Simulating Expert Discussions with Multi-agent for Enhanced Scientific Problem Solving

Ziyue Li, Yuan Chang, Xiaoqiu Le*
National Science Library, Chinese Academy of Sciences
Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences
{liziyue, changyuan, lexq}@mail.las.ac.cn

Abstract

Large Language Models (LLMs) have shown remarkable potential across various domains, yet their application in addressing complex scientific problems remains a formidable challenge. This paper presents a novel methodology to augment the problem-solving capabilities of LLMs by assigning them roles as domain-specific experts. By simulating a panel of experts, each LLM is tasked with delivering professional and cautious responses to scientific inquiries. Our approach involves querying multiple LLMs and assessing the consistency of their responses. High agreement among the LLMs suggests greater confidence in the proposed solution, whereas discrepancies prompt a collaborative discussion among the LLMs to reach a consensus. This method emulates real-world scientific problem-solving processes, fostering a more reliable and robust mechanism for LLMs to tackle scientific questions. Our experimental results show that assigning roles to multiple LLMs as domain-specific experts significantly improves their accuracy and reliability in solving scientific problems. This framework has the potential to advance the application of AI in scientific research, enhancing its effectiveness and trustworthiness.

1 Introduction

Large Language Models (LLMs) have achieved remarkable success in a wide range of natural language processing tasks, including text generation (Swanson et al., 2021; Yang et al., 2023), machine translation (Burda-Lassen, 2023; Alves et al., 2023), and text summarization (Laban et al., 2023). Despite their versatility and strong performance across various domains, the application of LLMs to solving complex scientific problems has remained a significant challenge. The primary obstacle lies not in the absence of domain-specific knowledge within these models, but rather in their limited ability to effectively harness this knowledge when confronted with intricate scientific problems that demand expert-level understanding and reasoning (Addlesee, 2024).

*Corresponding Author

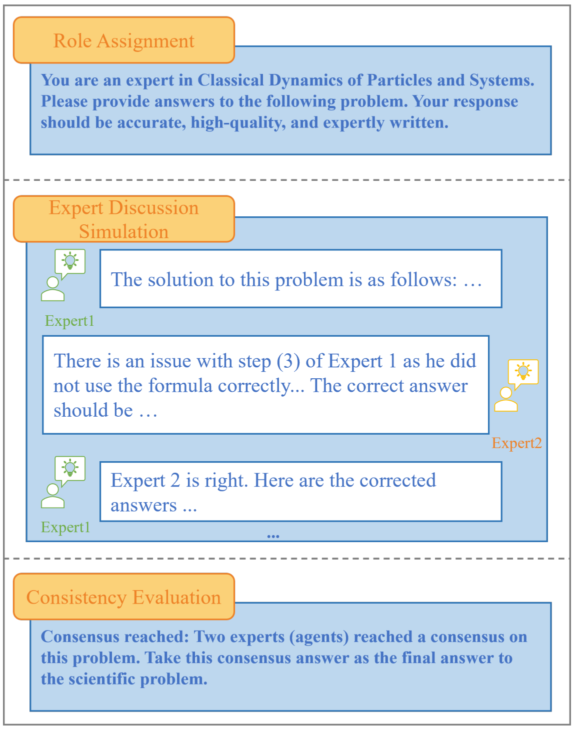

Figure 1: Simulating Expert Discussions with Multi-agent (SEDM).

The application of LLMs to scientific problem-solving presents a unique challenge due to the stringent requirements for precision and reliability in research. Minor inaccuracies can have far-reaching consequences, undermining the validity and trustworthiness of results. While LLMs possess extensive knowledge, their current architectures often struggle to consistently apply this knowledge to meet the rigorous demands of scientific inquiry. This limitation underscores the need for innovative approaches to enhance the problem-solving capabilities of LLMs in specialized domains. Improving the performance of LLMs in accurately and reliably

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 243–256

solving complex scientific problems could significantly advance their utility in research settings and unlock new potentials for artificial intelligence in science.

In this study, we introduce a novel methodology called **Simulating Expert Discussions with Multi-agent (SEDM)** that enhances the problem-solving capabilities of LLMs by assigning them roles as domain-specific experts, as illustrated in Figure 1. This approach involves simulating a panel of experts, where each LLM is tasked with providing professional and cautious responses to scientific inquiries. By querying multiple LLMs and evaluating the consistency of their responses, we can gauge the confidence in the proposed solutions. High agreement among the LLMs indicates greater reliability, while discrepancies trigger a collaborative discussion among the models to reach a consensus. This method mirrors real-world scientific problem-solving processes, fostering a more dependable mechanism for LLMs to address scientific questions. We evaluate the performance of SEDM on a range of problems across various scientific domains, including physics, chemistry, and mathematics. We use accuracy as the evaluation metric and compare SEDM with baseline methods such as direct LLM usage and few-shot learning.

In summary, our contributions are:

- We propose a novel multi-agent framework that assigns specific expert roles to LLMs, enabling them to collaboratively address scientific problems.

- We develop a discussion architecture for multi-agent systems, and experiments have shown that this architecture can effectively enable multiple agents to reach the correct consensus.

