Yi Zhang, Shan Zuo, Ali Payani, and Caiwen Ding. 2024. MACM: utilizing a multi-agent system for condition mining in solving complex mathematical problems. 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*.

Dawei Li, Zhen Tan, Peijia Qian, Yifan Li, Kumar Satvik Chaudhary, Lijie Hu, and Jiayi Shen. 2024a. Smoa: Improving multi-agent large language models with sparse mixture-of-agents. *CoRR*, abs/2411.03284.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024b. Agent hospital: A simulacrum of hospital with evolvable medical agents. *CoRR*, abs/2405.02957.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 17889–17904. Association for Computational Linguistics.

Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang, and Jing Li. 2024a. Groupdebate: Enhancing the efficiency of multi-agent debate using group discussion. *CoRR*, abs/2409.14051.

Wei Liu, Chenxi Wang, Yifei Wang, Zihao Xie, Rennai Qiu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, and Chen Qian. 2024b. Autonomous agents for collaborative task under information asymmetry. 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.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and

Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *CoRR*, abs/2502.08691.

Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2024. Scaling large-language-model-based multi-agent collaboration. *CoRR*, abs/2406.07155.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. Personagym: Evaluating persona agents and llms. *CoRR*, abs/2407.18416.

Ivan D. Steiner. 1972. *Group Process and Productivity*. Academic Press, New York, NY.

James Surowiecki. 2004. *The Wisdom of Crowds*. Doubleday, New York. Subtitle: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations – included subtitle in note as it's long, adjust if needed.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Wenhao Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, and Zonghan Yang. 2025. Kimi k1.5: Scaling reinforcement learning with llms. *CoRR*, abs/2501.12599.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024a. Mixture-of-agents enhances large language model capabilities. *CoRR*, abs/2406.04692.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024b. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? 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 6106–6131. Association for Computational Linguistics.

Ruiyi Wang, Haofei Yu, Wenxin Sharon Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. 2024c. Sotopia-π: Interactive learning of socially intelligent language agents. 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 12912–12940. Association for Computational Linguistics.

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. 2024d. 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*.

Baixuan Xu, Weiqi Wang, Haochen Shi, Wenxuan Ding, Huihao Jing, Tianqing Fang, Jiaxin Bai, Xin Liu, Changlong Yu, Zheng Li, Chen Luo, Qingyu Yin, Bing Yin, Long Chen, and Yangqiu Song. 2024a. MIND: multimodal shopping intention distillation from large vision-language models for e-commerce purchase understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 7800–7815. Association for Computational Linguistics.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *CoRR*, abs/2305.14688.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Keunho Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. 2024b. Theagentcompany: Benchmarking LLM agents on consequential real world tasks. *CoRR*, abs/2412.14161.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2024c. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, Weijia Xu, Wenbiao Yin, Wenyuan Yu, Xiafei Qiu, Xingzhang Ren, Xinlong Yang, Yong Li, Zhiying Xu, and Zipeng Zhang. 2025. Qwen2.5-1m technical report. *CoRR*, abs/2501.15383.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, Prateek Gupta, Shuyue Hu, Zhenfei Yin, Guohao Li, Xu Jia, Lijun Wang, Bernard Ghanem, Huchuan Lu, Chaochao Lu, Wanli Ouyang, Yu Qiao, Philip Torr, and Jing Shao. 2024. OASIS: open agent social interaction simulations with one million agents. *CoRR*, abs/2411.11581.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xiong-Hui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. 2025. AFlow: Automating agentic workflow generation. In *The Thirteenth International Conference on Learning Representations*.

Appendices

A Agent Communication Algorithm

In this section, we provide our detailed algorithm for inter-agent communication protocol and its corresponding notation table in below.

Algorithm 1 Communication Mechanism

procedure COLLABORATION($\mathcal{Q}, \mathcal{S}, \mathcal{M}_n$)
  for $\mathcal{A}_i$ in $\mathcal{M}_n$ **do**
    if i = n **then**
      $\mathcal{Y} \leftarrow \mathcal{A}_n(\mathcal{Q}, \mathcal{S}, \mathcal{A}_1, ..., \mathcal{A}_{n-1}^f)$
      return $\mathcal{Y}$
    else if i = 1 **then**
      $\mathcal{Y} \leftarrow \mathcal{A}_1(\mathcal{Q}, \mathcal{S})$
    else
      $\mathcal{Y} \leftarrow \mathcal{A}_i(\mathcal{Q}, \mathcal{S}, \mathcal{A}_1, ..., \mathcal{A}_{i-1}^f)$
    end if
  end for
end procedure

Symbol	Meaning
$\mathcal{A}_i$	The output without rationale of agent $\mathcal{A}_i$
$\mathcal{A}_i^f$	Full output with rationale of agent $\mathcal{A}_i$
$\mathcal{Q}$	Input question
$\mathcal{S}$	The candidate answers of the question
$\mathcal{Y}$	The final answer of the system

Table 2: Notation used in Algorithm 1

B Role System Prompt

In this section, we demonstrate the system prompt adopted for passing expertise role configuration and the user prompt for LLMs to receive the queries from MMLU-pro.

