JRE-L: Journalist, Reader, and Editor LLMs in the Loop for Science Journalism for the General Audience

Gongyao Jiang[1], **Xinran Shi**[1], **Qiong Luo**[1,2,*]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology
jianggongyao@gmail.com, luo@cse.ust.hk

Abstract

Science journalism reports current scientific discoveries to non-specialists, aiming to enable public comprehension of the state of the art. This task is challenging as the audience often lacks specific knowledge about the presented research. We propose JRE-L, a framework that integrates three LLMs mimicking the writing-reading-feedback-revision loop. In JRE-L, one LLM acts as the journalist, another LLM as the general public reader, and the third LLM as an editor. The journalist's writing is iteratively refined by feedback from the reader and suggestions from the editor. Our experiments demonstrate that by leveraging the collaboration of two 7B and one 1.8B open-source LLMs, we can generate articles that are more accessible than those generated by existing methods, including prompting single advanced models such as GPT-4 and other LLM-collaboration strategies. Our code is publicly available at github.com/Zzoay/JRE-L.

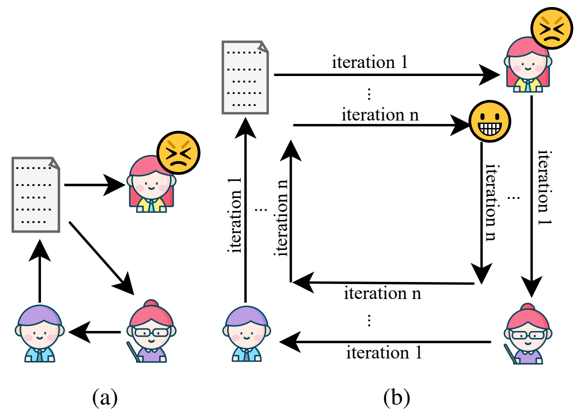


Figure 1: An article written by a science journalist may be challenging for the general reader without the reader's feedback to the editor in the revision cycle (a). Incorporating the reader's feedback into the journalism cycle can help enhance the readability of the article (b).

1 Introduction

Science journalism creates news articles that cover a wide range of scientific research, enhancing the public's understanding of science (Göpfert, 2008; Allan, 2011; Angler, 2017). With rapid advances in various disciplines, science journalism struggles to keep pace with the exponential growth of knowledge. In response, automatic science journalism (ASJ) has been proposed to expedite the filtering, learning, and communication of scientific knowledge (Dangovski et al., 2021).

The essence of ASJ lies in elucidating complex technical content for readers, thereby facilitating their comprehension of advanced research (Cardenas et al., 2023). The degree to which content is embraced depends on the reader's domain knowledge (August et al., 2024), thus scientific content may be challenging for the general reader, as illustrated in Figure 1 (a). Some researchers have developed parallel corpora (Dangovski et al., 2021; Goldsack et al., 2022; Cardenas et al., 2023), where the target content is extracted from online scientific news or journals. However, these press releases often remain technical, likely because they are originally tailored for researchers rather than the general audience.

Large language models (LLMs) have shown impressive proficiency in instruction adherence and content generation (Achiam et al., 2023; Bai et al., 2023), thereby making them potential tools for ASJ. Furthermore, LLMs have exhibited social intelligence, enabling them to play realistic roles and collaborate in real-world tasks (Park et al., 2023; Talebirad and Nadiri, 2023; Qian et al., 2024). Moreover, LLMs can iteratively improve their performance by context updates without training (Zhao et al., 2024; Senel et al., 2024). Motivated by these observations, we propose a com-

*Corresponding Author

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6579–6594
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

municative and iterative framework that leverages LLMs to gradually accomplish the ASJ task.

Our goal is to automatically generate a popular science article based on a technical paper and improve the accessibility of the generated article to the general audience. In real-world journalism, a journalist typically receives and learns from revision suggestions from a professional editor (Nip, 2006; Anderson et al., 2015), as reviews can lead to writing improvements (Bryant, 2002; Cho and MacArthur, 2011). Furthermore, in science communication, receiving feedback from the audience can improve scientists' communication skills (Brownell et al., 2013; Clark et al., 2016). Thus, the introduction of general readers' feedback into the journalism loop is promising to make the writing more accessible to the general public, as illustrated in Figure 1 (b).

Based on real-world practices and research results, we design the JRE-L framework, in which three LLMs collaborate in a loop of writing, reading, feedback, and revision, to generate highly accessible popular science articles. Concretely, we have an LLM serve as the journalist writing for readers who lack domain knowledge of the article. To help expose writing problems that might hinder the reading experience of general readers, we have another LLM, smaller than the journalist LLM, to serve as a general reader. This reader LLM reads the article written by the journalist and takes notes for giving feedback. As a less proficient model, the reader LLM needs material that is easily understandable to take comprehensive notes. Therefore, the more accessible the written article is, the greater the clarity and accuracy of the reader's notes will exhibit.

LLMs have shown the capability of evaluating the quality of text (Chan et al., 2023; Zheng et al., 2024a; Desmond et al., 2024). Therefore, we let an editor (the third LLM) evaluate the correctness and comprehensiveness of the reader's notes and then provide suggestions for the revision of the journalist's article. The journalist then revises the article based on the suggestions. By this iterative and parameter-free tuning process, the popular science article is enhanced and made more accessible to a general audience. To the best of our knowledge, our work is the first study on LLM collaboration for ASJ.

To assess our proposed method, we employ both automatic metrics and human evaluation on measures including readability, information conveyance, authenticity, and interestingness of our generated articles. Compared with other methods, including those with fine-tuning and prompting on various LLMs, our proposed method achieves the highest readability while remaining competitive on the other measures. We also provide a detailed analysis, including ablation studies of removing the editor LLM, removing the reader LLM, or removing both, as well as trend analysis and case studies, to offer a comprehensive understanding on LLMs in the ASJ task.

In brief, we make the following contributions:

- An ASJ framework with collaborative LLMs, generating content of high readability.
- Comprehensive experiments, analysis, and insights for applying LLMs in ASJ.

2 Related Work

Automatic Science Journalism. In recent years, some researchers have explored the application of LLM on authoring scientific articles (Wang et al., 2024a; Baek et al., 2024; Wang et al., 2024b). Dangovski et al. (2021) created a parallel corpus and provided a sequence-to-sequence method to generate news summaries from scientific articles. Goldsack et al. (2022) released two corpora, focusing on the biomedical and life science domains. Cardenas et al. (2023) constructed a dataset in various scientific fields and integrated the discourse structure of papers with their metadata to guide the generation. These methods of fine-tuning on small models can provide a good match with the reference, but there is still room for improvement in readability. In this work, we present an approach that integrates LLMs as agents to iteratively enhance readability.

