npj | artificial intelligence

Article

https://doi.org/10.1038/s44387-025-00045-3

Self-reflection enhances large language models towards substantial academic response

Check for updates

Baoxue Li & Chunhui Zhao ✉

Crafting response letters to reviewers' comments is a time-consuming yet critical part of academic peer review. The inexperience of researchers can hinder the publication of their work, exacerbating the Matthew effect in science. To address this, we design a large language model (LLM)-assisted writing framework. However, LLMs often output responses that are polished in structure and style but fail to address the core of the comment. Inspired by metacognition, we propose a dual-loop reflection method. First, the LLM critiques its own reasoning process against human reference responses (extrospection). The reflections gained from this process build a reflection bank. This bank is then retrieved during the reasoning process to facilitate introspection, allowing the LLM to overcome previous errors. The reflection bank was constructed using 4000 papers and 79,000 comments from Nature group journals. Validation on over 3700 comments from 200 papers demonstrates our method's effectiveness and superiority.

With the establishment of the first two academic journals, *Journal des sçavans* and *Philosophical Transactions of the Royal Society of London*, in 1665, the practice of public scrutiny and critique was introduced, although it did not fully resemble the peer review system we know today [1]. Over time, journals gradually adopted more rigorous review processes, eventually formalizing the peer review system. As part of the peer review process, reviewers provide valuable feedback on a manuscript's strengths and weaknesses, offering suggestions for improvement. In this context, the response letter serves as a key communication tool in this context, allowing authors to thoughtfully address the reviewer comments, communicate revisions, and demonstrate their professionalism. By providing clear explanations of how reviewer feedback has been incorporated, or why certain suggestions were not followed, the response letter increases the likelihood of acceptance and strengthens the academic value of the paper.

Although researchers recognize the importance of response letters, crafting a satisfactory one is far from a simple task and faces several challenges. First, writing response letters is time-consuming. According to a study[2], in 2020, a total of 21,800,126 reviewers provided feedback on academic papers globally. If it takes an average of 20 hours to respond to each reviewer's comments, researchers worldwide would need to spend approximately 436 million hours, or about 50,000 years, drafting response letters. Second, researchers in resource-constrained regions often lack access to guidance on how to write effective response letters. This disparity can contribute to an imbalance in the visibility and publication of research between marginalized and elite institutions, thereby exacerbating the Matthew effect in science[3]. This term describes how established scientists tend to accrue greater reputation and resources, while lesser-known researchers receive less recognition for comparable outcomes. In the context of our study, this effect is relevant, as inexperience in crafting convincing response letters may become a key obstacle, undermining the visibility of valid research from marginalized communities. Third, despite the availability of many open-source peer review documents, their utilization remains insufficient. Several journals within the *Nature* group have adopted a transparent peer review policy[4], allowing authors to publish reviewers' comments alongside accepted papers. By 2021, approximately 46% of Nature authors had chosen to make these reviewer discussions public[5]. Besides, the *OpenReview* platform (https://openreview.net/) has provided a flexible system for open peer review and has become widely adopted by computer science conferences. These resources not only enhance the transparency of the review process but also provide valuable learning opportunities through authors' response letters. However, with the sheer volume of papers being published across various fields, the vast number of peer review documents can be overwhelming, and extracting meaningful insights from them becomes a daunting task. Moreover, researchers often seek specific strategies for addressing particular types of reviewer comments. Given the abundance of response letters, finding relevant examples or addressing similar issues can be akin to searching for a needle in a haystack, with the effort involved often outweighing the potential benefits.

The State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou, China.
✉e-mail: chhzhao@zju.edu.cn

Recently, large language models (LLMs) have gained significant attention for their impressive capabilities in natural language understanding, generation, and reasoning[6,7]. They have become increasingly influential and may eventually become a cornerstone in a wide range of fields, including industry[8], healthcare[9], finance[10], education[11], VQA[12–14], and beyond[15,16]. In fact, LLMs have been gradually integrated into the writing workflow of researchers and are having a quiet influence on the research community[17]. For instance, the *Nature* Career Column article, titled "Three ways ChatGPT helps me in my academic writing", highlights the valuable applications of LLMs in academic writing[18]. Specifically, LLMs can assist in academic writing by polishing drafts for clarity and coherence, elevating peer review by organizing and articulating feedback, and optimizing editorial feedback through precise, actionable, and empathetic communication. In addition, scholars have explored the use of LLMs for paper quality assessment to address the issue of limited reviewer resources[19]. Through experiments, they have demonstrated that LLMs hold promise as assistants for specific review tasks, such as error detection and checklist verification[20]. Accordingly, LLMs offer new possibilities for response letter instruction and writing. LLM-assisted writing of response letters will significantly shorten the drafting cycle, thereby freeing up researchers' time for core tasks like experimentation. Furthermore, the vast amount of publicly available resources has the potential to be harnessed by LLMs, allowing the knowledge and experience contained within these resources to reach even peripheral researchers.

Despite the potential of LLMs to automate the writing of response letters, there are still significant challenges in ensuring that LLMs produce high-quality response letters. Writing a good response to a comment involves complex planning and reasoning[21], and many aspects need to be accomplished, such as understanding the reviewer's intent, deciding to accept the comment or rebut it, providing validation, ensuring clarity, and maintaining respect. In recent developments, the Chain-of-Thought (CoT) prompting has been introduced as a significant advancement of LLMs[22–24]. The central idea behind CoT is to guide LLMs to generate intermediate reasoning steps that mimic human-like thought processes, which enables LLMs to achieve impressive results in various tasks requiring logical reasoning. Although the CoT technique allows LLMs to break down and solve problems in chains, LLMs may output responses that are polished in structure and style but fail to address the core of the problem[25,26], which we denote the *shallow reasoning problem*, illustrated in Fig. 1a. Therefore, we point out that the LLM-assisted response letter writing is not a trivial task, but needs to guide the LLM to overcome the shallow reasoning problem, and really solve the reviewer's concerns as experienced human beings do. To realize this, an intuitive idea is to leverage an open-source peer review corpus to modify the LLMs using parameter-efficient fine-tuning (PEFT) methods[27–29], which aims to teach the LLMs the patterns and strategies humans typically use when crafting response letters. The flaws in this idea are twofold. First, it is hard to guarantee whether the LLMs learn the human routine or memorize it by rote, and also the LLM's ability in other tasks may be interfered with. Second, with the large and increasing amount of publicly available peer review material, it is not economical to modify the parameters of the LLMs.

In this paper, we propose a reflection bank-based LLM framework (RBB-LLM) to realize AI-assisted response letter writing and overcome the *shallow reasoning problem*. The core of our approach is a dual-loop reflection mechanism that involves both extrospection and introspection inspired by the metacognition in cognitive psychology[30–32]. Different from the existing reflection methods that focus solely on introspecting about the output[33,34], we first make the LLM detach from itself to act as an observer and critique its own reasoning process with human reference responses, i.e., extrospection. Reflections obtained from extrospection construct the reflection bank. Unlike just using good examples[35], it stores actionable insights about common failure points of LLMs. When online reasoning, these reflections obtained from extrospection facilitate the LLM to conduct introspection and overcome the shallow reasoning problem. In this article, the proposed reflection bank was constructed based on 4000 papers and peer review documents (approximately 79,000 comments) from *Nature Communications*, covering the fields of physical sciences, earth and environmental, biological sciences, health sciences, scientific community and society. Validation results on more than 3700 comments of 200 papers demonstrate the superiority of reflection and the effectiveness of LLM-assisted response letter writing. Despite its promising performance, we must emphasize that the framework is intended to support rather than replace human authors, and researchers still bear full responsibility for the responses.

Results

Here, we start by presenting peer review file collection results and summarizing the statistics of them. We then demonstrate the effectiveness of the proposed framework across various comment types and article domains, supported by quantitative results and two illustrative examples. Finally, we provide a detailed discussion on the impact of reflection bank size and the use of different LLMs.

Peer review file statistics

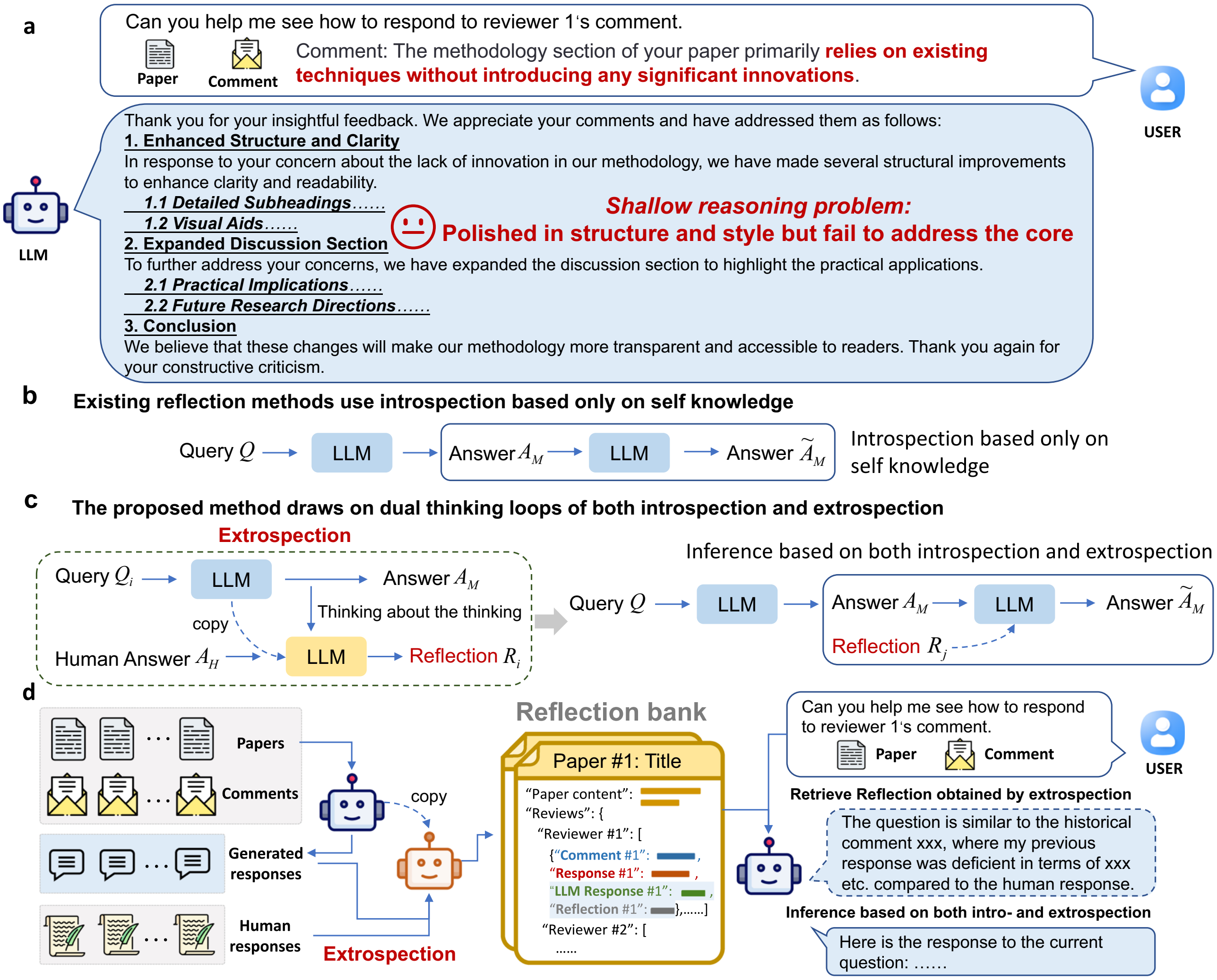
Our peer-reviewed documents were collected from papers published from January 1, 2024 to September 25, 2024 in *Nature Communications*, which were sourced directly from the website (https://www.nature.com/ncomms/). Articles without available peer-review documents or with non-selectable text were excluded, resulting in a total of 4259 articles. Among these, 4059 articles were utilized to construct the reflection bank, while 200 were randomly selected for framework validation. The 4059 articles were classified according to the *Nature Communications* taxonomy, with 1540 articles categorized as physical sciences, 204 as earth and environmental sciences, 1577 as biological sciences, 652 as health sciences, and 86 as scientific community and society. The test set of 200 articles was proportionally representative of these categories, ensuring consistency with the training set. Reviewer comments were categorized into six groups based on the guide[36]: grammar and language issues, methodological critique, novelty and originality, theoretical and conceptual critique, interpretation and conclusions, and citations and related work. Detailed statistics on the number of comments in each category are illustrated in Fig. 2.