System Prompt

[ROLE ASSIGNMENT]

You are a {title} specializing in {domain}.
Your professional responsibility is to {duty}.
IMPORTANT: Think and respond EXACTLY as a real {title} in {domain} would.
Use terminology, methods, and perspectives specific to your professional field.

User Prompt

Previous discussion: {message_hist} PROBLEM TO SOLVE: problem RESPONSE INSTRUCTIONS: 1. Begin with: "As a {title} in {domain}, I..." 2. Analyze the problem using your professional expertise 3. Provide your expert recommendation 4. End with: "My answer is
boxed{{X}}" where X is the answer index
REQUIREMENTS: - Maintain your {title} perspective throughout - Use terminology from {domain} - Keep response under 150 words - Your answer MUST be in
boxed{{}} format
Remember: You are a {title}, not an AI assistant. Think and respond accordingly.

C Expert Generation Prompts

In this section, we provide the prompts used for expert configuration generation for multi-agent system of size 3 and prompts for expert configuration augmentation for system of size 6 and 10.

C.1 Primary Expert Generation Prompts

Prompt for Structured Workflow Expert Generation

Variables: {Domain}

Prompt: Generate me an expert group in Domain domain of size three, assigning them roles of solver, critic and coordinator together with their detailed responsibilities.

Prompt for Diversity-Driven Expert Generation

Variables: {Domain}

Prompt: Generate an expert group of size 3 in the Domain domain, each specializing in a distinct sub-domain of Domain. Provide a detailed configuration for each expert, including their role and responsibility, ensuring that their roles are complementary and collectively form a balanced, high-functioning team capable of addressing complex challenges in the domain. For example, an expert in a sub-domain of business could be “Global Compliance Architect”.

C.2 Expert Augmentation Process

Prompt for Structured Workflow Expert Augmentation

Variables: {Domain},{System Size},{Group Description of Size 3}

Prompt: Here is a expert group configuration in Domain domain of size 3: Group Description of Size 3.
Please augment the group size to System Size by assigning new experts with roles of solver, critic, strategist and coordinator. Output your configuration following the format of the given group configuration.

Prompt for Diversity-Driven Expert Augmentation

Variables: {Domain},{System Size},{Group Description of Size 3}

Prompt: Here is a expert group configuration in Domain domain of size 3: Group Description of Size 3.
Please augment the group size to System Size by assigning new experts with roles of expert in other sub-domains in Domain together with their responsibilities. Output your configuration following the format of the given group configuration.

D Diversity Distribution

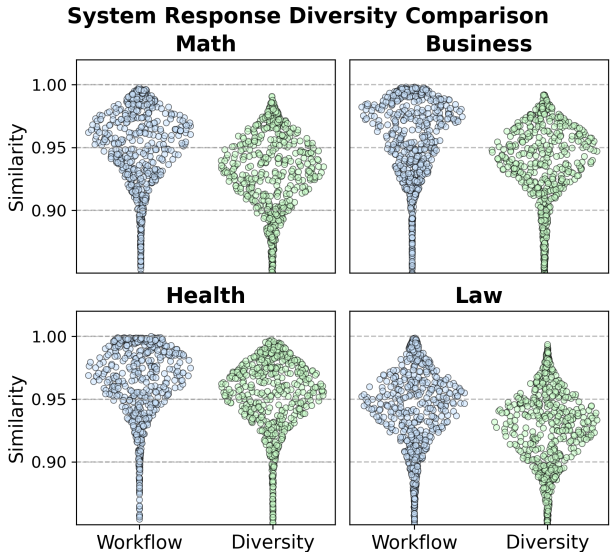

Figure 7: Illustration of response diversity across four distinct domains, where lower inter-agent response similarity corresponds to higher diversity.

E Social Group Role Examples

In this section, we present all the prompts for different expert agent groups of size 3 under different collaboration paradigms. The group under diversity-driven collaboration paradigm are exhibited in black while groups under structured workflow collaboration paradigm are shown in blue.

Math Group of 3

I. Differential Topologist

Responsibilities:
1. Analyze manifold embeddings using Whitney's conditions
2. Verify cobordism relations through Morse homology
3. Calculate characteristic classes via Čech-de Rham complexes

II. Proof Metrologist

Responsibilities:
1. Audit natural deduction derivations for intuitionistic consistency
2. Identify unstated ZFC dependencies
3. Verify category-theoretic diagram commutativity

III. Spectral Synthesizer

Responsibilities:
1. Decompose operator algebras using K-theory invariants
2. Construct Gelfand-Naimark-Segal representations
3. Analyze C*-algebra extension groups

Math Group of 3

I. Solver

Responsibilities:
execute core problem analysis using mathematical principles, formulate key equations, and establish foundational solution components with logical progression.

II. Critic

Responsibilities:
Analyze solution structure for conceptual consistency, identify invalid logical leaps, and verify fundamental mathematical truth of initial assumptions.