- We demonstrate through extensive experiments that our approach significantly improves the accuracy LLMs in scientific problem-solving. For instance, when using the GPT-4 model, SEDM achieves an average accuracy of 57.18% across all subjects, representing an improvement of 32 percentage points compared to direct query and 36 percentage points compared to few-shot learning. These results advance the application of AI in scientific research.

2 Related Work

Large Language Model Reasoning Large language models (LLMs) have demonstrated significant reasoning capabilities, especially when scaled to hundreds of billions of parameters (Ouyang et al., 2022; OpenAI et al., 2024). Various techniques, such as chain-of-thought prompting (Wei et al., 2022; Kojima et al., 2022; Shi et al., 2022) and rationale engineering (Fu et al., 2023; Zhou et al., 2022), have been proposed to further elicit and enhance the reasoning abilities of LLMs. However, despite these advancements, LLMs still struggle with complex reasoning tasks, particularly in the domain of scientific problem-solving (Chen et al., 2023; Wang et al., 2024; Ma et al., 2024). LLMs often struggle to provide reliable and consistent answers to intricate scientific questions (Wang et al., 2024), necessitating the development of novel approaches to improve their reasoning capabilities in this context.

Multi-Model Collaboration and Role-Playing Previous studies have explored the benefits of role-playing in LLMs, demonstrating that assigning distinct roles can lead to more specialized and accurate outputs (Lu et al., 2024; Guan et al., 2024; Tao et al., 2024; Bhattacharyya et al., 2024). Additionally, collaborative frameworks where multiple models interact and discuss to reach a consensus have shown promise in improving the robustness of the generated solutions (Du et al., 2024; Lu et al., 2024; Sadler et al., 2024; Mehta et al., 2024; Figueras et al., 2023; Xiong et al., 2023). Considering the complexity and rigor of scientific research, more effective methods are needed to stimulate the optimal intelligence of multi-agent systems.

Large Language Models in Solving Scientific Problems Recent studies have explored the potential of LLMs in scientific problem-solving, including theorem proof (Dong et al., 2023; Song et al., 2024), hypothesis generation (Qi et al., 2023; Yang et al., 2024) and scientific discovery (Boiko et al., 2023; AI4Science and Quantum, 2023). However, the understanding and reasoning capabilities of LLMs in fundamental STEM(Science, Technology, Engineering, and Mathematics) domains remain underexplored (Wang et al., 2024; Ma et al., 2024). While LLMs exhibit impressive performance on high-level scientific tasks, their ability to grasp complex scientific concepts, engage in rigorous logical reasoning, and provide reliable solutions to domain-specific problems is still uncertain. These challenges necessitate a more nuanced approach to harnessing the full potential of LLMs.

3 Method: Simulating Expert Discussions with Multi-agent

We propose a novel approach called Simulating Expert Discussions with Multi-agent (SEDM) to enhance the scientific problem-solving capabilities of large language models (LLMs) by simulating expert discussions. The overall framework of our methodology is illustrated in Figure 2. We assign multiple LLMs with domain-specific expert roles and simulate a panel discussion among these experts on a given scientific problem. By analyzing and evaluating the consistency of the LLM experts' responses, we derive reliable solutions.

3.1 Role Assignment

In our proposed approach, we assign domain-specific expert roles to multiple LLMs to address a given scientific problem within a particular domain. This assignment of expert roles is motivated by the following rationale:

- **Fostering Collaboration and Consensus** Scientific progress often relies on collaboration and consensus-building among experts within the same domain. By assigning identical roles, we encourage LLMs to engage in simulated collaborative discussions, challenging each other's assumptions, reconciling differences, and ultimately converging towards a consensus solution.

- **Enhancing Reliability through Ensemble Methods** Despite being instances of the same LLM architecture, each individual model may exhibit variations in its outputs due to factors such as random initialization, stochastic sampling, or sensitivity to input perturbations. By employing an ensemble of multiple LLMs with identical roles, we can leverage the collective wisdom of the group, mitigating the impact of individual model instabilities and enhancing the overall reliability of the proposed solutions.

- **Exploring Diverse Reasoning Paths** While sharing the same domain knowledge and expertise, each LLM may explore different reasoning paths and problem-solving strategies when presented with the same scientific problem. Assigning identical roles allows us to capture and analyze these diverse reasoning paths, potentially uncovering novel insights or alternative approaches that a single LLM might overlook.

Through this approach, we create a simulated panel of domain-specific experts with shared expertise but diverse reasoning perspectives. This setup emulates the real-world dynamics of scientific discourse, where experts from the same field evaluate and build upon each other's work, ultimately advancing our understanding of complex scientific problems.

3.2 Expert Discussion Simulation

At the heart of our proposed methodology lies the simulation of a panel discussion among multiple LLMs, each assuming the role of a domain-specific expert within the same scientific field. The overall overview of expert discussion simulation phase is shown in Figure 3. This approach aims to leverage the collective knowledge and diverse perspectives of the LLMs to tackle complex scientific problems effectively. The simulation process encompasses the following key steps:

Problem Presentation The initial step involves presenting a well-defined scientific problem or inquiry to the panel of LLMs.