Evaluation on validation articles and comments

In this part, we first present the quantitative performance of the proposed method compared to baseline methods across various article types and comment categories. Next, we provide two specific examples of comments and responses to analyze the characteristics of different methods. Furthermore, we evaluate the generated responses using different LLMs to showcase the fairness and consistency of the evaluation criteria. Finally, we demonstrate the generalizability of the proposed dual-loop reflection mechanism by employing different LLMs as the base model.

The used LLMs include GLM-4-Flash, DeepSeek-v2.5, and Qwen-Plus. As an efficient variant of the GLM-4 series developed by Zhipu AI, this model is widely recognized for its balance between inference speed and performance[37]. DeepSeek-v2.5[38] is a state-of-the-art open-source MoE (Mixture-of-Experts) model developed by DeepSeek-AI, which achieves top-tier performance across benchmarks. Qwen-Plus[39] is a proprietary flagship model from Qwen series, excelling in multilingual understanding (29+ languages) and long-context processing. In summary, the selected LLMs are popular and have differences in their capabilities, which can validate the generalization of the proposed method to different models.

Quantitative performance across various article types and comment categories. We compare the proposed RBB-LLM framework with four LLM reasoning approaches, including direct prompting (DP), CoT[22], self-refine (SR), and answer-based reasoning (ABR). The first three methods are general-purpose and popular prompting strategies, which we adapt to the task of generating responses to reviewer comments. The fourth method is a degraded variant of the proposed RBB, which is used to demonstrate the impact of removing the reflection. The five methods


a
Paper
Comment
USER
LLM
Shallow reasoning problem:
Polished in structure and style but fail to address the core
b
Existing reflection methods use introspection based only on self knowledge
Query Q
LLM
Answer A_M
LLM
Answer $\widetilde{A}_M$
Introspection based only on self knowledge
c
The proposed method draws on dual thinking loops of both introspection and extrospection
Extrospection
Query Q_i
LLM
Answer A_M
Thinking about the thinking
copy
Human Answer A_H
LLM
Reflection R_i
Inference based on both introspection and extrospection
Query Q
LLM
Answer A_M
LLM
Answer $\widetilde{A}_M$
Reflection R_j
d
Papers
Comments
Generated responses
Human responses
copy
Extrospection
Reflection bank
Paper #1: Title
Paper
Comment
USER
Retrieve Reflection obtained by extrospection
Inference based on both intro- and extrospection

Can you help me see how to respond to reviewer 1's comment.

Comment: The methodology section of your paper primarily **relies on existing techniques without introducing any significant innovations**.

Thank you for your insightful feedback. We appreciate your comments and have addressed them as follows:

1. Enhanced Structure and Clarity

In response to your concern about the lack of innovation in our methodology, we have made several structural improvements to enhance clarity and readability.

1.1 Detailed Subheadings……

1.2 Visual Aids……

2. Expanded Discussion Section

To further address your concerns, we have expanded the discussion section to highlight the practical applications.

2.1 Practical Implications……

2.2 Future Research Directions……

3. Conclusion

We believe that these changes will make our methodology more transparent and accessible to readers. Thank you again for your constructive criticism.

```
"Paper content":
"Reviews": {
  "Reviewer #1": [
    {"Comment #1": ,
    "Response #1": ,
    "LLM Response #1": ,
    "Reflection #1": },......]
  "Reviewer #2": [
    ......
```

Can you help me see how to respond to reviewer 1's comment.

The question is similar to the historical comment xxx, where my previous response was deficient in terms of xxx etc. compared to the human response.

Here is the response to the current question: ……

Fig. 1 | Overview of our motivations and contributions. a The figure depicts the *shallow reasoning problem* of the LLM in response writing. When LLM is asked to respond to the reviewer's comment, it might output a polished response that seems logical but does not actually answer the core concern. **b** illustrates the flow of existing reflection methods, where they introspect only through self-knowledge and focus on the output A_M rather than the entire reasoning process. **c** illustrates the proposed idea of dual-loop reflection that includes extrospection and introspection. The LLM is first instructed to detach from itself and critique its own reasoning process with human reference responses, i.e., extrospection. When reasoning, the LLM rewrites the initial response based on the reflection of the extrospection, i.e., introspection. **d** Illustrates the flow of the proposed method. First, we utilize publicly available peer review documents to instruct the LLM to generate responses and critique its own reasoning process with human reference responses, i.e., extrospection. When online reasoning, these reflections derived from extrospection will be retrieved, guiding the LLM to introspect on past shortcomings when solving similar comments at hand.

are validated on 3700 comments on 200 articles, covering 5 typical domains and 6 comment types.

Figure 3 shows the performance comparison of methods for different article types and comment types. Table 1 presents the performance comparison of methods by averaging the scores across different comment types for the same article types, providing an overall view of how each method performs on specific article categories without considering the variation in comment types. On the other hand, Table 2 averages the scores across different article types for the same comment types, highlighting the performance of methods based on specific comment categories. This dual perspective allows for a comprehensive understanding of the methods' effectiveness across varying contexts. Based on Fig. 3, Table 1, and Table 2, we can get the following observation.

First, the proposed method has the highest score with the lowest variance on different article types and different comment types, which demonstrates the adaptability, effectiveness, and stability of the proposed method.

Second, the partial order relationship between the different methods remains consistent across both article and comment types, with scores decreasing in the order of Ours, SR, CoT, ABR, and DP. This indicates that there is a performance gap between the different methods, and this gap is stable across various article and comment types.

Third, the CoT proves important in the academic response generation task. Compared to direct prompting, using the CoT strategy leads to score improvements of more than 3 points across both article types and comment types, as well as a reduction in variance compared to DP.

Fourth, prompting the LLM to self-refine the responses obtained from CoT is effective for academic response writing tasks. Compared to CoT, the quality of SR's responses improved by more than 1.5 points on average, accompanied by a further reduction in variance.

Fifth, the ABR method showed a counterintuitive performance degradation compared to CoT. Both ABR and SR responses were based on the initial CoT-generated responses. In contrast to SR, which does not rely on additional information to self-refine, ABR was provided with historically similar questions and human responses. This phenomenon suggests that for the LLM, the direct input of the human's reference response without reflection, where the model could fail to recognize and correct the gap between its own response and the human response, may even produce

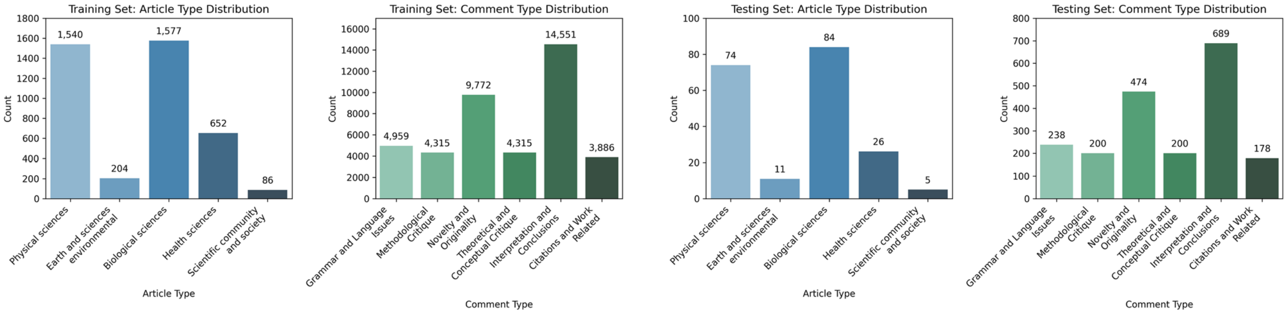


Fig. 2 | Distribution of article and comment types across the training and test sets.