Large Language Models. Our study involves three aspects related to LLMs. First, various studies show strong abilities of LLMs in writing scientific content (Pu et al., 2024; Kumar et al., 2024; Lee et al., 2024). Second, LLMs have demonstrated remarkable intelligence in social simulation (Park et al., 2023; Ziems et al., 2024) and real-world tasks (Liu et al., 2023; Chen et al., 2023; Ding et al., 2023; Qian et al., 2024; Qin et al., 2024). Inspired by this collection of work, we utilize LLMs as communicative agents to make content accessible to the general audience through a process resembling real-world practice. Third, previous studies have demonstrated that LLMs can be iteratively optimized by in-context learning without internal parameter tuning (Yang et al., 2023; Zhao et al.,

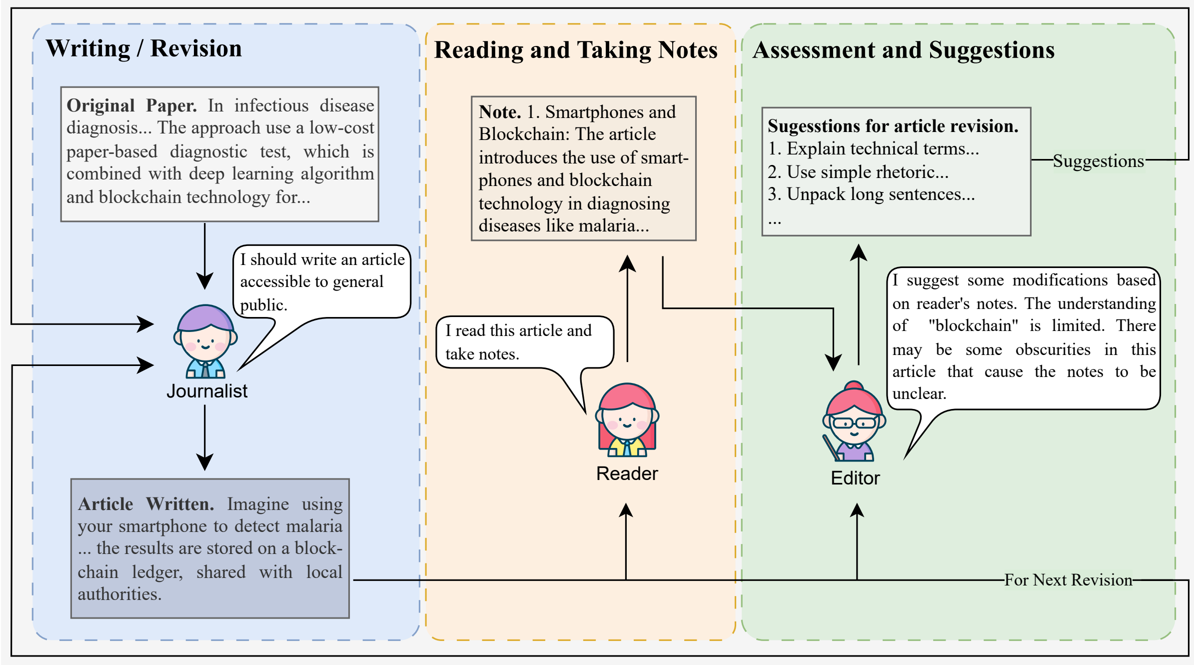


Figure 2: Overview of our iterative ASJ framework, JRE-L.

2024; Senel et al., 2024; Chen et al., 2024). Thus, this work proposes a parameter-free tuning framework that iteratively improves writing through interactions among LLMs.

3 Methodology

Following ASJ (Dangovski et al. (2021); Goldsack et al. (2022); Cardenas et al. (2023)) aims to automatically distill a scientific paper into an article accessible to a broader audience. Our ASJ framework JRE-L employs an iterative workflow of writing, reading, suggestion-making, and revision among three LLMs, as illustrated in Figure 2. All prompts for each LLM agent are listed in Appendix A.

3.1 The LLM Journalist

LLMs have shown strong writing abilities (Yuan et al., 2022; Wasi et al., 2024). Thus, they are promising tools for rewriting a given paper into a more accessible version. Following established strategies (Zheng et al., 2024a; Zhang et al., 2024), we start with prompting an LLM to assume the role of a journalist. Subsequently, the LLM journalist $\mathcal{J}$ is prompted that, given the paper x, its task is to compose an article p_0 for the general public.

$$p_0 = \mathcal{J}(x) \tag{1}$$

where journalist $\mathcal{J}$ is initialized from an LLM by a task prompt, one-shot demonstration, and fine-tuning.

3.2 The LLM Reader

In our initial attempts, we asked an LLM to directly assess the readability of the generated article. However, the results were unsatisfactory, probably due to the gap between human and model perceptions of reading difficulty. As illustrated in the two text boxes at the bottom of Figure 3, LLMs regarded both pieces of writing at a similar level of readability, as they all incorporated essential information, even if the terminology "low-cost paper-based microfluidic diagnostic tests" on the left side was not clearly explained. However, a human reader perceived the writing on the right is as more accessible.[1]

To address this readability assessment problem, we design a separate reader LLM to read the content and generate notes. Our idea is inspired by the accumulation of errors, a common phenomenon in pipeline systems (Caselli et al., 2015; Wu et al., 2018; Dziri et al., 2023). Specifically, we utilize the propagation from the textual readability of the journalist's article to the reader's comprehension in the writing-reading pipeline to induce the readability of the journalist's generated article.

Different from the LLM journalist, the reader LLM is of a smaller scale and thus has weaker reading comprehension skills, simulating a general

[1]We briefly document other unsuccessful attempts in Appendix B.

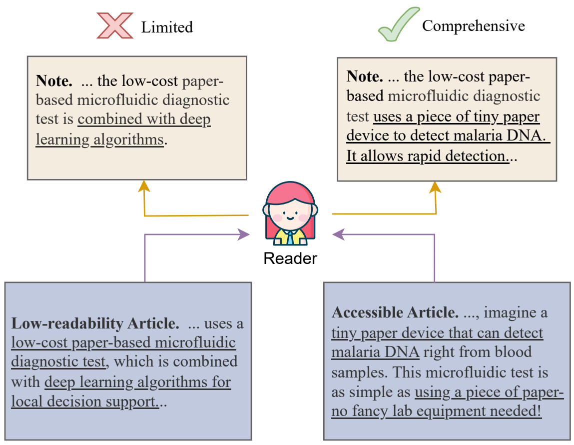


Figure 3: Accessible content helps the reader take comprehensive notes.

reader with limited domain knowledge. Once presented with the article crafted by the journalist, the reader LLM is employed to read the article and take notes. Specifically, we instruct the reader LLM to explain key terms in the article by extracting the explanations, if present, directly from the article or offering explanations for these terms, otherwise.

$$r_i = \mathcal{R}\left(p_{i-1}\right) \tag{2}$$

where r_i is the notes taken by the reader $\mathcal{R}$ on the current version of writing from the journalist $\mathcal{J}$, $i = 1, 2, \ldots, n$, and n is the number of iterations.