Individual Responses Upon receiving the problem, each LLM, operating within its assigned expert role, generates an independent response. This response is based on the LLM's knowledge and understanding of the specific sub-discipline or area of specialization it represents. By providing individual responses, the LLMs contribute their unique perspectives and insights to the problem-solving process, mimicking the diversity of opinions often encountered in real-world scientific discussions.

Response Analysis and Comparison Once all the LLMs have provided their individual responses, the next step involves collecting and analyzing these responses for consistency and complementarity. The analysis focuses on identifying areas of agreement and divergence among the LLMs' perspectives. High levels of agreement among the responses suggest a strong consensus and increased confidence in the proposed solution. Conversely, divergent viewpoints highlight areas that require further exploration, clarification, or synthesis, open-

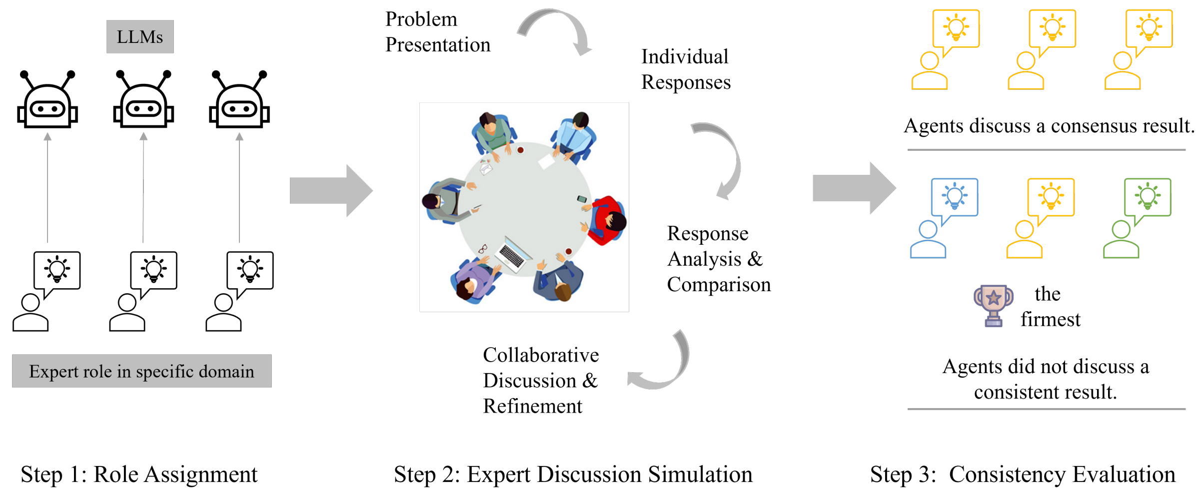

Figure 2: Framework of the SEDM (Simulated Expert Discussion with Multi-agent) approach to enhance LLM scientific problem-solving. Key steps: (1) Role Assignment of domain-specific experts to multiple LLMs; (2) Expert Discussion Simulation involving problem presentation, individual responses, response analysis, and collaborative discussion; (3) Consistency Evaluation - if consensus is reached, the agreed solution is adopted; otherwise, the solution of the most persistent expert (agent) is selected.

ing up opportunities for a more comprehensive understanding of the problem.

Collaborative Discussion and Refinement In cases where the initial responses reveal discrepancies or complementary insights, a collaborative discussion phase is initiated. During this phase, the LLMs engage in a simulated dialogue, exchanging their perspectives, challenging assumptions, and working towards reconciling any differences. This discussion process closely resembles the way experts within the same scientific domain would interact and collaborate in real-world settings, fostering a rigorous and iterative refinement of ideas. Through this iterative process of discussion, the LLMs aim to converge towards a more comprehensive and well-supported solution to the scientific problem at hand.

The simulated expert discussion within a specific domain, as outlined above, harnesses the collective knowledge and diverse perspectives of the LLMs to tackle complex scientific problems. By emulating the rigorous process of scientific inquiry, where ideas are scrutinized, refined, and synthesized through critical discourse among experts, our methodology aims to enhance the problem-solving capabilities of LLMs in scientific domains. This approach not only leverages the strengths of individual LLMs but also promotes a collaborative and iterative problem-solving process, ultimately leading to more reliable and comprehensive solutions to scientific challenges.

3.3 Consistency Evaluation

Upon completion of the expert discussion simulation, we conduct a consistency evaluation of the solutions proposed by the multiple LLMs assuming expert roles. This evaluation process is crucial for ensuring the reliability and robustness of the proposed solutions. Specifically, we employ the following strategies:

Consensus Determination When all experts reach a unanimous agreement during the discussion process, we consider that they have achieved consensus on the given problem. In such cases, we directly adopt the solution unanimously agreed upon by the experts as the final result. The attainment of consensus often indicates that the solution has undergone thorough discussion and argumentation, lending it higher credibility and reliability.

Maximum Discussion Round Limit Recognizing that expert discussions in the real world cannot continue indefinitely, we set a maximum number of discussion rounds for the expert deliberations. This limit serves to prevent the discussion process from entering an endless loop while also encouraging the experts to reach consensus or make decisions within a reasonable timeframe.