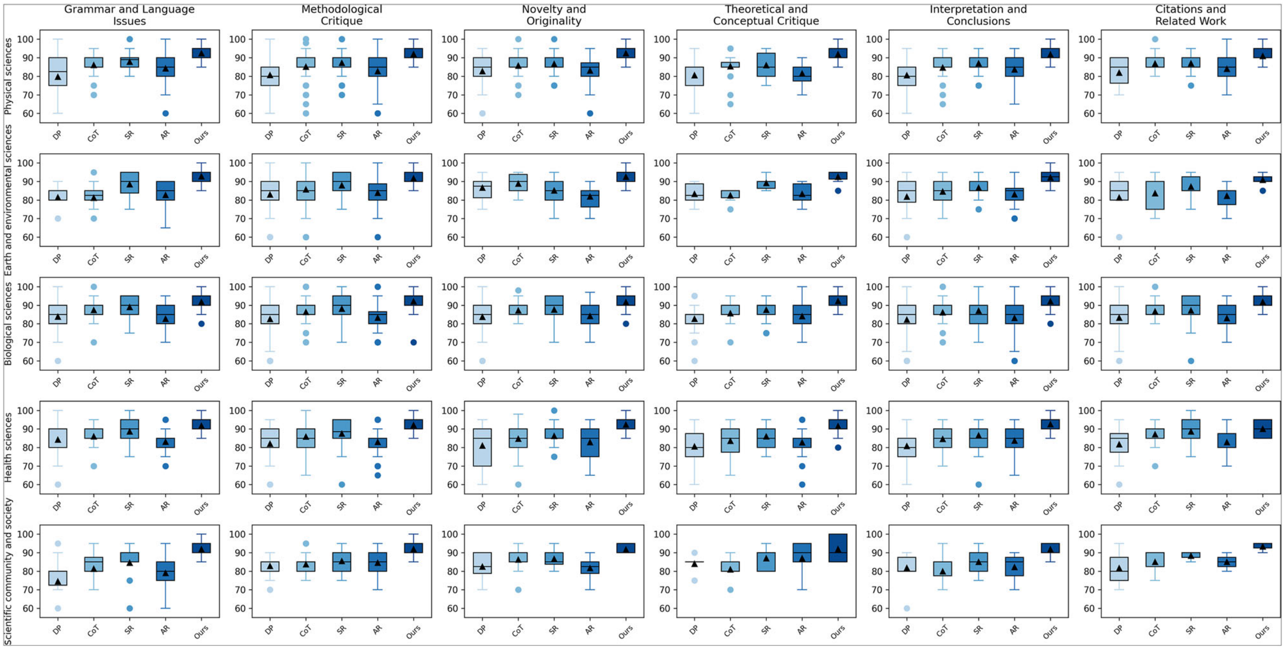


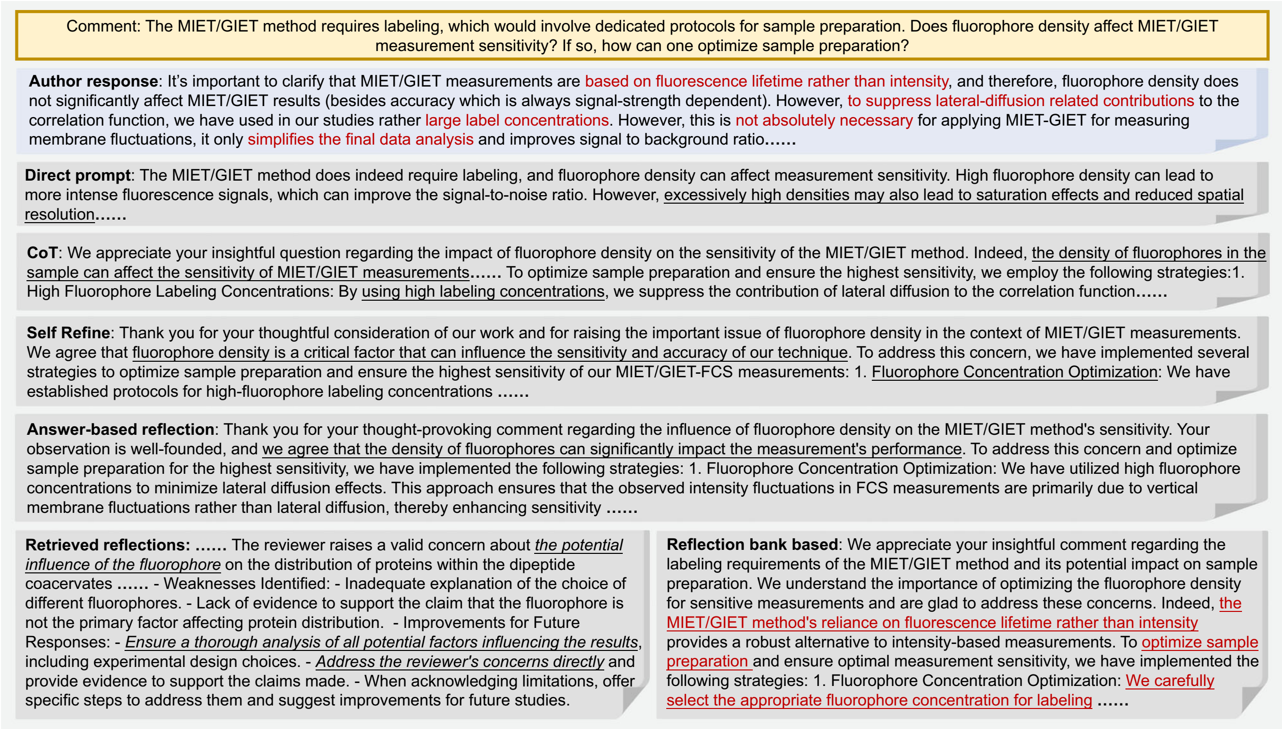
Fig. 3 | Box plots of methods for different article and comment types. The five methods in each subgraph are GLM-DP, GLM-CoT, GLM-SR, GLM-ABR, and the proposed method.

Table 1 | Performance comparison of methods on different article types

Methods	Physical sciences	Earth and environmental sciences	Biological sciences	Health sciences	Scientific community and society	Average	Δ
GLM-DP	80.99 ± 8.87	82.94 ± 8.24	82.74 ± 7.97	81.73 ± 8.26	81.43 ± 9.65	81.97	/
GLM-CoT	85.51 ± 7.46	85.19 ± 7.87	86.36 ± 7.61	85.44 ± 7.13	83.25 ± 8.76	85.15	3.18
GLM-SR	87.07 ± 6.43	87.45 ± 5.87	87.81 ± 6.27	87.18 ± 6.19	85.71 ± 6.65	87.04	5.08
GLM-ABR	83.05 ± 6.67	83.26 ± 6.62	83.35 ± 6.65	83.11 ± 6.25	82.94 ± 7.44	83.14	1.18
Ours	92.03 ± 4.03	91.95 ± 3.99	92.10 ± 4.08	92.03 ± 3.96	91.98 ± 3.97	92.02	10.05

Table 2 | Performance Comparison of Methods on Different Comment Types

Methods	Grammar and language issues	Methodological critique	Novelty and originality	Theoretical and conceptual critique	Interpretation and conclusions	Citations and related Work	Average	Δ
GLM-DP	82.00 ± 12.41	81.95 ± 8.06	83.07 ± 7.60	81.78 ± 7.50	82.61 ± 8.03	81.83 ± 9.07	82.21	/
GLM-CoT	86.27 ± 6.68	85.92 ± 7.62	86.45 ± 5.66	85.08 ± 8.42	85.48 ± 10.98	85.43 ± 7.40	85.77	3.56
GLM-SR	88.35 ± 5.87	87.72 ± 6.24	86.92 ± 6.07	86.90 ± 5.98	87.33 ± 8.91	86.88 ± 6.02	87.35	5.14
GLM-ABR	83.05 ± 8.76	83.15 ± 6.36	83.38 ± 6.74	83.23 ± 6.82	83.20 ± 6.88	83.35 ± 6.25	83.23	1.02
Ours	92.12 ± 4.06	92.06 ± 4.10	92.14 ± 4.02	92.13 ± 4.18	91.35 ± 4.10	92.13 ± 3.82	91.99	9.78



Comment: The MIET/GIET method requires labeling, which would involve dedicated protocols for sample preparation. Does fluorophore density affect MIET/GIET measurement sensitivity? If so, how can one optimize sample preparation?

Author response: It's important to clarify that MIET/GIET measurements are based on fluorescence lifetime rather than intensity, and therefore, fluorophore density does not significantly affect MIET/GIET results (besides accuracy which is always signal-strength dependent). However, to suppress lateral-diffusion related contributions to the correlation function, we have used in our studies rather large label concentrations. However, this is not absolutely necessary for applying MIET-GIET for measuring membrane fluctuations, it only simplifies the final data analysis and improves signal to background ratio……

Direct prompt: The MIET/GIET method does indeed require labeling, and fluorophore density can affect measurement sensitivity. High fluorophore density can lead to more intense fluorescence signals, which can improve the signal-to-noise ratio. However, excessively high densities may also lead to saturation effects and reduced spatial resolution……

CoT: We appreciate your insightful question regarding the impact of fluorophore density on the sensitivity of the MIET/GIET method. Indeed, the density of fluorophores in the sample can affect the sensitivity of MIET/GIET measurements…… To optimize sample preparation and ensure the highest sensitivity, we employ the following strategies:1. High Fluorophore Labeling Concentrations: By using high labeling concentrations, we suppress the contribution of lateral diffusion to the correlation function……

Self Refine: Thank you for your thoughtful consideration of our work and for raising the important issue of fluorophore density in the context of MIET/GIET measurements. We agree that fluorophore density is a critical factor that can influence the sensitivity and accuracy of our technique. To address this concern, we have implemented several strategies to optimize sample preparation and ensure the highest sensitivity of our MIET/GIET-FCS measurements: 1. Fluorophore Concentration Optimization: We have established protocols for high-fluorophore labeling concentrations ……

Answer-based reflection: Thank you for your thought-provoking comment regarding the influence of fluorophore density on the MIET/GIET method's sensitivity. Your observation is well-founded, and we agree that the density of fluorophores can significantly impact the measurement's performance. To address this concern and optimize sample preparation for the highest sensitivity, we have implemented the following strategies: 1. Fluorophore Concentration Optimization: We have utilized high fluorophore concentrations to minimize lateral diffusion effects. This approach ensures that the observed intensity fluctuations in FCS measurements are primarily due to vertical membrane fluctuations rather than lateral diffusion, thereby enhancing sensitivity ……

Retrieved reflections: …… The reviewer raises a valid concern about *the potential influence of the fluorophore* on the distribution of proteins within the dipeptide coacervates …… - Weaknesses Identified: - Inadequate explanation of the choice of different fluorophores. - Lack of evidence to support the claim that the fluorophore is not the primary factor affecting protein distribution. - Improvements for Future Responses: - *Ensure a thorough analysis of all potential factors influencing the results*, including experimental design choices. - *Address the reviewer's concerns directly* and provide evidence to support the claims made. - When acknowledging limitations, offer specific steps to address them and suggest improvements for future studies.

Reflection bank based: We appreciate your insightful comment regarding the labeling requirements of the MIET/GIET method and its potential impact on sample preparation. We understand the importance of optimizing the fluorophore density for sensitive measurements and are glad to address these concerns. Indeed, the MIET/GIET method's reliance on fluorescence lifetime rather than intensity provides a robust alternative to intensity-based measurements. To optimize sample preparation and ensure optimal measurement sensitivity, we have implemented the following strategies: 1. Fluorophore Concentration Optimization: We carefully select the appropriate fluorophore concentration for labeling ……

Fig. 4 | Responses from different methods on Case 1. Red font indicates key statements in the author's response. Black underlining indicates expressions in the response that relate to key statements. Italicized underlining indicates records in the reflection bank that are closely related to the comment. Red underlining indicates expressions in the proposed method that coincide with the key statements.

interference with the current question due to more contextual information. On the contrary, the proposed method not only provides historically similar comments and human responses, but also additionally provides the LLM's generated responses as well as the LLM's reflections on the generated responses and human responses. In this way, the LLM successfully identifies gaps from the reflection and migrates to the current question to generate higher-quality responses that better address the reviewers' concerns. The proposed method improves by 9 to 10 points compared to DP and by more than 4 points compared to the second-best SR.

Case presentation and analysis of responses from humans and different methods. In this part, responses generated by different methods for two specific comments are analyzed in detail. These two comments are from response letters of articles belonging to the biological sciences and sociology, respectively. Unlike simple grammatical problems or minor issues, they are methodological or motivational concerns. Interestingly, the comparison methods were sidetracked by the reviewers or did not answer the points that the reviewers really cared about. The proposed method, on the other hand, performed well, with a near-human level of quality and specificity.