Intuitively, if the article is more accessible, the reader's notes will be more comprehensive. For instance, in Figure 3, when the piece (left) lacks a detailed explanation of the term "paper-based microfluidic diagnostic test", the reader will only note this term as being "combined with deep learning". When the article (right) explains the usage and advantages of this text in more detailed and plain language, the reader LLM produces better notes. Through this readability propagation from the journalist's article to the comprehensiveness of the reader's notes, the editor LLM can better recognize issues in the journalist's writing and then provide suitable suggestions for revision.

3.3 Automated Suggestions and Revisions

LLMs have demonstrated strong capabilities in serving as evaluators, widely utilized in various generative tasks (Chan et al., 2023; Zheng et al., 2024a; Desmond et al., 2024). Therefore, we employ an LLM as an editor for automated evaluation of reader comprehension and providing recommendations for article enhancement. Given the article from the journalist and notes from the reader, the LLM editor $\mathcal{E}$ is tasked with assessing the quality of the reader's notes r and identifying issues in the journalist's writing p_{i-1} that may lead to reading obstacles. Next, the editor $\mathcal{E}$ offers suggestions s_i for the journalist's content revision.

Algorithm 1 JRE - L for ASJ

INPUT: a scientific paper $\boldsymbol{x}$; journalist $\mathcal{J}$; reader $\mathcal{R}$; editor $\mathcal{E}$; number of iterations n.
OUTPUT: a news article $\boldsymbol{p}_n$ for the general audience.

```
p_0 ← J(x)                         ▷ Initial writing
for i = 1 to n do
    r_i ← R(p_{i-1})               ▷ Reader's notes
    s_i ← E(p_{i-1}, r_i)          ▷ Editor's suggestions
    p_i ← J(x, p_{i-1}, s_i)       ▷ Revision
end for
Return p_n
```

$$s_i = \mathcal{E}\left(p_{i-1}, r_i\right) \tag{3}$$

For example, in Figure 2, the editor finds that the reader's understanding of the term "blockchain" is limited, possibly due to an insufficient explanation in the reading material. To address this perceived issue, the editor suggests that the article should "explain technical terms." These suggestions are then incorporated into the instructions that will guide the journalist in revising the article. Subsequently, with the strong ability to follow instructions, the journalist LLM $\mathcal{J}$ rewrites the article based on the suggestions:

$$p_i = \mathcal{J}\left(x, p_{i-1}, s_i\right) \tag{4}$$

Then, the revised piece is fed to the reader for reading and taking notes (Equation 2) to continue the process. By an iterative cycle encompassing writing, note-taking, suggesting modifications, and revision among three LLMs, the article tailored for the general readership undergoes steady enhancement. Algorithm 1 presents this process.

4 Experiments

4.1 Setting

Datasets. We use three publicly available corpora in different disciplines as benchmarks, namely *SCITech* (Cardenas et al., 2023), *eLife*, and *PLOS* (Goldsack et al., 2022). For a fair comparison, the data split strategy is the same as that in these previous studies. Appendix C briefly introduces these datasets.

Methods For Comparison. We study models with different sizes, a number of one-shot demonstra-

Approach	SCITech			eLife			PLOS			Avg.↓	Impr.(%)↑
	CLI↓	FKGL↓	DCRS↓	CLI↓	FKGL↓	DCRS↓	CLI↓	FKGL↓	DCRS↓		
LLaMA-2-7B	15.13	13.79	10.38	15.16	14.03	10.50	15.36	14.28	10.54	13.24	19.26††
Mistral-7B	14.90	13.54	10.82	14.61	11.72	10.85	15.38	11.98	11.21	12.78	16.35††
Qwen-1.5-7B	14.77	13.50	10.72	14.72	11.83	10.92	15.06	11.94	11.09	12.73	16.03††
Qwen-1.5-7B$_{OS}$	14.44	13.51	10.31	14.22	11.62	10.52	14.86	11.43	10.82	12.41	13.86††
LLaMA-3-8B	14.84	13.18	10.41	14.55	11.65	10.49	15.18	12.01	10.88	12.58	15.02††
LLaMA-3-8B$_{OS}$	14.65	13.01	10.21	14.13	11.57	10.35	14.76	11.81	10.79	12.36	13.51††
Mixtral-8x7B	13.98	13.25	10.36	14.21	12.01	10.28	15.34	11.58	10.98	12.44	14.07††
Qwen-1.5-72B	13.78	13.10	10.25	14.17	12.09	10.35	15.18	11.75	10.62	12.37	13.58††
GPT-3.5-Turbo	14.98	13.62	10.81	14.35	11.87	10.98	15.11	11.92	10.87	12.72	15.96††
GPT-4	13.48	12.13	10.14	13.96	10.87	10.11	14.86	11.78	10.47	11.98	10.77††
BART$_{FT}$	13.43	15.22	10.66	12.32	10.65	9.19	15.61	14.24	10.51	12.43	14.00††
Qwen$_{FT}$	13.37	14.79	10.48	12.15	10.63	9.12	15.54	13.95	10.58	12.29	13.02††
LLaMA$_{FT}$	13.31	14.52	10.12	12.01	10.31	9.14	15.03	13.64	10.37	12.05	11.29††
CollabStory	13.82	12.13	10.32	13.41	11.43	9.73	14.30	11.70	10.11	11.88	10.02††
ChatDev	13.51	12.21	10.49	13.07	10.92	9.77	13.91	10.97	9.92	11.64	8.16†
JRE-L$_{OS}$	12.74	10.37	9.89	11.86	10.08	9.26	12.84	10.03	9.87	10.77	0.74
JRE-L$_{FT}$	12.94	13.33	10.33	12.04	**9.85**	**9.04**	13.15	11.48	10.17	11.37	5.98†
JRE-L	**12.69**	**10.16**	**9.79**	**11.60**	10.10	9.46	**12.74**	**10.00**	**9.69**	**10.69**	0.00
Paper Abstracts	16.67	15.27	11.39	17.53	15.35	11.87	16.38	14.98	11.10	14.50	26.28††
Plain Summaries	14.23	14.79	11.13	12.52	10.91	8.94	15.90	14.76	10.91	12.68	15.69††

Table 1: Automated evaluation, including **single-LLM prompting**, **fine-tuning**, **multi-LLM prompting**, and **our JRE-L framework**. The option 'OS' and 'FT' denote 'one shot' and 'fine-tuning.' Symbols † and †† denote that the statistical significance of the comparison with JRE-L is $p < 0.05$ and $p < 0.01$, respectively.

tions and fine-tuning, as well as previous ASJ baselines and other LLM-collaboration frameworks. These methods are listed in Appendix D for brevity.