If the experts fail to reach complete agreement within the maximum number of discussion rounds, we resort to the following strategy: we select the solution proposed by the expert (agent) who most persistently defended their viewpoint throughout the discussion. The rationale behind this strategy is that

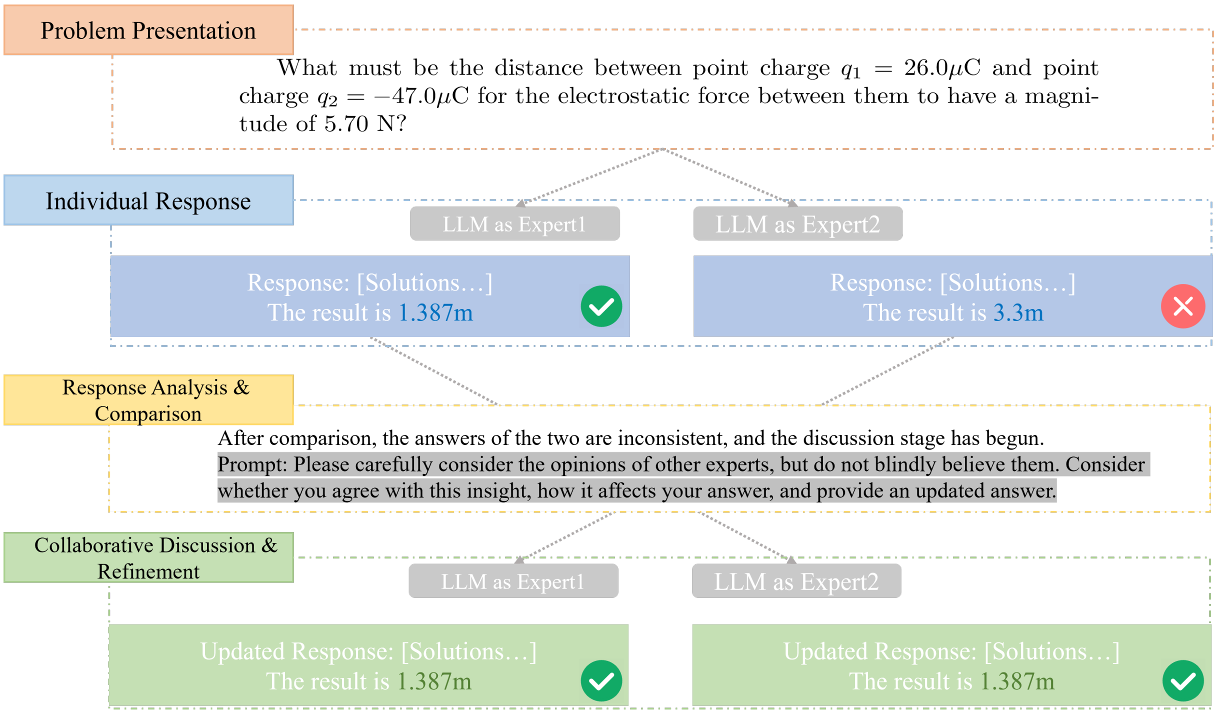

Figure 3: Overview of expert discussion simulation phase.

the expert who firmly maintains their stance likely possesses a deeper understanding of the problem and has provided more comprehensive arguments, rendering their proposed solution more convincing and reliable.

Discussion Convergence Analysis Although we establish a maximum number of discussion rounds, empirical evidence suggests that expert discussions often converge to consensus or a few primary viewpoints within a relatively small number of rounds. We conduct a convergence analysis of the discussion process, quantifying the frequency of achieving consensus or converging to main viewpoints at different round thresholds. Through extensive case studies, we observe that expert discussions typically reach consensus or converge to primary viewpoints within 2-3 rounds. This finding aligns with real-world expert discussion scenarios, demonstrating the effectiveness and practicality of our approach.

By employing these consistency evaluation strategies, we effectively synthesize the opinions of multiple experts to derive reliable and robust problem solutions. Moreover, by analyzing the convergence properties of the discussion process, we validate the efficacy of our method.

4 Experiment

4.1 Experimental Setup

Dataset In our experiment, we used the SciBench dataset (Wang et al., 2024), which is a comprehensive benchmark specifically designed to evaluate the scientific problem-solving ability of Large Language Models (LLMs). SciBench covers university level problems in various scientific disciplines, including mathematics, physics, and chemistry. This dataset includes open-ended questions from textbooks and open-ended questions from undergraduate exams, ensuring a rigorous evaluation of LLM's reasoning and computational skills. In this experiment, we use open-ended questions as testing. This dataset provides a solid foundation for testing and improving LLM's ability to solve problems in complex scientific environments.

Baseline To establish a baseline for our proposed multi-expert framework, we conducted experiments using two state-of-the-art LLMs: GPT-3.5 and GPT-4. Specifically, we utilized the gpt-3.5-turbo-0125 version for GPT-3.5 and the gpt-4-turbo version for GPT-4. For each model, we employed two query methods:

- Direct Querying: The scientific problem was directly presented to the LLM without any additional context or examples.

- Few-Shot Learning (Brown et al., 2020): We provided the LLM with a small set of representative examples of scientific problems and their solutions before presenting the target problem. This approach aims to prime the model with relevant context and improve its performance on the specific task.