The article of the first case presents a novel methodology combining metal/graphene-induced energy transfer (MIET/GIET) with fluorescence correlation spectroscopy (FCS) to measure out-of-plane fluctuations of biological membranes[40]. This method offers a powerful tool for studying the dynamic behavior of diverse biological membranes. Due to the incorporation of the fluorophore, the reviewers raised concerns about the effect of fluorophore density on the sensitivity of the measurements. As shown in Fig. 4, the reviewer commented, "The MIET/GIET method requires labeling, which would involve dedicated protocols for sample preparation. Does fluorophore density affect MIET/GIET measurement sensitivity? If so, how can one optimize sample preparation?" The key points in the authors' response are highlighted in red: MIET/GIET measurements rely on fluorescence lifetime rather than intensity, and high label concentrations are used to enhance the signal-to-background ratio, but this is not absolutely necessary. It can be seen that the reviewer misunderstood that the method relied on fluorophore density, and the authors first clarified that the method relies on fluorescence lifetime. By looking at the responses of DP, CoT, SR, and ABR, we can see that the responses of these methods are biased by the reviewers, affirming that fluorophore density can affect measurement sensitivity. Although this is correct, the reviewer's misunderstanding of the MIET/GIET has not been clarified. The follow-up responses of CoT, SR, and ABR talk about the setting of fluorophore density preparation in this article, which is a reply to the reviewer's question about how to "optimize sample preparation". The DP method, on the other hand, goes further and further away, and the topic shifts to saturation effects, which is not mentioned by the reviewer. For the proposed method, we first present the retrieved texts from the reflection bank based on the reviewer's comment, which is a quadruple consisting of the comment, human response, LLM response, and reflection. The article corresponding to the retrieved texts is also related to fluorophores, which are utilized as a tool for detection and tracking to study the functionality and reactions of dipeptide coacervates within cells. The comment is about the potential influence of the fluorophore on the distribution of proteins within the dipeptide coacervates. Due to space constraints, only some of the key content is shown in Fig. 4, mainly the reflection part. In reflection, the LLM summarizes its weaknesses and future enhancements by comparing the generated response with the author's response. Among them, ensuring "a thorough analysis of all potential factors influencing the results" and addressing "the reviewer's concerns directly" are also applicable to the current response writing. Prompted by the retrieved reflection, the generated response of our method begins by clarifying that "the MIET/GIET method relies on fluorescence lifetime rather than intensity", which helps to dispel the reviewer's misunderstanding. And it mentions the appropriate choice of fluorophore concentration in order to optimize sample preparation, which answers the reviewer's second question.

The article of the second case explores whether small-scale social interactions can affect large-scale economic inequalities in under-resourced contexts, using population-level data from one of the poorest settings in

Table 3 | Taxonomy and frequency of failure modes in baseline methods and our method

Method	Ignoring the core concern	Misinterpreting reviewer intent	Hallucination	Inappropriate tone	Verbose and unfocused filler
ChatGLM-DP	45.15%	25.62%	22.48%	6.51%	10.23%
ChatGLM-CoT	28.45%	22.08%	18.13%	6.03%	15.31%
ChatGLM-SR	22.79%	21.25%	15.41%	5.97%	17.55%
ChatGLM-ABR	30.58%	24.12%	21.45%	6.43%	12.35%
Ours	**16.75%**	**15.20%**	**12.55%**	**5.89%**	**8.29%**

Values represent the percentage of responses where the specified error occurs. Bold numbers indicate the best performance among all methods.

South Africa[41]. The authors leverage a context-specific measure of overlapping household memberships derived from census data to map inter-household relationships. The reviewer is questioning the logical consistency of the paper's motivation. Specifically, the paper mentions in the introduction that "Survey-based studies, which are the most commonly employed approach for such investigations, are often prone to inaccurate self-report responses and can be costly at scale." However, the reviewer points out that the authors themselves rely on data filled in by household members, which is not administratively generated but self-reported. The key points in the authors' response are highlighted in red: First, inter-household relationships derived from census data can reduce self-reporting bias to a certain extent. Second, survey-based approaches often encounter challenges in remote populations due to high costs and inadequate infrastructure. Third, authors have revised the entire paragraph for clear elaboration. By looking at the responses of DP, CoT, SR, and ABR, none of them could fully cover these three points. In particular, DP and SR erroneously believe the article's methodology is founded on administratively generated data, not self-reported data. While CoT and ABR mention that the methodology is designed with a unique census indicator to reduce self-reported bias, they do not mention the latter two key points. For the proposed method, we also present the retrieved quadruple from the reflection bank. The article corresponding to the retrieved texts assesses the relationship between prenatal socioeconomic disadvantage and changes in cortical network segregation. The comment questions the inclusion of maternal nutrition and insurance status and suggests that the term "socioeconomic status" may not fully capture the broader construct of social disadvantage. In reflection, LLM recognizes the need for a more comprehensive understanding of the theoretical framework, addressing the reviewer's concerns directly and theoretically, and offering examples of how to revise the manuscript. The reflection is derived from a specific comment, but is generic and can contribute to the comment at hand. Prompted by the retrieved reflection, the generated response points out, on the one hand, that social interaction mining based on census data can be more reliable than self-reporting and, on the other hand, that the original expression emphasizing survey-based approaches faces challenges when carried out at scale in remote and under-resourced settings. In addition, our response points to the need for modification of the Introduction to enhance the justification.

Failure mode analysis of baseline methods and our method. To provide a deeper understanding of the limitations of existing methods and to better show the advantages of our RBB-LLM framework, we conducted an error analysis on the responses generated by the baseline methods (DP, CoT, SR, and ABR) and our method. We define five primary failure modes based on our analysis of the 3700 comments in the validation set: We define five primary failure modes based on our analysis of the 3700 comments in the validation set: ignoring the core concern, where the generated response appears polished and relevant on the surface but fails to address the central point or critical question raised by the reviewer; misinterpreting reviewer intent, where the LLM misunderstands the intent of the reviewer's comment, such as mistaking a suggestion for a criticism; hallucination, where the response includes incorrect information and justifications that are not supported by the manuscript; inappropriate tone, where the response adopts a tone that is overly defensive, too informal, or otherwise unprofessional for academic correspondence, potentially harming the author-reviewer dialogue; and verbose and unfocused filler, where the response is unnecessarily long and contains redundant or irrelevant information that dilutes the core message and fails to be direct and concise.

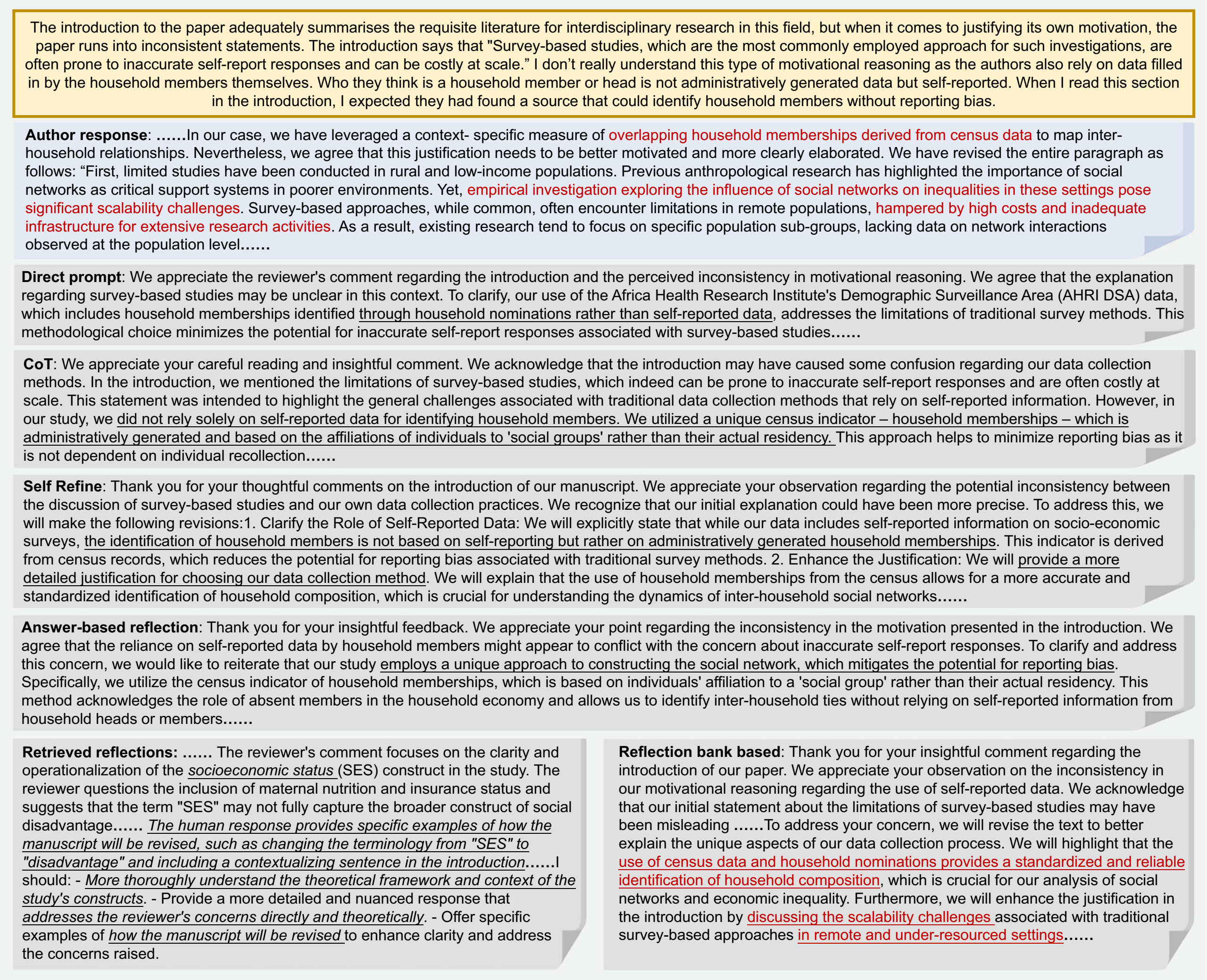
We use DeepSeek-R1 to annotate the failure modes for each response generated by different methods. Table 3 summarizes the distribution of these errors, offering a quantitative view of each method's weaknesses. It is worth noting that a response may have multiple error modes.

As shown in Table 3, DP often ignores the reviewer's core concern, confirming its tendency for shallow reasoning. While CoT and SR improve on this, they still struggle significantly with misinterpreting the reviewer's intent and verbose filler. Interestingly, the ABR method, which sees human answers without reflection, sometimes leads to an increase in hallucination as the model struggles to integrate new information without proper context. Our RBB-LLM framework mitigates these failures. The retrieved reflections from extrospection explicitly warn the LLM against its own past tendencies to produce shallow responses, thus directly combating "Ignoring the core concern". By providing a complete quadruple–including the comment, human response, past LLM attempt, and reflection–our method gives the LLM crucial context, which reduces cases of misinterpreting reviewer intent and avoids other failure modes. This structured learning from past mistakes is key to the performance gains illustrated in our case studies (Figs. 4 and 5) and quantitative results.

Consistency of assessment results across different LLM-based evaluators. In the previous part, the evaluation results of the responses were derived from GLM-4-Flash[37]. Notably, both the construction of the reflection bank and the generation of responses were also conducted using GLM-4-Flash. Although the LLM used for evaluation was reset to clear memory prior to each assessment, there remains uncertainty about whether the same model may exhibit biases toward content it has generated. To address this concern, we employed additional LLMs, including DeepSeek-v2.5[38] and Qwen-plus[39], as evaluators to assess the consistency of the assessment results.

Figure 6 shows the average rankings of five methods for response writing under different evaluators. For each evaluation method, we first determine the rankings based on the scores of different responses under a specific comment and then compute the average ranking across all comments to obtain the overall performance of each method. A smaller ranking indicates higher response quality for the corresponding method. Taking GLM-4-Flash as an example, the average rankings of the five methods, from smallest to largest, are RBB, SR, CoT, ABR, and DP. This result aligns with the findings presented in Tables 1 and 2. A comparison of the three evaluators–GLM-4-Flash, DeepSeek-v2.5, and Qwen-Plus–shows that, although the average rankings of the methods exhibit slight variations, the relative ordering of the five methods remains consistent. This consistency highlights the robust superiority of the proposed methods across different evaluators.