Automatic Evaluation. Following Goldsack et al. (2022); Cardenas et al. (2023), we use the Coleman-Liau Index (CLI, Coleman and Liau (1975)), the Flesch-Kincaid Grade Level (FKGL, Kincaid et al. (1975)) and the Dale-Chall Readability Score (DCRS, Dale and Chall (1948)) to automatically assess readability. CLI considers the number of sentences, words, and characters, whereas FKGL is based on the number of sentences, words, and syllables. DCRS analyzes the average sentence length and the presence of familiar words from a list of the most commonly used words[2].

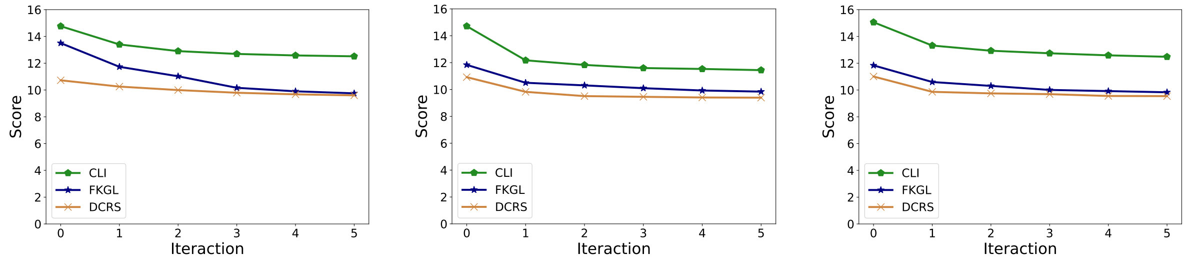
Human Evaluation. Four human participants are enlisted for evaluation. All of them are pursuing or possess master's degrees, two from computer science and two from biomedical science. We sample 10 pairs of original papers in computer science from SCITech and their corresponding popular science articles, as well as 10 pairs in biomedical science, 5 pairs from eLife and 5 pairs from PLOS. This quantity is comparable to previous studies (Goldsack et al., 2022; Cardenas et al., 2023), taking into account both the reliability of the results and the workload of the annotators.

Approach	Read.	Info.	Auth.	Intr.
Within Field				
Plain Summaries	2.95††	2.90††	3.35††	2.70††
Qwen1.5-7B	3.50†	3.35†	3.40†	3.10†
GPT-4	3.80	**3.75**	**3.80**	3.40
JRE-L	**3.95**	3.60	3.70	**3.55**
Outside Field				
Plain Summaries	2.75††	2.85††	3.25††	2.65††
Qwen1.5-7B	3.35†	3.10†	3.30†	3.10†
GPT-4	3.40	**3.55**	**3.70**	3.15
JRE-L	**3.65**	3.40	3.55	**3.20**

Table 2: Results of human evaluation. Symbols "†" and "††" indicate that the statistical significance of the comparison with JRE-L is $p < 0.05$ and $p < 0.01$. The higher the scores, the better.

Four representative methods are chosen for human evaluation: (1) plain summaries by human writers, (2) Qwen1.5-7B generation, (3) GPT-4 generation, and (4) generation by our JRE-L. The human evaluation encompasses multiple dimensions, namely Readability (Read.), Information Conveyance (Info.), Authenticity (Auth.), and Interestingness (Intr.). Participants are tasked with evaluating the articles using a 1-5 Likert scale (Likert, 1932), grounded on specific questions. Each participant is assigned to assess all articles both in

[2] https://help.readable.com/en/article/dale-chall-words-list-w877fe



(a) SCITech (b) eLife (c) PLOS

Figure 4: Performance improvement over iterations, with iteration 0 producing the initial writing.

Approach	SCITech↓	eLife↓	PLOS↓	Avg.
JRE-L	**10.88**	<u>10.39</u>	10.81	**10.69**
Reader→7B	<u>10.95</u>	10.44	**10.79**	10.73
LLM→LLaMA	10.99	**10.37**	<u>10.81</u>	<u>10.72</u>
– Reading	11.39	11.00	11.40	11.26
– Suggestions	11.44	11.03	11.49	11.32
– Collaboration	11.74	11.29	11.75	11.59

Table 3: Ablation results, reporting average scores. The scores in bold are the best, and the underlined scores are the second best.

the field they are familiar with (Within Field) and those they are not familiar with (Outside Field), to provide a genuine evaluation from readers within the specific discipline and general readers. Appendix E shows the details of these measures and the questionnaire form.

Hyperparameters. We list hyperparameters in Appendix C for brevity.

4.2 Automatic Evaluation

Different Models and Sizes. Table 1 show the comparison of different methods. Recent LLMs of similar scales have shown comparable performance, surpassing LLaMA-2. Larger models such as Mixtral-8x7B and Qwen1.5-72B show even better performance, indicating that performance improves as the model scale increases. Additionally, the formidable LLM GPT-4 outperforms all the other single LLMs.

One-shot. Among the methods of single-LLM prompting, one-shot demonstration brings some improvements of the performance (Avg.) by 0.32 in Qwen and 0.22 in LLaMA. In our LLM framework, adding one-shot examples as a warm start shows close performance to the original version. A possible reason is that these examples can be used as references for the first revision and the iterative process narrows the performance gap.

Fine-tuning. The fine-tuning methods exhibit competitive performance, slightly better than prompting the open-source LLMs. Fine-tuning models with a larger scale (Qwen1.5-7B and LLaMA-3-8B) outperforms BART. Interestingly, in our JRE-L framework, fine-tuning the journalist LLM to warm up does not lead to a noticeable improvement, possibly because the LLM does not follow instructions as well as before after such specialized fine-tuning, thereby affecting cooperation among LLMs.

Previous baselines. Previous ASJ work (Goldsack et al., 2022; Cardenas et al., 2023) took BART as their baseline. Compared to their fine-tuning method, our framework brings an average of $(12.43 - 10.69)/12.43 = 14\%$ improvement on all the three datasets. Moreover, compared with lay summaries written by humans, our method shows higher readability, as those summaries were crafted for technical audiences.

LLM-collaboration frameworks. We adapt two other LLM-collaboration methods to the ASJ task for comparison. Our methods attain average enhancements of 8.16% and 10.02% over them respectively. This performance improvement is due to the reader's feedback in our framework, which can better expose writing issues and is not available in the other two paradigms. This experimental result also demonstrates the effectiveness of our proposed iterative workflow.

