Overview of Changes:

1. **We have done additional experiments on the Creative Writing task and updated the experimental results in Figure 2.**
2. **We have presented additional experiment data on the ablation studies of the training convergence trend in Section 4.4. We also did additional analysis clarifying and discussing improvements amounts from line 474-483.**
3. **We have presented additional case studies on the generated graph on the Creative Writing task in Appendix E to demonstrate the interaction and propagation pattern differences on different tasks.**
4. **We have presented additional experiment data on the ablation studies of the evolution of connection scores during training in Appendix F.**
5. **Corrected several typos and formatting errors.**

Reviewer SnKE:

1. The key hypothesis proposed in the paper is not empirically validated. Section 3.2 hypothesizes that interaction quality and the best connection order among profiled agents would differ across tasks, but the experiments do not show any task-specific differences in optimal connection order or information propagation patterns, nor provide supporting analysis or case studies for this hypothesis.

Author's Response: We have provided case studies and examples to demonstrate the interaction and propagation patterns in the math reasoning tasks in Figure 4 and Appendix E. We present another case study on the task of Creative Writing to demonstrate interaction patterns and present that in Appendix E.

2. The experimental coverage of models is limited. The main experiments are conducted only on closed-source models (GPT-3.5-turbo and GPT-4o), lacking systematic validation on open-source models of different sizes (such as 7B, 32B) and types (reasoning-oriented or non-reasoning), which is insufficient to prove the generalizability and effectiveness of the method.

Author's Response: We have tested our framework on other model families, such as LLaMa 3.1-70B, and reported our performance in Table 2. For smaller model sizes, such as Llama-7B, they exhibit poor performance as a base single-agent on complex reasoning tasks (for example, on AdvGSM-M1, Llama 3.1-7B achieves an accuracy of 5 percent using direct prompting). That's why we build the framework on large language models. This is

also a typical setting to explore large language models on complex reasoning tasks. Several works [1, 2, 3] have explored multi-agent LLM reasoning using only large language models.
For the reasoning and non-reasoning models, we benchmark our main results against all baselines on GPT-3.5 to have a fair comparison, as the baselines all used GPT-series models. In addition, we also provided results on GPT-4o as a more advanced model. For the newer ones, such as GPT-o1, they are significantly more costly, so we were unable to run the experiments on those models. However, our framework is directly transferable to more advanced and costly models like OpenAI's o1. We believe that we have provided a comprehensive and fair comparison with all the baselines on the same model.

3. The distinction between agent types is not sufficiently clear. In the experiments, both profiled agents and meta agents use the same backbone LLMs, making it difficult to highlight their different roles in the reasoning process, and there is no experimental analysis of how different agent structures affect overall performance, which reduces the persuasiveness of the framework design.

Author's Response: We have tested our framework using mixture of agents as backbones. For example, in Table 2, we showed that using using a mixture of LLaMa 3.1 and GPT 3.5 as backbone perform worse than using three same GPT models with different profiling as backbone. Based on our findings, our conclusion is that the most effective way to build a multi-agent system is to use the same superior model with multiple copies, but initialized with different profiles.

We also demonstrated the effectiveness of profiling in Table 3, where we compare the performance of the variation with no profiling to other variations and baseline methods. We showed the details of different profiles in the ablation studies. Overall, using profiling greatly enhances framework performance by an average of 3 points across datasets.

4. The rationale and validation for parameter settings are lacking. While Section 4.1 specifies the generation parameters, it does not explain the basis for these choices or provide ablation or sensitivity experiments, which impacts the reproducibility of the results and the applicability of the method in different settings.

Author's Response: To ensure a fair comparison and consistency with all the baselines, we directly used the default hyperparameter settings for each model. This is a common practice in the LLM agent work [1,2,3].

[1] https://arxiv.org/pdf/2305.19118 [2] https://arxiv.org/pdf/2309.13007 [3] https://arxiv.org/pdf/2402.18272

Reviewer yDsh:

1. The paper hypothesizes that debating quality plays an important role in multi-agent systems (line 15-16 and line 87-88). This hypothesis should be highlighted and supported in the experiments or baseline comparisons.

Author's Response: To support our key claim, we demonstrate the effectiveness of considering debating quality by comparing our framework with other baseline methods that do not account for this trait, and our framework outperforms the baselines. We have also conducted an ablation study using our framework, excluding debate quality from the training time. The results are available in Table 1 of Section 4. Compared with not considering debating quality ("Train Without Interaction Quality" row in Table 1), our full framework performs better at all datasets by an average of 0.5 points. We have made necessary highlights in Table 1 of Section 4 and Section 4.4 to make it more clear.