Generalization of the dual-loop reflection mechanism across different LLMs. In this part, the generalization of the reflection bank across different LLM bases will be discussed. The results in Tables 1 and 2



The introduction to the paper adequately summarises the requisite literature for interdisciplinary research in this field, but when it comes to justifying its own motivation, the paper runs into inconsistent statements. The introduction says that "Survey-based studies, which are the most commonly employed approach for such investigations, are often prone to inaccurate self-report responses and can be costly at scale." I don't really understand this type of motivational reasoning as the authors also rely on data filled in by the household members themselves. Who they think is a household member or head is not administratively generated data but self-reported. When I read this section in the introduction, I expected they had found a source that could identify household members without reporting bias.

Author response: ……In our case, we have leveraged a context- specific measure of overlapping household memberships derived from census data to map inter-household relationships. Nevertheless, we agree that this justification needs to be better motivated and more clearly elaborated. We have revised the entire paragraph as follows: "First, limited studies have been conducted in rural and low-income populations. Previous anthropological research has highlighted the importance of social networks as critical support systems in poorer environments. Yet, empirical investigation exploring the influence of social networks on inequalities in these settings pose significant scalability challenges. Survey-based approaches, while common, often encounter limitations in remote populations, hampered by high costs and inadequate infrastructure for extensive research activities. As a result, existing research tend to focus on specific population sub-groups, lacking data on network interactions observed at the population level……

Direct prompt: We appreciate the reviewer's comment regarding the introduction and the perceived inconsistency in motivational reasoning. We agree that the explanation regarding survey-based studies may be unclear in this context. To clarify, our use of the Africa Health Research Institute's Demographic Surveillance Area (AHRI DSA) data, which includes household memberships identified through household nominations rather than self-reported data, addresses the limitations of traditional survey methods. This methodological choice minimizes the potential for inaccurate self-report responses associated with survey-based studies……

CoT: We appreciate your careful reading and insightful comment. We acknowledge that the introduction may have caused some confusion regarding our data collection methods. In the introduction, we mentioned the limitations of survey-based studies, which indeed can be prone to inaccurate self-report responses and are often costly at scale. This statement was intended to highlight the general challenges associated with traditional data collection methods that rely on self-reported information. However, in our study, we did not rely solely on self-reported data for identifying household members. We utilized a unique census indicator – household memberships – which is administratively generated and based on the affiliations of individuals to 'social groups' rather than their actual residency. This approach helps to minimize reporting bias as it is not dependent on individual recollection……

Self Refine: Thank you for your thoughtful comments on the introduction of our manuscript. We appreciate your observation regarding the potential inconsistency between the discussion of survey-based studies and our own data collection practices. We recognize that our initial explanation could have been more precise. To address this, we will make the following revisions:1. Clarify the Role of Self-Reported Data: We will explicitly state that while our data includes self-reported information on socio-economic surveys, the identification of household members is not based on self-reporting but rather on administratively generated household memberships. This indicator is derived from census records, which reduces the potential for reporting bias associated with traditional survey methods. 2. Enhance the Justification: We will provide a more detailed justification for choosing our data collection method. We will explain that the use of household memberships from the census allows for a more accurate and standardized identification of household composition, which is crucial for understanding the dynamics of inter-household social networks……

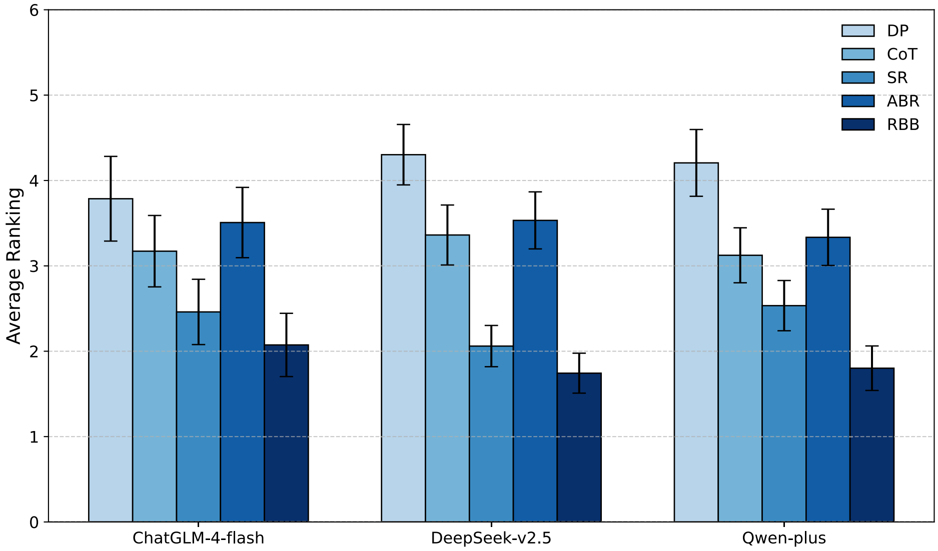
Answer-based reflection: Thank you for your insightful feedback. We appreciate your point regarding the inconsistency in the motivation presented in the introduction. We agree that the reliance on self-reported data by household members might appear to conflict with the concern about inaccurate self-report responses. To clarify and address this concern, we would like to reiterate that our study employs a unique approach to constructing the social network, which mitigates the potential for reporting bias. Specifically, we utilize the census indicator of household memberships, which is based on individuals' affiliation to a 'social group' rather than their actual residency. This method acknowledges the role of absent members in the household economy and allows us to identify inter-household ties without relying on self-reported information from household heads or members……

Retrieved reflections: …… The reviewer's comment focuses on the clarity and operationalization of the *socioeconomic status* (SES) construct in the study. The reviewer questions the inclusion of maternal nutrition and insurance status and suggests that the term "SES" may not fully capture the broader construct of social disadvantage…… *The human response provides specific examples of how the manuscript will be revised, such as changing the terminology from "SES" to "disadvantage" and including a contextualizing sentence in the introduction*……I should: - *More thoroughly understand the theoretical framework and context of the study's constructs*. - Provide a more detailed and nuanced response that *addresses the reviewer's concerns directly and theoretically*. - Offer specific examples of *how the manuscript will be revised* to enhance clarity and address the concerns raised.

Reflection bank based: Thank you for your insightful comment regarding the introduction of our paper. We appreciate your observation on the inconsistency in our motivational reasoning regarding the use of self-reported data. We acknowledge that our initial statement about the limitations of survey-based studies may have been misleading ……To address your concern, we will revise the text to better explain the unique aspects of our data collection process. We will highlight that the use of census data and household nominations provides a standardized and reliable identification of household composition, which is crucial for our analysis of social networks and economic inequality. Furthermore, we will enhance the justification in the introduction by discussing the scalability challenges associated with traditional survey-based approaches in remote and under-resourced settings……

Fig. 5 | Responses from different methods on Case 2. Red font indicates key statements in the author's response. Black underlining indicates expressions in the response that relate to key statements. Italicized underlining indicates records in the reflection bank that are closely related to the comment. Red underlining indicates expressions in the proposed method that coincide with the key statements.

Fig. 6 | The average rankings of five methods for response writing under different evaluators. Error bars in the figure represent one standard deviation (SD) to indicate the variability of the data within each group.



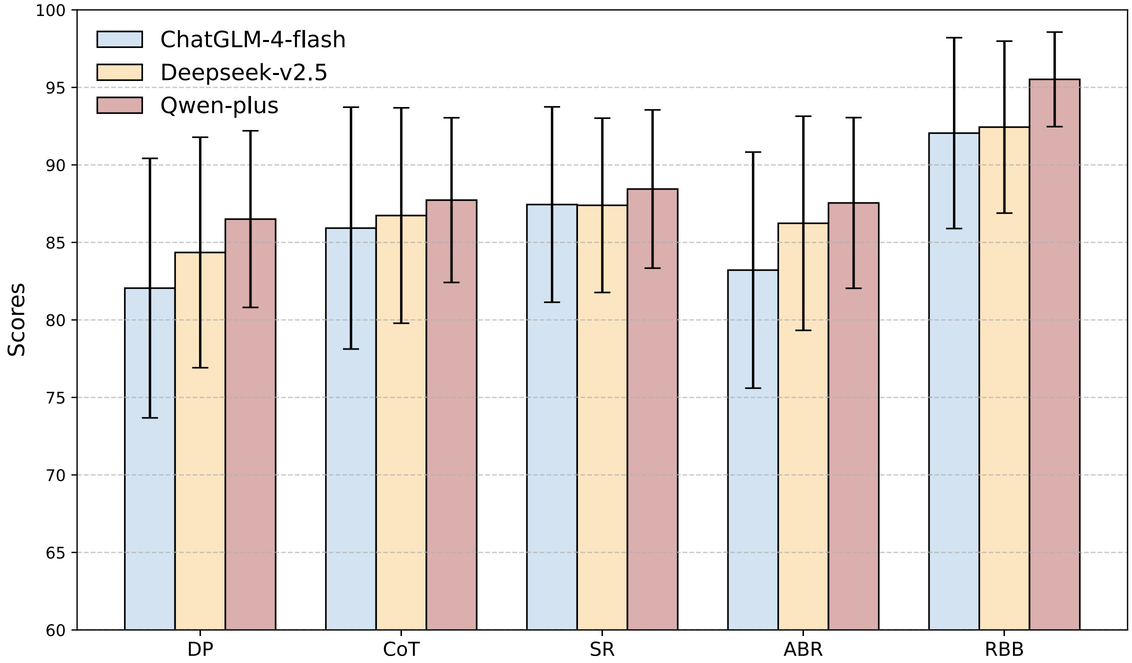


Fig. 7 | The evaluation scores of different approaches across various LLM bases. For the proposed RBB method, the reflection bank is constructed by GLM-4-Flash, and the reasoning based on the reflection bank is implemented by three LLM bases. Error bars in the figure represent one standard deviation (SD) to indicate the variability of the data within each group.

were obtained using GLM-4-Flash as a base, i.e., the reflection bank is constructed by GLM-4-Flash, as well as the reasoning based on the reflection bank is implemented by GLM-4-Flash. However, the construction of the reflection bank and the reasoning process can, in fact, be decoupled. The construction of the reflection bank is a computationally intensive process that can be efficiently handled by inexpensive and simpler LLMs. In contrast, reflection bank-based reasoning places higher demands on the performance of the LLM, as stronger reasoning capabilities yield higher-quality responses. To evaluate the generalization capability of the reflection bank constructed using GLM-4-Flash, this section explores its application to different model bases during the inference phase. Since the consistency of the assessment results of the different evaluators has been shown in the previous part, only the GLM-4-Flash is used here as the evaluator.

Figure 7 illustrates the evaluation scores of different approaches across various LLM bases, with the proposed RBBs utilizing the reflection bank constructed by GLM-4-Flash. From the figure, several key observations can be made. First, there is a noticeable gap in response-writing capabilities among the LLMs. For instance, the CoT performance of DeepSeek-v2.5 and the DP performance of Qwen-Plus are comparable to the SR performance of GLM-4-Flash. Second, both CoT and SR demonstrate positive effects across different LLM bases; however, their improvements exhibit a saturation effect. For example, the performance gain from CoT to SR is marginal, as neither approach introduces additional information. The improvements achieved solely through their inherent capabilities reach an upper limit. Third, although ABR incorporates historical human responses for reference, its performance is inferior to both CoT and SR. This is attributed to the inability of human responses to alert the LLMs to potential issues in its output and, in some cases, the interference caused by human responses. This underscores that the introduction of additional information does not necessarily yield positive effects. Fourth, the RBB demonstrates consistent and significant improvements across all LLM bases. This highlights, on the one hand, the generalizability of the reflection bank across different LLM bases and, on the other hand, the ability of reflection bank-based reasoning to enable the LLM to effectively leverage insights from additional information, thereby further enhancing the quality of responses.