4.2 Main Results

We evaluate the performance of our proposed Simulating Expert Discussions with Multi-agent (SEDM) approach against two baselines: direct querying of the LLM and few-shot learning. The experiments are conducted across three scientific domains: physics, chemistry, and mathematics. Each domain is further divided into subdomains, such as thermodynamics and classical mechanics for physics, to assess the model's performance on a diverse range of scientific problems. In the main experiment, we used the setting of 2 experts and 2 discussion rounds.

Table 1 presents the accuracy scores of the models on the test set. The results demonstrate that SEDM significantly outperforms the baselines across most domains and subdomains. For GPT-3.5, SEDM achieves an average accuracy of 33.25%, markedly higher than the 9.59% for Direct response and 9.60% for few-shot learning. Similarly, for GPT-4, SEDM attains an average accuracy of 57.18%, compared to 25.09% for Direct and 21.46% for few-shot.

However, it is important to note that SEDM does not always achieve the highest scores in every subdomain. For instance, in GPT-4's performance in statistics domain, the direct querying approach slightly outperforms SEDM. This may be attributed to the nature of statistical problems, which are often more standardized and formulaic compared to other subdomains. Many statistical problems can be solved by applying specific formulas or algorithms, which aligns well with the strengths of language models. Consequently, direct querying may be sufficient to handle these relatively standard problems.

Despite these few exceptions, SEDM consistently demonstrates robust performance improvements across the majority of subdomains, highlighting its effectiveness and adaptability in enhancing the problem-solving capabilities of LLMs. It is worth noting that the performance of few-shot learning is comparable to or slightly worse than direct querying. This may be due to the limited ability of the selected prompt examples to fully capture the diversity of the domain, leading to a decrease in the performance of few-shot learning.

The results also reveal some variation in performance across subdomains. For instance, in physics, the models achieve higher accuracy in fundamental concepts compared to thermodynamics and classical mechanics. This suggests that the complexity and specificity of the subdomain can influence the model's performance. Nevertheless, SEDM consistently outperforms the baselines in almost all subdomains, demonstrating its robustness and adaptability.

4.3 Further Analysis

Solution Quality In addition to evaluating the accuracy of the models, we also assess the quality of the generated solutions. We randomly sample 100 problems and evaluate the solutions using LLMs and human evaluation based on three criteria: (1) the correctness of the reasoning steps, (2) the clarity of the explanations, and (3) the appropriateness of the mathematical notations and symbols used. Each criterion was rated on a scale of 1 to 5, with 5 being the highest quality.

For the human evaluation, we employed three expert annotators. To ensure reliability, we calculated the inter-annotator agreement using Fleiss' kappa (Fleiss, 1971) for each of the three criteria:

- Correctness of reasoning steps: $\kappa = 0.71$
- Clarity of explanations: $\kappa = 0.62$
- Appropriateness of mathematical notations and symbols: $\kappa = 0.55$

The overall average kappa value was 0.63, indicating substantial agreement among the annotators.

The detailed prompts for LLM evaluation and the specific guidelines for human evaluation are provided in Appendix B. Table 2 presents the average quality scores for solutions from GPT-4. Compared to baseline, SEDM consistently achieves higher quality scores in both LLM and human evaluations. The solutions generated by SEDM demonstrate clearer reasoning steps, more coherent explanations, and more precise use of mathematical notations. This suggests that the multi-expert discussion framework not only improves the accuracy of the solutions but also enhances their overall quality and readability.

Subject		Physics			Chemistry				Math			Avg
		fund	thermo	class	quan	chemmc	atkins	matter	calc	stat	diff	
GPT-3.5	Direct	10.96	2.94	2.13	8.82	20.51	4.67	2.04	9.30	28.00	6.00	9.59
	Few Shot	8.22	1.49	0.00	11.76	15.38	5.61	4.08	13.95	26.67	**10.00**	9.60
	SEDM *	**40.85**	**36.36**	**25.00**	**30.30**	**55.26**	**37.14**	**17.02**	**38.10**	**44.44**	8.00	**33.25**
GPT-4	Direct	15.07	11.94	8.51	14.71	23.08	27.10	22.45	42.86	**56.00**	18.00	25.09
	Few Shot	26.03	5.97	12.77	17.65	30.77	15.87	12.24	33.33	49.33	8.00	21.46
	SEDM *	**81.69**	**27.27**	**37.50**	**57.58**	**81.58**	**59.05**	**53.19**	**78.57**	51.39	**44.00**	**57.18**

Table 1: The accuracy scores (%) of different baseline methods and our proposed SEDM approach across various scientific domains using GPT-3.5 and GPT-4 models under the setting of 2 experts and 2 discussion rounds. The best results for each subject are in **bold**.

Eval. Method	LLM Evaluation			Human Evaluation		
	(1)	(2)	(3)	(1)	(2)	(3)
Direct	3.20	3.40	3.80	4.00	4.25	4.25
SEDM *	4.20	3.60	4.60	4.75	4.25	4.50

Table 2: The average quality score of solutions from GPT-4 evaluated by LLMs and humans.