4.3 Human Evaluation

Our human participants assessed articles relevant to their fields as well as those in their unfamiliar fields. The results are reported in Table 2, where the scores are obtained by averaging scores on the corresponding metric and are categorized into within-field and outside-field articles. The Krippendorff's alpha value (Krippendorff, 2011) between annotators is 0.52, slightly lower than the 0.57 value in Cardenas et al. (2023), indicating an acceptable agreement. Notably, participants assign lower ratings to articles

Scientific Paper. In infectious disease diagnosis, ***results need to be communicated rapidly to healthcare professionals once testing has been completed so that care pathways can be implemented***. This represents a particular challenge when testing in remote, low-resource rural communities, in which such diseases often create the largest burden. Here, we report a ***smartphone-based end-to-end platform for multiplexed DNA diagnosis of malaria***. The approach uses a low-cost paper-based microfluidic diagnostic test, which is combined with deep learning algorithms for local decision support and blockchain technology for secure data connectivity and management. We validated the approach via field tests in rural Uganda, where it correctly identified more than 98% of tested cases. Our platform also provides secure geotagged diagnostic information, which creates the possibility of integrating infectious disease data within surveillance frameworks.
Article Written by JRE-L. In the fight against malaria, ***timely diagnosis is crucial***, especially in rural communities where these diseases often hit hardest. Researchers have come up with a brilliant solution using smartphones! It's a ***smartphone-based system that combines a low-cost test with deep learning and blockchain technology***. First, ***imagine a tiny paper device that can detect malaria DNA right from blood samples***. ***This microfluidic test is as simple as using a piece of paper - no fancy lab equipment needed!*** Then, the results are analyzed by smart algorithms on your phone. ***These "AI helpers" give instant feedback to healthcare workers without delay.*** All this diagnostic data is stored securely on blockchain. This means information is protected and cannot be tampered with. In Uganda, field tests showed that this system was almost perfect - it correctly identified 98% of cases! ***The best part? The platform also records each diagnosis's location through geotags.*** This opens up the possibility to ***connect local disease data to bigger surveillance networks, helping us track and combat malaria more efficiently.*** So next time you hear about remote villages facing health challenges due to lack of fast diagnosis tools, remember: there's hope with these smartphone-powered innovations!

Table 4: An example of the original paper and the article generated by our method. We highlight statements that we regard ***easy to understand*** and ***explanations for technical details***.

outside their expertise than those within their expertise areas, indicating the effect of topic familiarity on reader experience. Nevertheless, the relative performance between methods appears consistent across both within expertise and outside expertise settings.

Interestingly, all LLM-based methods outperform the plain summaries written by humans ($p < 0.01$), probably because these summaries are still too technical for readers, leading to lower ratings. The LLM-generated contents, targeted at general audience, are easier for both in-domain and out-of-domain readers to read. Furthermore, our JRE-L surpasses the single Qwen in all dimensions ($p < 0.05$), demonstrating the effectiveness of LLM collaboration. Our JRE-L achieves high ratings in all dimensions, close to those of GPT-4, even though our model is much smaller. Collectively, these findings attest to the effectiveness of our proposed approach.

5 Analysis

5.1 Performance Over Iterations

As our system is an iterative enhancement framework, we study its performance over iterations. The entire writing-reading-suggestion-revision process is carried out for five iterations, with readability scores depicted in Figure 4. The figure shows a pronounced decline in reading difficulty in the first iteration and the decline continues in the next two iterations. This pattern demonstrates the efficacy of our iterative revision methodology. Following the third iteration, the performance levels off, indicating diminishing improvements from subsequent suggestion and editing efforts.

5.2 Ablation Study

We conduct an ablation experiment to evaluate the effectiveness of each component by removing or replacing components. Table 3 presents the averaged results on each dataset, and detailed results are included in Appendix F. In the "*Reader→7B*" setting, we substitute the 1.8B reader model with the 7B version. This substitution leads to a minor performance fluctuation. On one hand, the 7B model tolerates low readability of content and highlights fewer writing issues. On the other hand, it excels in instruction following, enhancing task execution and reducing intermediate errors in the workflow. The dynamics between gains and losses render the 7B reader comparably advantageous at times and disadvantageous at others. As such, we recommend using the 1.8B model for its higher resource efficiency and throughput. To test the generalizability of our framework, we replace each agent with LLaMA models of 3-8B in the "*LLM→LLaMA*" setup. Experimental results show that the performance after replacement is close to that based on Qwen, demonstrating that our framework can be adapted to different LLMs.

In $-$*Reading Notes*, we eliminate the requirement for the reader LLM to read the article and make notes. Instead, we have the editor LLM of-

1st Iteration. 1. The mobile lab on your phone is a portable, paper-based test that uses deep learning algorithms to analyze blood samples for the presence of malaria. 2. The end-to-end system combines a smartphone app, allowing healthcare workers to make informed decisions about treatment without delay. 3. The smartphone app uses ***deep learning algorithms to interpret the test results***, providing accurate diagnoses over 98% in field tests. 4. The platform ensures secure data management by using ***blockchain technology, which records transparently and cannot be tampered with***.
2nd Iteration. 1. Smartphones and Blockchain: The article introduces the use of smartphones and ***blockchain technology in diagnosing diseases like malaria, particularly in remote areas where traditional diagnostic methods may not be available or accessible***. 2. Malaria Diagnosis: The low-cost paper-based microfluidic diagnostic test uses a piece of tiny paper device to detect malaria DNA. It allows rapid detection with the support of AI systems. 3. Low-Cost Test: The AI-based test is referred to as a "smart decision support" system, which uses advanced technology such as artificial intelligence to interpret the results right on the user's device. 4. ***Artificial Intelligence (AI): The AI acts like a smart detective, quickly identifying potential cases with an accuracy rate of 98% in real-world tests***, enabling healthcare workers to make informed decisions about treatment without needing fancy labs or waiting too long.
3rd Iteration. 1. Smartphone: A device with advanced features such as cameras, sensors, and processors that can perform various tasks, including medical diagnosis. 2. Blood test: A diagnostic procedure where blood is collected from a finger or other body part through a small needle and analyzed for specific signs of malaria (e.g., anemia or fever). 3. ***AI as personal health assistant: An artificial intelligence system designed to assist individuals in managing their health by providing personalized recommendations based on medical data analysis (e.g., detecting early signs of disease).*** 4. Secure digital diary: A digital record containing sensitive information about an individual's health status stored on a secure blockchain network to ensure accuracy and prevent unauthorized access or manipulation. 5. ***Blockchain technology: A distributed ledger system that allows secure sharing of data across multiple parties without the need for intermediaries or centralized authorities (e.g., storing patient records in hospitals). In this case, it serves as the secure digital ledger for tracking where diseases are spreading in remote areas due to malaria prevalence.***

Table 5: Notes taken by the reader, ***two technical terms and their explanations*** are highlighted. The notes become more detailed and comprehensive through the iterative process.

fer suggestions directly, matching the situation in Figure 1 (a). In −*Suggestions*, the editor is omitted, and the journalist revises the article based on the reader LLM's reading. In −*Collaboration*, the journalist revises the article based on the previous writing without any input from the reader or the editor. As depicted in Table 7, our approach exhibits a decrease in performance when each module and the collaboration is removed, underscoring the significance of each module and the collaboration.