Table 4 | GSB evaluation results comparing model outputs to author responses (*N* = 200)

Method	Good	Same	Bad
ChatGLM-DP	2.5%	15.5%	82.0%
ChatGLM-CoT	4.0%	35.0%	61.0%
ChatGLM-SR	6.5%	42.5%	51.0%
ChatGLM-ABR	3.0%	31.0%	66.0%
Ours	9.0%	55.5%	35.5%

Evaluation of post-editing effort with GSB score

Evaluating the post-editing or refinement efforts of LLM-generated outputs, such as measuring time saved, error reduction, or user satisfaction, is an important aspect of assessing practical utility. However, conducting such a study is challenging in terms of standardization and objectivity. To provide a quantitative proxy for the post-editing effort, we conducted the widely used GSB (Good, Same, Bad) evaluation[42].

This evaluation compares the outputs of each method against the ground-truth author responses, where a higher proportion of "Same" or "Good" ratings indicates a lower need for manual correction. Specifically, we sample 200 comments from our test set and use DeepSeek-R1 to classify the generated response from each method relative to the human response. A response is labeled "Same" if it captured the core logic and addressed the reviewer's concern with a quality comparable to the author version. As presented in Table 4, our proposed RBB-LLM framework demonstrates a clear advantage. The results show that our method produces responses of "Same" quality as humans in 55.5% of the cases. This means that more than half of the responses are already close to human results, an improvement of 13% over the second-best algorithm. It implies a greatly reduced workload for researchers during the post-editing phase.

Discussion

In this article, we design an LLM-assisted response letter writing framework to improve the efficiency of response letter writing for researchers and alleviate the lack of response experience for scholars in marginalized areas. We point out that LLM-assisted response letter writing is not a trivial task, but faces the *shallow reasoning problem*, which means that the LLM sometimes outputs responses that are polished in structure and style but fail to address the core of the comment. Inspired by the metacognition in cognitive psychology, we design a dual-loop reflection mechanism that involves both extrospection and introspection. Different from the existing reflection methods that focus solely on introspecting about the output, the LLM is instructed to detach from itself and critique its own reasoning process with human reference responses, i.e., extrospection. When reasoning, the LLM rewrites the initial response based on the reflection of the extrospection, i.e., introspection. The proposed reflection bank based on dual-loop reflections was constructed based on 4000 papers and peer review documents (approximately 79,000 comments) from *Nature Communications*, covering the fields of physical sciences, earth and environmental, biological sciences, health sciences, scientific community and society. The proposed dual-loop reflection method successfully mitigates the *shallow reasoning problem* and improves the quality of response writing. Validation results on more than 3700 comments of 200 papers demonstrate the effectiveness and superiority of the proposed method on different article types and comment types. The average performance of the proposed method across all article types exceeds the direct prompting by 10.05 points and outperforms the second-highest performing self-refine by 4.97 points. Additional experiments are implemented to demonstrate the stability of the proposed method's performance on different evaluators and the generalization of the dual-loop reflection mechanism across LLM bases.

Although the proposed framework shows potential in improving the drafting process for response letters, we emphasize that our framework is designed to assist, not replace human authors. First, this framework is designed as a time-saving tool that frees researchers to devote more time to the core scientific thinking and experimentation that truly advance knowledge. Second, it serves as an educational "teacher" and "guider" by providing structural suggestions, which is particularly beneficial for non-native English-speaking researchers and early-career or resource-constrained researchers. Finally, this framework must be utilized as a human-in-the-loop system, the final responsibility for content, tone, and scientific reasoning always remains with the human author. We advocate for the responsible use of this framework as a tool that improves, rather than undermines, the integrity of the peer review process.

Although our work utilizes open-sourced peer review files to build the reflection bank and reflection bank-based reasoning methods, and demonstrates their generalizability across different review types, article types, and large model bases, we encounter certain limitations.

First, in the construction of the reflection bank, we only applied the first round of review information, i.e., the first round of review comments followed by the authors' responses. In fact, some papers have multiple rounds of review, and the reviewers' satisfaction with the authors' responses in the first round is actually implied in the questions in the second round, which is a part of the information we did not utilize. In future work, we would extend our framework to model these dialogue-based reviewer-author chains. This would involve structuring the data as conversational threads (e.g., Comment 1 -> Response 1 -> Follow-up Comment 2 -> Final Response 2). By doing so, the reflection process can evolve to become dialogue-aware, enabling the LLM to learn more complex strategies. Besides, it is promising to extend our collection to include peer-reviewed documents from different academic venues. For example, incorporating materials from computer science conference proceedings on OpenReview (e.g., ICLR, NeurIPS) would expose our model to different review styles and formats. Moreover, expanding the dataset to include cross-lingual peer reviews from other major publishers represents a significant next step.

Second, searching for relevant quadruples from the reflection bank uses only reviewers' comments and the standard RAG technique, which can make the retrieved texts potentially not optimal. One direction that could be improved is to introduce LLM to rewrite or split the reviewer questions for parallel retrieval[43], which would improve the relevance of the retrieved information. Meanwhile, the proposed reasoning process based on the retrieved texts is based on the classical CoT. In fact, CoT has many variants and upgraded versions, such as self-consistency CoT[44], Tree of Thoughts (ToT)[45], and Meta-CoT[46]. Incorporating these more advanced CoT techniques or LLMs with reasoning capability, e.g., DeepSeek-R1[47], can further improve the quality of response writing.

Third, one promising direction for future work is using different LLMs for the extrospection process. In the current framework, both the initial response and the reflection are generated by the same model. This approach carries a potential risk of confirmation bias or homogenization, as the LLM might have its own inherent biases or tendencies in reasoning[48]. To mitigate this, a future enhancement could use a heterogeneous setup where a separate "critic" LLM generates the reflection. It could produce more objective feedback, further strengthening the framework's ability to overcome shallow reasoning.

Last, the potential of the dual-loop reflection mechanism extends beyond academic writing. The "shallow reasoning problem" is critical in any domain where LLMs must deliver precise and subtle insights to humans. Our framework can be directly extended to such scenarios, such as grant rebuttals, peer review, and code review. Adapting the framework to these scenarios requires constructing a customized reflection bank. Building the quadruples can follow a unified structure: (1) the problem or item to be reviewed (e.g., a critical comment or code snippet), (2) the expert human example (e.g., a successful rebuttal or a senior engineer's review), (3) the LLM's initial response, and (4) the reflection that identifies the strategic or conceptual gap between the LLM's output and the expert's. In this way, the reflection bank guides the LLM to overcome its specific weaknesses in each context and generate outputs that approach an expert level.

Methods

In this part, we provide details on our methodology. We begin by providing an overview of the proposed framework. Next, we describe the collection of peer review documents and the content filtering process. Following this, we detail the construction of the reflection bank. Subsequently, we outline the implementation process of the LLM using the reflection bank for response writing. Finally, we present the comparison and evaluation methods to evaluate the quality of generated responses.

Overview of the RBB-LLM framework

As shown in Fig. 8, the proposed framework consists of three components, including peer-review corpus collection and filtering, reflection bank construction, and LLM reasoning based on the reflection bank. In the first component, publicly available peer review documents are collected and filtered to produce structured data, comprising articles, comments, and human responses. In the second component, based on this structured data, the LLM performs reasoning using CoT and generates reflections informed by human responses, ultimately producing quadruples for constructing the reflection bank. The third component utilizes the reflection bank to retrieve similar comments, enabling the LLM to enhance the accuracy and depth of the current comment by leveraging retrieved quadruples. In summary, the first component provides the foundational corpus for the second, and the third performs the reasoning based on the reflection bank obtained in the second. Detailed descriptions of these components follow in subsequent sections.

Peer review corpus collection and filtering

We source peer-reviewed documents from papers published from January 1, 2024 to September 25, 2024 in *Nature Communications* website (https://www.nature.com/ncomms/). If the authors have agreed to make the peer review documents publicly available, these files are hosted in the Supplementary Information section, from which we download them using a Python script. The downloaded articles and peer review files are in pdf format, and we use pypdf2 (https://pypdf2.readthedocs.io/en/3.x/) to convert them to text form.

The peer review documents are unstructured, making it challenging for the program to locate a specific comment from a particular reviewer or match it with the corresponding response. This lack of structure hinders the program's ability to efficiently read, modify, and augment the peer review file. To obtain structured peer review files, we design a matching-based filtering algorithm that separates each individual comment and its corresponding reply. This algorithm efficiently converts each article's peer review file into a JSON format, enabling the program to extract key information, including the article's title, links, the number of reviewers, each reviewer's comments, and the corresponding author responses. This file structure also serves as the basis for the construction of the reflection bank.

Algorithm 1. Comment-Response Matching Algorithm

```
function MatchCommentsQ, L
  Tokenize questions ($Q$) & responses ($L$) into sentences: $\{q_1, \ldots, q_n\}$, $\{r_1, \ldots, r_m\}$
  Initialize:
    $current\_group \leftarrow \emptyset$, $result \leftarrow \emptyset$, $last\_pos \leftarrow -1$
  for each question sentence $q_i \in Q$ do
      Find response sentence $r_j$ with maximum similarity $\mathrm{sim}(q_i, r_j) > \tau$
      if $j = last\_pos + 1$ then          ▷ Continuous match: current matches next position
          $current\_group \leftarrow current\_group \cup \{q_i\}$          ▷ Extend current comment group
      else          ▷ Non-consecutive match: boundary detection
          if $current\_group \neq \emptyset$ then
              Extract response: $\bigcup_{k=last\_pos+1}^{j-1} r_k$          ▷ Response between last match & current
              Add ($current\_group$, $response$) to $result$
          end if
          $current\_group \leftarrow \{q_i\}$          ▷ Start new comment group
      end if
      $last\_pos \leftarrow j$          ▷ Update anchor position
  end for
  return $result$          ▷ Structured {comment group, response} pairs
end function
```

The challenge with this process is that the style of peer review documents is diverse, making it difficult to extract text using uniform rules. Specifically, some authors label the questions as "Comment #*n*" or "Question #*n*", where *n* represents the question number. Some authors rely on different text colors to distinguish between comments and responses. However, this color information is lost when the program processes the file as text, making the questions and responses indistinguishable.

To address this problem, we subtly utilize the pattern of content in the peer review document, where the comment letter and the response letter appear sequentially. Our algorithm relies on a matching strategy, leveraging the following two key observations: The content from the comment letter also appears in the response letter, as authors usually restate the reviewers' questions. The unmatched content within the matched content of the response letter corresponds to the authors' responses.

First, we design regular expressions for matching reviewers, and chunk the comment letter and response letter based on reviewer identity.

Second, each sentence from the reviewer's comments is compared against the sentences in the response letter. If consecutive sentences from the reviewer's comments match consecutive sentences in the response letter, they are grouped into a single "comment section".

Third, when the match is discontinuous (e.g., a gap between matched positions), the current comment section is closed, and a new section starts. The content following the last matched position of the current comment section and preceding the next matched position is identified as the author's response to the current comment.

In this way, the unstructured text of a peer review file is transformed into a structured JSON file, as illustrated in Algorithm 1 and Fig. 8b. This format organizes reviewer comments and human responses, grouped by reviewers, facilitating easy retrieval and processing. In this work, images in comments or replies are not taken into consideration. In the future, multimodal LLM or VQA technology [49,50] can be used to utilize this modality.