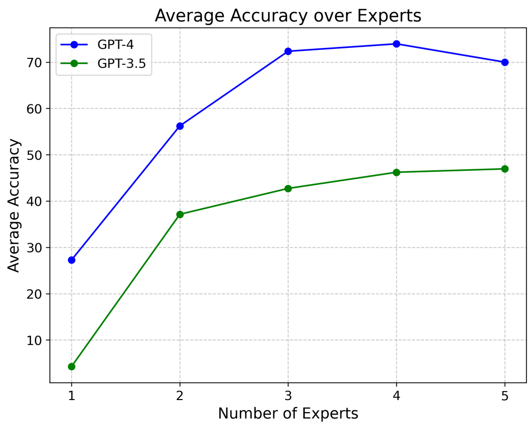

Figure 4: The relationship between the number of experts and the average accuracy of GPT-3.5 and GPT-4.

Number of Experts We investigate the impact of the number of experts in the panel on the performance of SEDM. We vary the number of experts from 2 to 5 and evaluate the accuracy of the generated solutions.Figure 4 shows the relationship between the number of experts and the average accuracy of GPT-3.5 and GPT-4. The results reveal that increasing the number of experts generally leads to higher accuracy. However, the performance gains diminish as the number of experts exceeds 4. This suggests that a panel of 2-4 experts strikes a balance between performance improvement and computational efficiency. Having too many experts may introduce redundancy and increase the computational overhead without significant performance benefits.

Number of Discussion Rounds We also investigate the impact of the number of discussion rounds on the performance of SEDM. We conducte experiments with varying numbers of discussion rounds, ranging from 1 to 5, and measure the accuracy of the generated solutions, as illustrated in Figure 5. The results indicate that increasing the number of discussion rounds generally improves the accuracy, but the performance gains plateau after 3 rounds. This suggests that 2-4 discussion rounds provide a good trade-off between performance and efficiency.

Ablation Study The ablation study results presented in Table 3 demonstrate the effectiveness of each component in our proposed SEDM framework. By comparing the performance of the full SEDM framework with its variants, we can gain insights into the contributions of the expert role assignment and the expert discussion components.

When the expert role assignment is removed, the performance of both GPT-3.5 and GPT-4 drops sig-

	GPT-3.5	GPT-4
w/o Expert role	11 / 100	42 / 100
w/o Expert discussion	17 / 100	55 / 100
Full SEDM*	**45** / 100	**87** / 100

Table 3: Ablation study results showing the effectiveness of each component in our proposed SEDM framework. We report the number of correctly answered questions out of 100 test samples. "w/o" denotes the removal of the corresponding component from the full SEDM framework.

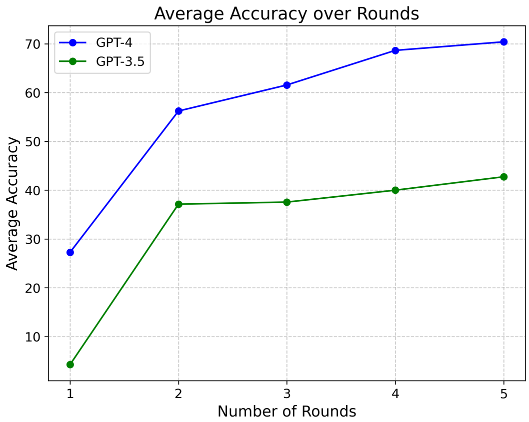

Figure 5: The relationship between the number of discussion rounds and the average accuracy of GPT-3.5 and GPT-4.

nificantly. GPT-3.5 achieves only 11 out of 100 correctly answered questions, while GPT-4 manages to answer 42 out of 100 questions correctly. This substantial decrease in performance highlights the importance of assigning domain-specific expert roles to the LLMs, as it enables them to provide more accurate and reliable responses to scientific inquiries.

Similarly, the removal of the expert discussion component also leads to a notable decline in performance. GPT-3.5 correctly answers 17 out of 100 questions, and GPT-4 achieves 55 out of 100 correct answers. This finding suggests that the collaborative discussion among the LLMs plays a crucial role in reaching a consensus and improving the overall accuracy of the system.

The ablation study provides strong evidence for the effectiveness of our proposed SEDM framework. By assigning domain-specific expert roles to LLMs and facilitating collaborative discussions among them, we can significantly enhance their performance in addressing complex scientific questions. This finding underscores the potential of our approach to advance the application of AI in scientific research, offering a more reliable and trustworthy solution for tackling scientific problems.

5 Conclusion

In this paper, we have introduced a novel approach called Simulating Expert Discussions with Multi-agent (SEDM) to enhance the scientific problem-solving capabilities of LLMs. By assigning domain-specific expert roles to multiple LLMs and simulating a panel discussion, our method leverages the collective knowledge and diverse perspectives of these models to tackle complex scientific problems effectively.

The proposed SEDM framework represents a significant step forward in harnessing the potential of LLMs for scientific problem-solving. By simulating expert discussions and leveraging the collective intelligence of multiple models, we can enhance the accuracy, reliability, and robustness of LLM-generated solutions. This approach opens up new avenues for applying artificial intelligence in scientific research, enabling more effective and trustworthy problem-solving.