5.3 Writing Case Analysis

To facilitate an intuitive assessment of our method, we present a case study on one writing sample. As shown in Table 4, our method can generate articles that are more readable, with concise and vivid narratives, along with explanations for technical details. For instance, our generated article states that "timely diagnosis is crucial" rather than "results need to be communicated rapidly to healthcare professionals once testing has been completed so that care pathways can be implemented", making it more brief and accessible for readers to grasp the research objective. Moreover, our generated article details that the proposed system is a "smartphone-based system that combines a low-cost test with deep learning and blockchain technology" rather than "smartphone-based end-to-end platform for multiplexed DNA diagnosis of malaria", enhancing the comprehensibility for a broader audience.

5.4 Case of Reading Notes

We further present a case study on the notes taken by the reader LLM during the first three iterations. As shown in Table 5, detailed explanations for technical terms increase with iterations. For example, the technical terms "AI" and "blockchain" are more thoroughly explained in the third iteration than in the first two iterations. This phenomenon suggests that as the readability of writing improves, readers can understand the content better and provide more elaborate explanations, aligning with human reading behaviors.

5.5 Suggestions for Revision

We further examine the suggestions given by the editor LLM for the revision over the iterative process, as shown in Table 6. In the initial stage, the editor offers general advice. Following the revision, the advice becomes more specific and detailed. Subsequent rounds of suggestions highlight specific content in the article and recommend revisions in various aspects. In the 3rd iteration, the feedback from the editor suggests on phrases in the previous version. This revision behaviour is similar to real-world science journalism, showcasing the efficacy of our framework and the social intelligence of LLMs.

1st Iteration. 1. Simplify technical terms 2. Break down processes 3. Emphasize benefits 4. Conclusion statement	
2nd Iteration. 1. **Emphasize technology**: Instead of listing specific platforms (Ethereum or Hyperledger Fabric), mention that a secure digital ledger is used without specifying which one. 2. **Simplify medical terms**: For "Plasmodium falciparum," provide a simpler explanation or use a common term like "malaria parasite type." 3. **Break down process**: Explain how smartphone analysis works in more detail - perhaps by describing how it compares to traditional lab methods. 4. **Connect to equity**: Highlight how this technology addresses health disparities by providing quick diagnosis in remote areas.	
3rd Iteration. 1. **Emphasize simplicity**: For accessibility, rephrase "low-cost paper-based microfluidic diagnostic test" as "affordable, easy-to-use test with a paper strip." 2. **Explain AI in simpler terms**: Instead of "AI instantly interprets results," say "The smartphone app quickly analyzes the data to give a diagnosis." 3. **Break down data security**: Highlight that information is stored securely on a phone or cloud server with strong passwords or encryption. 4. **Quantify success**: Mention that 98% accuracy rate is exceptional but could be framed as an impressive achievement ("This system detected almost all cases correctly!"). 5. **Cite real-life impact**: Share examples of how this technology has made a difference in remote communities to connect it emotionally with readers.	

Table 6: Suggestions provided by the editor LLM are becoming increasingly specific through the iterative process.

6 Conclusion

This study proposes the collaboration of LLMs in the loop for ASJ aimed at general readers. Initially, an LLM functions as a journalist by composing an explanatory article for the general public. Subsequently, another LLM, acting as a general audience, reads these articles and takes notes, helping reveal the readability issues in the generated article. Then, an editor LLM assesses the reader's notes and offers suggestions for improvement. Following the suggestions, the journalist revises its article, and then passes it to the reader to continue the iterative process. Extensive experiments are conducted to evaluate the effectiveness of our framework, including both automatic and human evaluation. In comparison to prompting and fine-tuning LLMs as other ASJ systems do, our method achieves the highest readability while maintaining high quality.

Limitations and Future Work

We identify the following five limitations of our work. First, following previous work, we have defined ASJ as the process of transforming a single paper into an article intended for a general audience. In practice, a popular science article may encompass multiple studies. Therefore, an extension can be the consolidation of several papers into a single article. Second, we have utilized some statistical indexes for automatic assessment, but these statistical measures may miss semantic information. LLM-driven evaluation could offer a solution. While there remains a gap between LLMs and humans in evaluating text on readability and authenticity, efforts such as human-preference optimization could be made to minimize this gap. Third, given our exploratory approach in utilizing LLMs for ASJ, we strategically chose abstracts as opposed to full papers as input to maintain both simplicity and resource efficiency. Nevertheless, long-context ASJ is an intriguing task with higher impact. Fourth, due to the limit of space and effort, we study our framework on settings with up to three LLMs. It could be interesting to study collaborative writing between writers, receiving feedback from multiple readers of different backgrounds, and considering revision suggestions from a hierarchy of editors or editors from different areas. Lastly, all components of our framework are powered by LLMs. In addition to our efforts to make each LLM simulate humans, it will be interesting to incorporate genuine human preferences to enhance the generated content.

Ethics Statement

The experiments in this study were conducted on publicly available datasets. Any information involving privacy was removed. All annotators have been properly paid for their efforts.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments, which help improve this work. This research was partially supported by the Guangzhou Municipality Big Data Intelligence Key Lab (2023A03J0012).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, and Sam Altman et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Stuart Allan. 2011. Introduction: Science journalism in a digital age. *Journalism*, 12(7):771–777.

Christopher William Anderson, Emily Bell, and Clay Shirky. 2015. Post-industrial journalism: Adapting to the present. *Geopolitics, History and International Relations*, 7(2):32.

Martin Angler. 2017. *Science journalism: an introduction*. Routledge.

Tal August, Kyle Lo, Noah A Smith, and Katharina Reinecke. 2024. Know your audience: The benefits and pitfalls of generating plain language summaries beyond the" general" audience. *arXiv preprint arXiv:2403.04979*.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Sara E Brownell, Jordan V Price, and Lawrence Steinman. 2013. Science communication to the general public: why we need to teach undergraduate and graduate students this skill as part of their formal scientific training. *Journal of undergraduate neuroscience education*, 12(1):E6.

John Bryant. 2002. *The fluid text: A theory of revision and editing for book and screen*. University of Michigan Press.

Ronald Cardenas, Bingsheng Yao, Dakuo Wang, and Yufang Hou. 2023. 'don't get too technical with me': A discourse structure-based framework for automatic science journalism. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1186–1202.

Tommaso Caselli, Piek Vossen, Marieke van Erp, Antske Fokkens, Filip Ilievski, Rubén Izquierdo, Minh Le, Roser Morante, and Marten Postma. 2015. When it's all piling up: investigating error propagation in an nlp pipeline. In *WNACP@ NLDB*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Dake Chen, Hanbin Wang, Yunhao Huo, Yuzhao Li, and Haoyang Zhang. 2023. Gamegpt: Multi-agent collaborative framework for game development. *arXiv preprint arXiv:2310.08067*.

Sijia Chen, Baochun Li, and Di Niu. 2024. Boosting of thoughts: Trial-and-error problem solving with large language models. In *The Twelfth International Conference on Learning Representations*.