Reflection bank construction based on extrospection

In this part, the construction process of the reflection bank is presented.

First, based on the structured JSON files, we iterate through the reviewer questions for the papers sequentially, instructing the LLM to generate initial responses through chain reasoning. Chain reasoning is realized through the basic CoT technique. The LLM is instructed to thoroughly understand the paper and the reviewer's comment by identifying key questions, concerns, or suggestions. Then, it is directed to perform an internal reasoning process, where it breaks down the response logically: identifying the main issue raised, reflecting on how the concern relates to the author's work, deciding whether the author agrees or disagrees with the reviewer, justifying the stance, and proposing actions to address the concern. After this reasoning, the LLM is instructed to generate a clear, concise, and professional response. Specifically, for a reviewer's question Q and a chain-of-thought (CoT) prompt P_{CoT}, the LLM M generates the initial response A_M through conditional sampling:

$$A_M \sim M(Q \mid P_{\mathrm{CoT}}) \tag{1}$$

Second, we provide human responses to the LLM, prompting it to reflect on its initial responses. This process is akin to extrospection in cognitive psychology, where LLM is instructed to detach from itself and critique its own reasoning process with a human reference response. The LLM is asked to compare its previous response with a human-written response, analyze differences in tone, logic, and structure, and identify areas for improvement. The reflection process includes understanding the reviewer's comment, evaluating how well the LLM's response addresses the comment compared to the human response, and identifying weaknesses such as unclear logic, insufficient detail, or tone issues. The LLM is then tasked with documenting these reflections for future reference, ensuring it can learn from the mistakes and refine its reasoning for future tasks. Specifically, given the initial LLM response A_M, a human-written response A_{H}, and a extrospection prompt P_{Extro}, the LLM generates a reflection R by

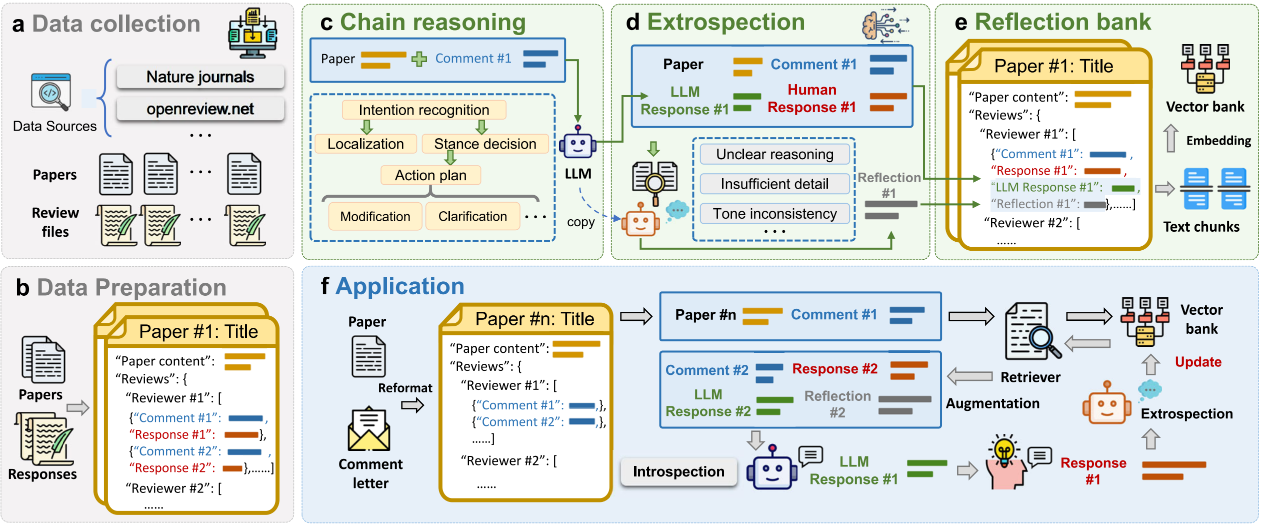


Fig. 8 | Overview of the proposed framework with the dual-loop reflection mechanism. The framework consists of three key steps. The (**a**) and (**b**) demonstrate peer-review corpus collection and preparation. The (**c**–**e**) are about the LLM extrospection and reflection bank construction. The (**f**) shows the introspection and reasoning process based on the reflection bank.

comparing A_M and A_H:

$$R \sim M(Q, A_M, A_\mathrm{H} \mid P_\mathrm{Ex}) \tag{2}$$

Once all the reviews in a JSON file (corresponding to a paper) have been processed, the two key-value pairs under each comment are updated to quadruples: the reviewer's comment Q, the human response A_H, the LLM's response A_M, and the LLM's reflection R on its own response. To build the reflection bank, we transform the JSON files of all the articles into embedding vectors, which can efficiently capture semantic meaning in a compact, high-dimensional space, enabling faster and more accurate retrieval. In this work, all-MiniLM-L6-v2(BERT)[51] is adopted as the embedding model to obtain a 384-dimensional dense vector, and Chroma is used as the vector database to store embedding vectors generated from all-MiniLM-L6-v2. Converting sentences into embedding vectors requires chunking the text to ensure that the length of the input sentences does not exceed the embedding model's limitations. We set the chunk size to 200, a widely used and efficient choice. To avoid mixing different comments, we define the smallest unit of chunking to be a quadruple, meaning that no chunk will contain more than one quadruple. Since we require the complete quadruple for retrieval, rather than just the most similar chunk, we use the ParentDocumentRetriever module (https://python.langchain.com/docs/how_to/parent_document_retriever/). This approach first fetches the smaller chunks for precision matching, then looks up the parent IDs for those chunks, and returns the complete quadruple for comprehensive recall. This module resolves the conflict between precision matching and comprehensive recall caused by the chunk size.

Introspection and reasoning based on the reflection bank

In this part, LLM-assisted response writing based on the reflection bank is described.

First, the introspection and reasoning process begins by retrieving relevant quadruples from the reflection bank. After the reflection bank is constructed, it stores numerous embedding vectors, with each vector corresponding to a chunk of text. The parent document ID for each chunk is recorded to locate the original quadruple to which the chunk belongs. When the target article and reviewer's comment are given, the reviewer's comment is first retrieved as a query in the reflection bank according to similarity[52]. The parent document corresponding to the most matching chunk, i.e., the quadruple whose slice yields the current chunk, is recalled. The target comment, the paper of the target comment, and the recalled quadruple are fed together into the LLM to complete the response writing task.

Second, we design prompts to guide the LLM in leveraging past reflections and current materials. The key steps involve employing CoT to obtain the initial response and leveraging historical reflections to introspection and give an improved response. For the former step, the initial response is obtained similarly to Formula (1). For the latter step, LLM is instructed to focus on the areas where AI responses historically diverged from human responses. This could be in tone (e.g., AI being too direct or formal), depth of explanation (e.g., missing key contextual details), or misunderstanding of the critique. By introspecting these differences, the LLM is prompted to produce an improved reply to the current comment that aligns more closely with the expected human-like reasoning and communication style, effectively addressing the reviewer's concerns. This process can be formulated as:

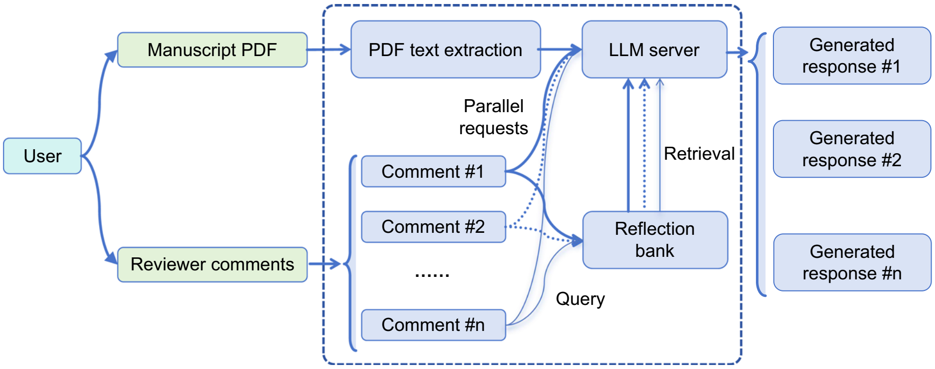
$$\begin{aligned} A_M &\sim M(Q \mid P_\mathrm{CoT}) \\ A'_M &\sim M\left(Q, \mathscr{H}, A_M \mid P_\mathrm{In}\right) \end{aligned} \tag{3}$$

where P_In denotes the introspection prompt, and $\mathscr{H} = \{Q', A'_H, A'_M, R'\}$ denotes the retrieved quadruple, including past reviewer questions Q', human responses A'_H to Q', LLM's initial responses A'_M to Q', and Reflection records R'.

Once the output of the LLM is obtained, humans can refine it by making adjustments or drawing inspiration from the model's output, ultimately producing a version that is suitable for submission. Since the reflection bank is essentially an embedding vector bank, we can easily update or add new vectors to it as needed. On the one hand, after we have refined the human responses for the current comment, we can have the LLM compare its own responses with the human responses and output the reflection, thereby constructing the quadruple as shown in Fig. 8d. These quadruples can be incremented into the reflection bank by obtaining the embedding vectors along with parent-child document relationships in the way of Fig. 8e. On the other hand, after obtaining the quadruple, we can add issues not recognized by the LLM or annotate them from a new perspective in the reflection part. As a result, author or research group preferences can also be included and recalled, which can help the LLM to meet the needs of personalization and diversity.

Another significant advantage of the proposed method is the low computational power requirement. Instead of locally downloading and running the LLM, only local maintenance of the designed reflection bank

Fig. 9 | Practical application workflow of the RBB-LLM framework.



and calls to the LLM API are required to realize the use of the proposed method on a laptop.

Practical application workflow of our framework

To further clarify the practical usability of our framework from a researcher's perspective, we have designed a user-centric workflow diagram. This process highlights how a user would interact with the RBB-LLM framework in a real-world scenario to draft a response letter. As shown in Fig. 9, the workflow begins with the user providing the manuscript and reviewer comments. Individual comments are processed through parallel requests. The reviewers' questions are simultaneously used as queries to search the reflection bank, and the retrieved quadruples are fed into the LLM server. The LLM server generates responses in parallel for multiple comments based on the proposed RBB-LLM framework combined with the retrieved content from the reflection bank. For our framework, the heavy computational burden of building the reflection bank is a one-time and centralized effort, while the day-to-day use is lightweight and accessible to any researcher with a standard computer. This entire process can be integrated into a researcher's workflow, for instance, through a simple web interface, requiring only API calls to the LLM and local maintenance of the reflection bank.

Comparison and evaluation methods

To validate the general applicability of the proposed RBB-LLM across different LLMs, we selected three LLM bases, including GLM-4-Flash[37], DeepSeek-v2.5[38], and Qwen-Plus[39], with differences in their capabilities. To ensure fairness, all methods are compared using the same LLM base. The reflection bank was constructed using GLM-4-Flash, based on cost considerations. We have discussed the generalization and effectiveness of using the reflection bank across different LLM bases in the Results section.

Direct prompting (DP): In this approach, the paper and the reviewer's comment are directly provided to the LLM, and the model is tasked with generating a response without additional guidance or structured reasoning. This method serves as a baseline for evaluating the LLM's ability to generate accurate responses without any pre-structured prompts or external reasoning mechanisms.