6 Limitations

While the Simulating Expert Discussions with Multi-agent (SEDM) approach has shown promise in enhancing the scientific problem-solving capabilities of LLMs, several limitations warrant further investigation.

Firstly, the current study is limited to a subset of scientific domains, namely physics, chemistry, and mathematics. Future research should explore the generalizability of SEDM to a broader range of disciplines to assess its adaptability and effectiveness across diverse problem types and domain-specific challenges.

Secondly, the current implementation of SEDM employs fixed LLMs assuming expert roles within a specific domain. Although effective, this approach may not fully capture the complexity of real-world scientific collaborations. Future work could investigate more dynamic role assignment strategies, allowing for the inclusion of interdisciplinary experts to enrich discussions.

Acknowledgments

The authors thank all the anonymous reviewers for their valuable comments and constructive feedback. The authors acknowledge financial support from the National Social Science Fund of China (No. 23BTQ102). Xiaoqiu Le is the corresponding author.

References

Angus Addlesee. 2024. Grounding LLMs to in-prompt instructions: Reducing hallucinations caused by static pre-training knowledge. In *Proceedings of Safety4ConvAI: The Third Workshop on Safety for Conversational AI @ LREC-COLING 2024*, pages 1–7, Torino, Italia. ELRA and ICCL.

Microsoft Research AI4Science and Microsoft Azure Quantum. 2023. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *Preprint*, arXiv:2311.07361.

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.

Abhidip Bhattacharyya, Martha Palmer, and Christoffer Heckman. 2024. ReCAP: Semantic role enhanced caption generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13633–13649, Torino, Italia. ELRA and ICCL.

Daniil Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624:570–578.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Olena Burda-Lassen. 2023. Machine translation of folktales: small-data-driven and LLM-based approaches. In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 68–71, Gothenburg, Sweden. Association for Computational Linguistics.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. TheoremQA: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901, Singapore. Association for Computational Linguistics.

Qingxiu Dong, Li Dong, Ke Xu, Guangyan Zhou, Yaru Hao, Zhifang Sui, and Furu Wei. 2023. Large language model for science: A study on p vs. np. *Preprint*, arXiv:2309.05689.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate.

Blanca Figueras, Irene Baucells, and Tommaso Caselli. 2023. Dynamic stance: Modeling discussions by labeling the interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6503–6515, Singapore. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. *Preprint*, arXiv:2210.00720.

Weihong Guan, Shi Feng, Daling Wang, Faliang Huang, Yifei Zhang, and Yuan Cui. 2024. Improving role-oriented dialogue summarization with interaction-aware contrastive learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8913–8924, Torino, Italia. ELRA and ICCL.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Philippe Laban, Wojciech Kryscinski, Divyansh Agarwal, Alexander Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. SummEdits: Measuring LLM ability at factual reasoning through the lens of summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9662–9676, Singapore. Association for Computational Linguistics.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung-yi Lee, and Shao-Hua Sun. 2024. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *arXiv preprint arXiv:2405.06373*.

Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, Aixin Sun, Hany Awadalla, and Weizhu Chen. 2024. Sciagent: Tool-augmented language models for scientific reasoning. *Preprint*, arXiv:2402.11451.

Nikhil Mehta, Milagro Teruel, Xin Deng, Sergio Figueroa Sanz, Ahmed Awadallah, and Julia Kiseleva. 2024. Improving grounded language understanding in a collaborative environment by interacting with agents through help feedback. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1306–1321, St. Julian's, Malta. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers. *Preprint*, arXiv:2311.05965.

Philipp Sadler, Sherzod Hakimov, and David Schlangen. 2024. Sharing the cost of success: A game for evaluating and learning collaborative multi-agent instruction giving and following policies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14770–14783, Torino, Italia. ELRA and ICCL.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

Peiyang Song, Kaiyu Yang, and Anima Anandkumar. 2024. Towards large language models as copilots for theorem proving in lean. *Preprint*, arXiv:2404.12534.

Ben Swanson, Kory Mathewson, Ben Pietrzak, Sherol Chen, and Monica Dinalescu. 2021. Story centaur: Large language model few shot learning as a creative writing tool. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 244–256, Online. Association for Computational Linguistics.

Yufei Tao, Ameeta Agrawal, Judit Dombi, Tetyana Sydorenko, and Jung In Lee. 2024. ChatGPT role-play dataset: Analysis of user motives and model naturalness. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3133–3145, Torino, Italia. ELRA and ICCL.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024. Scibench: Evaluating college-level scientific problem-solving abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7572–7590, Singapore. Association for Computational Linguistics.

Fangkai Yang, Pu Zhao, Zezhong Wang, Lu Wang, Bo Qiao, Jue Zhang, Mohit Garg, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. 2023. Empower large language model to perform better on industrial domain-specific question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 294–312, Singapore. Association for Computational Linguistics.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. Large language models for automated open-domain scientific hypotheses discovery. *Preprint*, arXiv:2309.02726.

Hattie Zhou, Azade Nova, Hugo Larochelle, Aaron Courville, Behnam Neyshabur, and Hanie Sedghi. 2022. Teaching algorithmic reasoning via in-context learning. *Preprint*, arXiv:2211.09066.