2. It is suggested to use case studies or additional experiments to demonstrate and discuss whether the agents' interaction quality improves during the optimization process.

Author's Response: We would like to clarify some common misunderstandings. We are not enhancing the pairwise agent interaction quality, but instead optimizing the search and pruning of selected interaction pairs to improve the overall framework's interaction quality. We have provided the stepwise performance enhancement in Figure 3. We present another ablation study on the evolution of the connection scores of the top-5 connection scores in each round on the 5-agent scenario on the GSM-Plus dataset in Appendix F.

3. In Figure 3, compared to the performance improvement on easy datasets like GSM8K (84 -> 85 -> 87), the performance gain on the more challenging MATH dataset is minimal as the epochs increase (33 -> 34 -> 34). Intuitively, a multi-agent system should help integrate the strengths of different agents to tackle more complex problems, but the experiments show that OPTAGENT yields more significant gains on simpler problems like GSM8K. The authors should discuss this observation.

Author's Response: We did additional analysis clarifying and discussing improvements amounts from line 474-483. We believe that on harder datasets, the agents have difficulties forming high-quality answers and interactions. We observe that on MATH and AdvGSM-M1, there are many cases where none of the agents give out the correct answer, which prohibits correct answer propagation when there's no correct answer present. Some other research works on Multi-Agent LLM frameworks [1,2,3] also exhibit this phenomenon, where the improvement amount, when compared with simple debating, in GSM-8K is higher than that of MATH.

4. It is recommended to add more epochs in Figure 3 to observe the performance upper bound of OPTAGENT on GSM8K, and to check whether the model's performance continues to improve with additional epochs.

Author's Response: We have presented additional experiment data on the ablation studies of the training convergence trend in Section 4.4.

[1] https://arxiv.org/pdf/2412.01928v2 [2] https://arxiv.org/pdf/2311.17371 [3] https://arxiv.org/pdf/2402.05120

Reviewer a1PL:

1. The constructed graph may vary depending on the domain or task, yet the experiments do not clearly explain the unit of training, whether the framework is trained separately on each domain-specific dataset or a single trained model is reused across tasks.

Author's Response: We have made changes in Section 4.1, in line 380-381, to highlight that our framework was trained separately on each dataset. Our framework consists of multiple agents instead of a single trained model. During the training process, we optimize the search and pruning of interaction pairs to improve the overall framework's interaction quality. We have made necessary changes to highlight the relevant sections.

2. Using only 3 data points for training the framework is insufficient.

Author's Response: For the number of training data points, we reported the evolution of performance as we added training data points in Figure 3. We chose to use 3 data points for training across all baseline methods that required training, and presented additional experiment data on the ablation studies of the training convergence trend in Section 4.4.

3. The explanation of the creative writing and sorting tasks is lacking. In particular, for creative writing, the use of only 10 data points raises concerns about the significance and generalizability of the results.

Author's Response: We present the experiment results on Creative Writing using 100 points on OPTAGENT in Figure 2. Our 5-agent framework consistently outperforms baseline methods over the larger testing set on Creative Writing.

4. The performance on sorting and ARC tasks is lower than that of the 0-shot CoT baseline, which raises doubts about the overall effectiveness of the proposed method. The method appears to show strong results only for math reasoning tasks.

Author's Response: We would like to clarify some misunderstandings. In the sorting task, we adopt the settings from the GoT paper [1]. As explained in section 4.2, lower error numbers represent better performance; we will hightlight this in the revised version of our paper. Our framework outperforms 0-Shot CoT and simple debating methods in the 16-number and 32-number scenarios. The improvement amount is most obvious in the 32-number scenario, where we have a performance gap of 0.9 errors per case. For ARC, we acknowledged that multi-agent framework do not bring benefit, and we discussed this in Section 4.6. In detail, we believe that ARC is leaning towards more direct knowledge recall rather than reasoning; the model's knowledge base and understanding of these questions are more important than the logical reasoning process. We did the experiments on ARC for a more comprehensive picture and insights to share with the research community. Our framework outperforms baseline methods at GPQA, which is also a science reasoning dataset and is significantly more challenging.

5. It is necessary to report how the connection scores between agents evolve over each data point to understand the dynamics of collaboration better.

Author's Response: We report the evolution of the connection scores of the top-5 connection scores in each round on the 5-agent scenario on the GSM-Plus dataset in Appendix F.