CoT[22]: The CoT method enhances the prompting process by encouraging the model to engage in step-by-step reasoning before generating output. For our task, when providing the paper and one comment to the model, we use prompts to direct the model through the following reasoning steps: identifying the main issue or suggestion, reflecting on the author's work and its relation to the concern, deciding whether to agree or disagree with the reviewer and justifying the decision, and proposing actions to address the concern. This approach aims to produce more thoughtful and logically sound responses by breaking down the reasoning process.

Self-refine (SR)[33]: The SR method improves initial outputs from LLMs through iterative feedback and refinement. For our task, the SR method builds upon CoT by incorporating a self-reflection step, where the model reviews and refines its initial response, ensuring that it fully addresses the reviewer's concerns. The model is encouraged to reflect on whether the response sufficiently resolves the issues raised by the reviewer, paying particular attention to the alignment with the reviewer's intent and offering improvements where necessary.

Answer-based reasoning (ABR): This method is a degraded variant of the proposed framework. After CoT reasoning on a given article and reviewer comment, the LLM is provided with historical, similar comments and their corresponding human-crafted responses (reference answers) to learn potentially useful response strategies. Unlike the proposed framework, this approach does not include LLM responses to historically similar comments or reflection on the gap between LLM-generated and human responses.

The quality of the generated response letter is assessed using both the LLM and human evaluators. The LLM is responsible for scoring responses to all comments on a scale of 1 to 100, providing a comprehensive and systematic evaluation, commonly referred to as "LLM-as-a-Judge"[53–55]. In parallel, human evaluators perform detailed analyses of specific cases, offering deeper insights and validating the accuracy and relevance of the responses. Notably, we do not evaluate the generated responses alongside human-crafted responses to directly compare their scores or aim for the LLM to surpass human performance. The purpose of our work is not to create responses indistinguishable from human ones—an outcome that might invite academic misuse—but to provide meaningful assistance and guidance to human authors. It is sufficient for the LLM to produce responses aligned with the main focus of human responses and address the reviewer's concerns. Therefore, for the LLM evaluator, we provide human responses as a reference for the LLM to determine whether the current response is consistent with the core focus of the human response and address the reviewer's concerns, which also avoids the LLM's preference for overly polished but superficial content. In this work, three LLMs—GLM-4-Flash, DeepSeek-v2.5, and Qwen-Plus—are chosen as evaluators. In addition, to ensure unbiased scoring, the LLM's memory is reset before evaluating responses to each comment. This prevents any interference between comments and any memory of generating the response letter. Besides, to enhance the reliability of our evaluation, we adopt the widely used Good/Same/Bad (GSB) evaluation. In this setup, we present the response generated by LLMs and the human-authored response to a judge LLM (DeepSeek-R1). Nevertheless, we admit that even this method cannot fully match the detailed and context-aware judgment of human experts, which is a limitation of the present work. Therefore, future work would greatly benefit from incorporating large-scale evaluations with human experts.

Data availability

The processed TXT files of the papers and peer review files used in this study, the JSON files containing quadruplets for constructing the reflection bank, and the vector base of the reflection bank are available at https://github.com/chunhuiz/ResponseLLM.

Code availability

The peer review corpus collection and filtering code, the code for the proposed method, and the prompts are available at https://github.com/chunhuiz/ResponseLLM.

Received: 8 May 2025; Accepted: 7 October 2025;
Published online: 01 December 2025

References

1. Mudrak, Ben. Scholarly publishing: A brief history. *Am. J. Exp*. https://www.aje.com/arc/scholarly-publishing-brief-history (accessed November 19, 2024).
2. Aczel, B., Szaszi, B. & Holcombe, A. O. A billion-dollar donation: estimating the cost of researchers' time spent on peer review. *Res. Integr. Peer Rev.* **6**, 14 (2021).
3. Merton, R. K. The Matthew Effect in Science: The reward and communication systems of science are considered. *Science* **159**, 56–63 (1968).
4. Nature will publish peer review reports as a trial. *Nature* 578, 8, https://doi.org/10.1038/d41586-020-00309-9 (2020).
5. Nature is trialling transparent peer review - the early results are encouraging. *Nature* 603, 8, https://doi.org/10.1038/d41586-022-00493-w (2022).
6. Zhao, W. X. et al. A Survey of Large Language Models. Preprint at arXiv.2303.18223 (2024).
7. Naveed, H. et al. A Comprehensive Overview of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **16**, 105 (2025).
8. Zhao, J., Yue, J., Zhao, C. & Chen, C. Adjust to reality: LLM-driven test-time semantic adjustment for zero-shot fault diagnosis. *Control Eng. Pract.* **164**, 106406 (2025).
9. Qiu, P. et al. Towards building multilingual language model for medicine. *Nat. Commun.* **15**, 8384 (2024).
10. Xie, Q. et al. PIXIU: A Comprehensive Benchmark, Instruction Dataset and Large Language Model for Finance. *Adv. Neural Inf. Process Syst.* **36**, 33469–33484 (2023).
11. Kasneci, E. et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **103**, 102274 (2023).
12. Chowdhury, S. & Soni, B. R-VQA: A robust visual question answering model. *Knowl. -Based Syst.* **309**, 112827 (2025).
13. Chowdhury, S. & Soni, B. Beyond Words: ESC-Net Revolutionizes VQA by Elevating Visual Features and Defying Language Priors. *Comput. Intell.* **40**, e70010 (2024).
14. Chowdhury, S. & Soni, B. ENVQA: Improving Visual Question Answering model by enriching the visual feature. *Eng. Appl. Artif. Intell.* **142**, 109948 (2025).
15. Romera-Paredes, B. et al. Mathematical discoveries from program search with large language models. *Nature* **625**, 468–475 (2024).
16. Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **630**, 625–630 (2024).
17. Liang, W. et al. Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews. *Proc. Int. Conf. Mach. Learn.* **235**, 29575–29620 (2024).
18. Gruda, D. Three ways ChatGPT helps me in my academic writing. *Nature*. https://www.nature.com/articles/d41586-024-01042-3 (2024).
19. Liang, W. et al. Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis. NEJM AI. 1, 8 (2024).
20. Liu, R. & Shah, N. B. ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing. Preprint at http://arxiv.org/abs/2306.00622 (2023).
21. Hunt, M. J. et al. How to write an effective response letter to reviewers. *Med. Sci. Pulse* **13**, 60–63 (2019).
22. Wei, J. et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Adv. Neural Inf. Process Syst.* **35**, 24824–24837 (2022).
23. Kojima, T., Gu, S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large Language Models are Zero-Shot Reasoners. *Adv. Neural Inf. Process Syst.* **35**, 22199–22213 (2022).
24. Sprague, Z. et al. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. Preprint arXiv:2409.12183 (2024).
25. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 610-623 (Association for Computing Machinery, New York, NY, USA, 2021).
26. Wu, M. & Aji, A. F. Style Over Substance: Evaluation Biases for Large Language Models. *International Conference on Computational Linguistics*, **21**, 297–312 (2023).
27. Han, Z., Gao, C., Liu, J., Zhang, J. & Zhang, S. Q. Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. Preprint at arXiv.2403.14608 (2024).
28. J, M. R., VM, K., Warrier, H. & Gupta, Y. Fine Tuning LLM for Enterprise: Practical Guidelines and Recommendations. Preprint at arXiv.2404.10779 (2024).
29. Mao, Y. et al. A survey on LoRA of large language models. *Front. Comput. Sci.* **19**, 197605 (2024).
30. Foos, P. W., Mora, J. J. & Tkacz, S. Student study techniques and the generation effect. *J. Educ. Psychol.* **86**, 567–576 (1994).
31. Bertsch, S. et al. The generation effect: A meta-analytic review. *Mem. Cognition* **35**, 201–210 (2007).
32. Dunlosky, J. & Metcalfe, J. Metacognition. SAGE Publications (2008).
33. Madaan, A. et al. Self-Refine: Iterative Refinement with Self-Feedback. *Adv. Neural Inf. Process Syst.* **36**, 46534–46594 (2023).
34. Renze, M. & Guven, E. Self-Reflection in LLM Agents: Effects on Problem-Solving Performance. Preprint at arXiv.2405.06682 (2024).
35. Xu, D. et al. Does Few-Shot Learning Help LLM Performance in Code Synthesis? Preprint at arXiv.2412.02906 (2024).
36. Writing your report. *Nature Portfolio*. https://www.nature.com/ncomms/for-reviewers/writing-your-report (accessed November 19, 2024).
37. Team GLM. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. Preprint at arXiv.2406.12793 (2024).
38. DeepSeek-AI. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. Preprint at arXiv.2401.02954 (2024).
39. Bai, J. et al. Qwen Technical Report. Preprint at arXiv.2309.16609 (2023).
40. Chen, T., Karedla, N. & Enderlein, J. Measuring sub-nanometer undulations at microsecond temporal resolution with metal- and graphene-induced energy transfer spectroscopy. *Nat. Commun.* **15**, 1789 (2024).
41. Yu, S.-T. et al. Local Network Interaction as a Mechanism for Wealth Inequality. *Nat. Commun.* **15**, 5322 (2024).
42. Cai, Y. & Yuan, Y. CAR-Transformer: Cross-Attention Reinforcement Transformer for Cross-Lingual Summarization. *AAAI* **38**, 17718–17726 (2024).
43. Zhao, S. et al. Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely. Preprint at arXiv.2409.14924 (2024).
44. Wang, X. et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models. Preprint at arXiv.2203.11171 (2023).
45. Yao, S. et al. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *Adv. Neural Inf. Process Syst.* **36**, 11809–11822 (2023).

46. Zou, A., Zhang, Z., Zhao, H. & Tang, X. Generalizable Chain-of-Thought Prompting in Mixed-task Scenarios with Large Language Models. Preprint at arXiv.2310.06692 (2024).
47. Guo, D. et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **645**, 633–638 (2025).
48. Wataoka, K., Takahashi, T. & Ri, R. Self-Preference Bias in LLM-as-a-Judge. Preprint at arXiv.2410.21819 (2025).
49. Chowdhury, S. & Soni, B. QSFVQA: A Time Efficient, Scalable and Optimized VQA Framework. *Arab J. Sci. Eng.* **48**, 10479–10491 (2023).
50. Chowdhury, S. & Soni, B. Handling language prior and compositional reasoning issues in Visual Question Answering system. *Neurocomputing* **635**, 129906 (2025).
51. Koroteev, M. V. BERT: A Review of Applications in Natural Language Processing and Understanding. Preprint at arXiv.2103.11943 (2021).
52. Fan, W. et al. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. in Proc. 30th ACM SIGKDD, 6491-6501 (2024).
53. Raina, V., Liusie, A. & Gales, M. Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM Assessment. In *Empirical Methods in Natural Language Processing*. **427**, 7499–7517 (2024).
54. Gu, J. et al. A Survey on LLM-as-a-Judge. Preprint at arXiv.2411.15594 (2025).
55. Li, D. et al. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. Preprint at arXiv.2411.16594 (2025).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62450020), the National Natural Science Foundation of China (No. 62125306), and the State Key Laboratory of Industrial Control Technology, China (ICT2025C01).

Author contributions

Baoxue Li and Chunhui Zhao contributed to the conception or design of the work, and further performed the acquisition, analysis, or interpretation of data for the work. In writing, Baoxue Li drafted the work, and Chunhui Zhao reviewed it critically for important intellectual content. All authors approved the version to be published and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Chunhui Zhao is the corresponding author.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Chunhui Zhao.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025