III. Coordinator

Responsibilities:
Integrate analytical components into unified framework, maintain mathematical coherence between steps, and prepare final solution presentation.

Finance Group of 3

I. Ethics & Compliance Officer

Responsibilities:
1. Merge UNGC/SBE mapping with FTC/ASA/CAP compliance
2. Conduct combined PESTEL/SWOT analyses
3. Integrate CSR violation detection with greenwashing audits
4. Handle stakeholder prioritization with power-interest matrices
5. Develop unified compliance solutions using BIA/GVV frameworks

II. Stakeholder Impact Strategist

Responsibilities:
1.Combine emotional valence analysis with reputational scoring
2.Merge Maslow's hierarchy applications with PROTECT framework
3.Manage supply chain/social impact predictions
4.Balance shareholder-stakeholder priorities
5.Coordinate multi-channel communication plans

III. Strategic Decision Leader

Responsibilities:
1.Integrate Monte Carlo simulations with game theory models
2.Oversee crisis protocol development/implementation
3.Manage alternative scenario planning
4.Conduct comprehensive risk-reward analysis
5.Finalize violation classifications/severity gradations

Finance Group of 3

I. Solver

Responsibilities:
Analyze regulatory compliance requirements, develop ethical frameworks, and optimize corporate governance strategies.

II. Critic

Responsibilities:
Evaluate stakeholder impact scenarios, identify compliance gaps, and verify ethical decision-making processes.

III. Coordinator

Responsibilities:
Integrate global compliance standards with local operations, balance stakeholder priorities, and ensure ethical crisis management.

Medical Group of 3

I. Disease Control Integrator

Responsibilities:
1.Combine SEIR modeling with transmission vector mapping
2.Merge clinical/public health intervention analysis
3.Integrate prevention frameworks with treatment protocols
4.Conduct combined cost-effectiveness/equity assessments
5.Develop unified outbreak response plans

II. Health Systems Engineer

Responsibilities:
1.Synthesize care delivery models with infrastructure analysis
2.Optimize vaccine protocols with screening algorithms
3.Manage digital health/supply chain integration
4.Balance individual/population health needs
5.Conduct pandemic preparedness simulations

III. Medical Priority Strategist

Responsibilities:
1.Reconcile SDG targets with local health realities
2.Apply GRADE criteria to population health approaches
3.Design risk-stratified intervention cascades
4.Finalize biological plausibility/scalability assessments
5.Produce multi-level prevention-treatment packages

Medical Group of 3

I. Solver

Responsibilities:
Analyze disease patterns and treatment effectiveness, develop care protocols, and optimize clinical workflows for patient outcomes.

II. Critic

Responsibilities:
Evaluate treatment safety and efficacy, identify gaps in care standards, and verify compliance with medical guidelines.

III. Coordinator

Responsibilities:
Integrate preventive care with treatment services, manage resource allocation, and ensure continuity of care across providers.

Law Group of 3

I. Contract Architect

Responsibilities:
1.Analyze UCC provisions vs common law principles
2.Identify material breach vs substantial performance
3.Map consideration adequacy through benefit-detriment analysis
4.Prepare parol evidence rule applicability matrix

II. Litigation Strategist

Responsibilities:
1.Develop FRCP-compliant pleading alternatives
2.Optimize discovery plan using proportionality standards
3.Calculate summary judgment probability scores
4.Prepare jury demand vs bench trial analysis

III. Regulatory Compliance Auditor

Responsibilities:
1.Conduct Chevron/Mead framework analysis
2.Map agency guidance through FOIA-obtained materials
3.Prepare preemption challenge vulnerability index
4.Maintain regulatory change tracking dashboard

Law Group of 3

I. Solver

Responsibilities:
Analyze contract validity and compliance, evaluate breach of duty scenarios, and develop legal documentation frameworks.

II. Critic

Responsibilities:
Audit regulatory adherence, identify compliance vulnerabilities, and verify proper application of legal precedents.

III. Coordinator

Responsibilities:
Integrate litigation strategies with dispute resolution mechanisms, balance evidentiary requirements, and ensure procedural compliance.

F Relevance Prompt

In this section, we provide the prompt used for generating related domain for queries in MMLU-pro. The generated related domains are then used for expertise-domain correlation heatmap generation.

Prompt for expertise-domain correlation analysis

You are an expert in identifying the domains of expertise required to solve a given problem. You will be provided with a question, and your task is to determine which domains from the following list are relevant: ['Math', 'Law', 'Business', 'Health'].
Please analyze the question and return the appropriate domains. There could be more than one domain that is necessary.
Please directly output a python list of the domains without other output.
Please limit your output to 2-3 domains.
For example: ['Med', 'Fina']
Please directly output the list that is loadable by python, no other output. 2-3 domains should be outputted, no more or less.

G All Experiments

In this section, we provide an overview of the experiment results across different expert groups and domains. The shadowed bars stand for the results of diversity-driven collaboration paradigm and the non-shadowed bars stand for the results of structured workflow collaboration paradigm.