Kwangsu Cho and Charles MacArthur. 2011. Learning by reviewing. *Journal of educational psychology*, 103(1):73.

Greg Clark, Josh Russell, Peter Enyeart, Brant Gracia, Aimee Wessel, Inga Jarmoskaite, Damon Polioudakis, Yoel Stuart, Tony Gonzalez, Al MacKrell, et al. 2016. Science educational outreach programs that benefit students and scientists. *PLoS Biology*, 14(2):e1002368.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Rumen Dangovski, Michelle Shen, Dawson Byrd, Li Jing, Desislava Tsvetkova, Preslav Nakov, and Marin Soljačić. 2021. We can explain your research in layman's terms: Towards automating science journalism at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12728–12737.

Michael Desmond, Zahra Ashktorab, Qian Pan, Casey Dugan, and James M Johnson. 2024. Evalullm: Llm assisted evaluation of generative outputs. In *Companion Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 30–32.

Shiying Ding, Xinyi Chen, Yan Fang, Wenrui Liu, Yiwu Qiu, and Chunlei Chai. 2023. Designgpt: Multi-agent collaboration in design. In *2023 16th International Symposium on Computational Intelligence and Design (ISCID)*, pages 204–208. IEEE.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaïd Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604.

Winfried Göpfert. 2008. The strength of pr and the weakness of science journalism. In *Journalism, science and society*, pages 227–238. Routledge.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

J Peter Kincaid, RP Fishburne, RL Rogers, and BS Chissom. 1975. Derivation of new readability formulas (automated reliability index, fog count and flesch reading ease formula) for navy enlisted personnel (research branch report 8-75). memphis, tn: Naval air station; 1975. *Naval Technical Training, US Naval Air Station: Millington, TN*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Sandeep Kumar, Guneet Singh Kohli, Tirthankar Ghosal, and Asif Ekbal. 2024. Longform multimodal lay summarization of scientific papers: Towards automatically generating science blogs from research articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10790–10801.

Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A Alghamdi, et al. 2024. A design space for intelligent and interactive writing assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–35.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. 2023. Awq: Activation-aware weight quantization for llm compression and acceleration. *arXiv preprint arXiv:2306.00978*.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2023. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*.

Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date.

Joyce YM Nip. 2006. Exploring the second phase of public journalism. *Journalism studies*, 7(2):212–236.

OpenAI. 2023. Introducing ChatGPT.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Dongqi Pu, Yifan Wang, Jia Loy, and Vera Demberg. 2024. Scinews: From scholarly complexities to public narratives–a dataset for scientific news report generation. *arXiv preprint arXiv:2403.17768*.

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

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. Questeval: Summarization asks for fact-based evaluation. *Empirical Methods in Natural Language Processing*.

Lütfi Kerem Senel, Besnik Fetahu, Davis Yoshida, Zhiyu Chen, Giuseppe Castellucci, Nikhita Vedula, Jason Ingyu Choi, and Shervin Malmasi. 2024. Generative explore-exploit: Training-free optimization of generative recommender systems using LLM optimizers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5396–5420, Bangkok, Thailand. Association for Computational Linguistics.

Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Saranya Venkatraman, Nafis Irtiza Tripto, and Dongwon Lee. 2024. Collabstory: Multi-llm collaborative story generation and authorship analysis. *arXiv preprint arXiv:2406.12665*.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024a. Scimon: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 279–299. Association for Computational Linguistics.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024b. Autosurvey: Large language models can automatically write surveys. *CoRR*, abs/2406.10252.

Azmine Toushik Wasi, Rafia Islam, and Raima Islam. 2024. Llms as writing assistants: Exploring perspectives on sense of ownership and reasoning. *arXiv preprint arXiv:2404.00027*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Lijun Wu, Xu Tan, Di He, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. Beyond error propagation in neural machine translation: Characteristics of language also matter. *arXiv preprint arXiv:1809.00120*.

Zeyuan Yang, Peng Li, and Yang Liu. 2023. Failures pave the way: Enhancing large language models through tuning-free rule accumulation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1751–1777.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*, pages 841–852.

Meishan Zhang, Gongyao Jiang, Shuang Liu, Jing Chen, and Min Zhang. 2024. Llm–assisted data augmentation for chinese dialogue–level dependency parsing. *Computational Linguistics*, pages 1–24.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024a. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024b. Llamafactory: Unified efficient fine-tuning of 100+ language models. *Preprint*, arXiv:2403.13372.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Prompts for LLMs

We list all prompts in Table 10. All prompts follow a similar format. First, we assign a role to each LLM agent by a sentence. We then specify the task and background in one or two sentences. Next, we give each LLM step-by-step instructions. After that, we input the rules to be followed for each LLM. Finally, the format of the output is specified as a "markdown" style to facilitate the extraction and support the information flow among LLMs. In our preliminary study, this pattern works well in prompting various LLMs, with strong task completion performance and format adherence.

B Failed Attempts

This appendix outlines our unsuccessful attempts. We hope that it will help follow-up research. First, as a commonly used mechanism in LLM agents, reflection can support iterative enhancement by consolidating prior experiences (Yao et al., 2022; Park et al., 2023; Yang et al., 2023). In our pilot experiments, however, this approach did not succeed in refining the journalist's writing. A potential explanation could be that the ASJ task for general audience requires specific revision instructions, whereas solely summarizing prior writing experiences results in general guidance only.

Within our framework, an LLM acts as a reader, reading an article and taking notes. How about having this reader perform a reading comprehension task instead of simply taking notes? Intuitively, it can demonstrate the reader's understanding and induce content complexity. Our preliminary investigations, however, revealed that the reading comprehension approach yields less efficacy compared to the note-taking strategy. It might be that the quality of question generation greatly affects the efficiency of the reading comprehension results. Also, asking a fixed number of questions narrows the space of textual exploration, thereby restricting a comprehensive perception of content complexity.

We have tried to let an LLM (either Qwen 1.5 or GPT-3.5) do the writing-reading-editing task with

Statistics	SCITech	eLife	PLOS
# pairs	2431	4828	27525
# words^{ori}	216.8	166.3	268.3
# sentences^{ori}	5.7	6.8	10.2
# words^{pln}	176.1	347.6	175.6
# sentences^{pln}	7.9	15.7	7.8

Table 7: Statistics of benchmark datasets. Numbers of words and sentences are average values. The "ori" superscript denotes abstracts of original papers, and "pln" represents plain summaries written by humans.

a single prompt, but the LLM did not seem to follow the instructions successfully, e.g., it generated writing without giving or taking any feedback. We suspect that these models have limited ability to follow complex instructions.

Next, we briefly discuss the modules in our framework that may go wrong at runtime. We observe that, in some cases, one of the LLMs did not follow editing instructions but just simply copied the input to the generated article. This problem may be solved as the capabilities of the model increase.