A Experiment Prompts

The prompts employed in this work for the Direct querying, few-shot learning, and Simulating Expert Discussions with Multi-agent (SEDM) are illustrated in Figures 6, 7, and 8.

B Evaluation Criteria and Prompts for Solution Quality Assessment

The prompts used for LLM evaluation and the guidelines provided for human evaluation of solution quality are presented in Figure 9 and Table 4 respectively. These criteria focus on assessing the correctness of reasoning steps, clarity of explanations, and appropriate use of mathematical notations.

Please provide answers to the following problem.

Question: [scientific problem]

Please reiterate the pure numerical answer at the end of the answer.

Figure 6: The prompt for Direct querying.

Criterion	**1 (Poor)**	**2 (Fair)**	**3 (Good)**	**4 (Very Good)**	**5 (Excellent)**
Correctness of Reasoning Steps	Most steps are incorrect or missing	Several major errors in reasoning	Minor errors in reasoning, but overall approach is correct	Reasoning is correct with very minor oversights	All reasoning steps are perfectly correct and complete
Clarity of Explanations	Explanations are confusing or absent	Explanations are unclear and difficult to follow	Explanations are mostly clear but some points are ambiguous	Explanations are clear with minor areas for improvement	Explanations are exceptionally clear, concise, and easy to understand
Appropriateness of Mathematical Notations and Symbols	Incorrect or missing notations and symbols throughout	Several major errors in notation and symbol usage	Minor errors in notation and symbol usage, but generally appropriate	Notations and symbols are correct with very minor inconsistencies	All mathematical notations and symbols are perfectly appropriate and consistently used

Table 4: Human Evaluation Guidelines for Scientific Problem Solutions.

Please provide answers to the following problem.

Question: [scientific problem]

Please reiterate the pure numerical answer at the end of the answer.

Task Example 1:
Question: The logistic model has been applied to the natural growth of the halibut population in certain areas of the Pacific Ocean. [12] Let y, measured in kilograms, be the total mass, or biomass, of the halibut population at time t. The parameters in the logistic equation are estimated to have the values $r = 0.71/$ year and $K = 80.5 \times 10^6$ kg. If the initial biomass is $y_0 = 0.25K$, find the biomass 2 years later.

Solution: It is convenient to scale the solution (11)

$$y = \frac{y_0 K}{y_0 + (K - y_0)\, e^{-rt}}$$

to the carrying capacity K; thus we write Eq. (11) in the form

$$\frac{y}{K} = \frac{y_0/K}{(y_0/K) + [1 - (y_0/K)]\, e^{-rt}}$$

Using the data given in the problem, we find that

$$\frac{y(2)}{K} = \frac{0.25}{0.25 + 0.75e^{-1.42}} \cong 0.5797.$$

Consequently, $y(2) \cong 46.7 \times 10^6$ kg.

Task Example 2:
Question: Find the bonding and antibonding Hückel molecular orbitals for ethene.

Solution: The equations for c_1 and c_2 associated with Equation

$$\begin{vmatrix} H_{11} - ES_{11} & H_{12} - ES_{12} \\ H_{12} - ES_{12} & H_{22} - ES_{22} \end{vmatrix} = 0$$

are

$$c_1(\alpha - E) + c_2\beta = 0 \quad \text{and} \quad c_1\beta + c_2(\alpha - E) = 0$$

For $E = \alpha + \beta$, either equation yields $c_1 = c_2$. Thus,

$$\psi_{\mathrm{b}} = c_1\left(2p_{z1} + 2p_{z2}\right)$$

The value of c_1 can be found by requiring that the wave function be normalized. The normalization condition on ψ_π gives $c_1^2(1 + 2S + 1) = 1$. Using the Hückel assumption that $S = 0$, we find that $c_1 = 1/\sqrt{2}$.
Substituting $E = \alpha - \beta$ into either of the equations for c_1 and c_2 yields $c_1 = -c_2$, or

$$\psi_{\mathrm{a}} = c_1\left(2p_{z1} - 2p_{z2}\right)$$

The normalization condition gives $c^2(1 - 2S + 1) = 1$, or $c_1 = 1/\sqrt{2}$.

Figure 7: The prompt for few-shot learning.

You are an expert in Physical Chemistry. Please provide answers to the following problem. Your response should be accurate, high-quality, and expertly written.

Question: [scientific problem]

Please reiterate the pure numerical answer at the end of the answer.

These are the opinions from other experts:
One expert response:[One expert response]
One expert response:[One expert response]
...

Please carefully consider the opinions of other experts, but do not blindly believe them. Consider whether you agree with this insight, how it affects your answer, and provide an updated answer.

Figure 8: The prompt for Simulating Expert Discussions with Multi-agent.

Please evaluate the quality of the following solution to the given scientific problem, on a scale of 1-5 (with 5 being the highest) for each of these criteria:
1. Correctness of the reasoning steps
2. Clarity of the explanations
3. Appropriateness of the mathematical notation and symbols used
Provide a score from 1-5 for each criterion, along with a brief justification for each score.

Scientific problem: [scientific problem]

Solution: [solution]

Figure 9: The prompt for LLM evaluation of solution quality.