C Datasets and Hyperparameters

Table 7 shows some statistics of the three datasets used in our experiments. *SCITech(News)* is released by Cardenas et al. (2023), who gathered press releases from ACM Technews as well as their source articles from various publishers, involving fields of computer science, engineering, astrophysics, biology, and others. *eLife* is an open-access journal that focuses on biomedical and life sciences. Goldsack et al. (2022) collected some eLife articles as well as digests written by journal editors based on both the article itself and questions answered by the author. Similarly, *PLOS* hosts journals across areas of science and medicine. Some of these articles, also collected by Goldsack et al. (2022), come with the author's summary. For resource saving, the original paper's abstract serves as the scientific content input. The same as previous studies, in *SCITech*, we use 1,431 instances for training and validation, and the remaining 1,000 for testing, and each of *eLife* and *PLOS* datasets is separated into training, validation, and testing splits at a ratio of 90%/5%/5%.

Our local LLM service runs on a machine with eight GTX 4090 GPUs. We utilize the Huggingface platform (Wolf et al., 2019) for downloading and loading checkpoints. For rapid inference and memory efficiency, we utilize the vLLM library[3] to develop API services. We deploy agents from the Qwen-1.5 series, for their good performance and diverse model scales (Bai et al., 2023). In particular, Qwen-1.5-7B is employed for the steps of writing, providing suggestions, and revision, whereas Qwen-1.5-1.8B serves as the reader for taking notes. To improve memory efficiency, we implement activation-aware weight quantization (AWQ, Lin et al. 2023) for model quantization. For one-shot learning settings, we randomly take one sample from the parallel datasets as a demonstration injected in the prompt for the journalist LLM. For fine-tuning, we utilize LoRA (Hu et al., 2021) with Llama-Factory (Zheng et al., 2024b), adopting the default setting with the number of epochs set to 10. We use the default temperature setting and empirically set top_p to 0.4, both frequency penalty and repetition penalty to 1, ensuring the stability of the LLMs' output while retaining diversity. We iterate five times and empirically select the output from the third iteration as the final result, for a balance of performance and cost-effectiveness. The maximum number of tokens in the model output is 4,096. Our preliminary tuning showed that such hyperparameters exhibited a relatively good and robust performance in various LLMs. For the balance of fairness and resource expenditure, all approaches in our comparison study share the same set of hyperparameters setting.

D Methods Under Comparison

We list detailed descriptions of all methods under comparison in Table 8.

E Details of Human Evaluation

Automatically assessing content authenticity and informativeness has been a challenging task. Cardenas et al. (2023) used QuestEval (Scialom et al., 2021) to assess the faithfulness of ASJ-generated content, yet the results exhibited significant variances. Therefore, human evaluation remains the main method for such assessments. We created a questionnaire for human evaluation using a 1-5 Likert scale, as shown in Figure 5. All participants were informed that their assessments would be used for research purposes. We utilized Label Studio (Tkachenko et al., 2020-2022) to construct the annotation platform. Initially, participants indicated their familiarity with the given topic. They

[3] https://docs.vllm.ai

Approach	Description
LLaMA-2-7B, LLaMA-3-8B	Two generations of LLaMA (Touvron et al., 2023; Meta, 2024) are tested for comparison. We also adopt one-shot learning (OS) and fine-tuning (FT) on LLaMA-3-8B, as well as test these variants in our collaborative framework.
Mistral-7B, Mixtral-8x7B	We test two scales of Mistral (Jiang et al., 2023) to investigate the impact of model size.
Qwen-1.5-7B, Qwen-1.5-72B	We use three implementations of Qwen-1.5-7B (Bai et al., 2023) for testing, i.e., the original version, with one-shot demonstration, and after fine-tuning, for a systematic study. The 72B version is for testing the impact of model size.
GPT-3.5-Turbo, GPT-4	We test the performance of two closed LLMs, GPT-3.5-Turbo-1106 (OpenAI, 2023), and GPT-4-1106-preview (OpenAI, 2023).
BART	Previous ASJ studies (Goldsack et al., 2022; Cardenas et al., 2023) took BART (Lewis et al., 2020) as baselines. We also include it in our study.
CollabStory	Venkatraman et al. (2024) proposed a method that involves three individual LLMs to generate the beginning, middle, and end of stories in sequence. We adopt this collaborative paradigm for ASJ to compare with our framework.
ChatDev	Qian et al. (2024) introduced an LLM-powered software development framework, where LLM-driven agents communicate with each other through a prompt-based workflow, including design, coding, and testing. We adapt this framework to ASJ by changing phases to outline design, writing, and reviewing and editing. As each phase involves an instructor and an assistant, there are six agents in total, doubling the number of our framework. We take Qwen-1.5-7B as each agent for a fair comparison.
JRE-L	We test three versions of our framework. The first is 2×Qwen1.5-7B+Qwen1.5-1.8B. The second is with the journalist replaced by a one-shot demonstrated version as a warm start (OS). The third uses a fine-tuned journalist (FT) as a warm start.

Table 8: Description of methods under comparison.

were then tasked with answering four questions on Readability, Information Conveyance, Authenticity, and Interestingness respectively. Readability assesses how easily the article can be read. Information conveyance determines how the rewritten content accurately and comprehensively conveys the information from the original paper. Similarly, authenticity assesses the correctness of the content, as a high-quality article should contain minimal factual or common sense errors. Finally, the level of interestingness is also a crucial factor; content of high appeal will provide a better reading experience.

F Ablation Results

Table 9 presents the detailed results of the ablation study, demonstrating the effectiveness of our proposed framework. Our framework exhibits stability when component LLMs are replaced by different models. Additionally, upon the removal of one or more modules of our framework, the performance undergoes a significant decline, highlighting the efficacy of our framework design. Moreover, we analyze the impact of changing LLMs to the latest Qwen version (2.5) and LLaMA (3.2) in a complementary experiment, as main studies including automatic evaluation and human evaluation have been finished before the latest models were released. The result shows that newer models can improve the performance, but the improvement is limited, possibly due to marginality through the iterative process.

G Use of AI Assistants

We use ChatGPT for correcting grammar and improving expressions in this manuscript.

Approach	SCITech			eLife			PLOS			Avg.↓
	CLI↓	FKGL↓	DCRS↓	CLI↓	FKGL↓	DCRS↓	CLI↓	FKGL↓	DCRS↓	
JRE-L	**12.69**	**10.16**	9.79	**11.60**	10.10	9.46	12.74	10.00	**9.69**	**10.69**
Reader→7B	12.81	10.35	**9.68**	11.82	10.01	9.51	12.67	**9.93**	9.78	10.73
LLM→LLaMA	12.77	10.24	9.97	11.95	**9.83**	**9.34**	**12.51**	10.07	9.84	10.72
− Reading	13.21	10.63	10.33	12.22	10.78	10.02	13.35	10.59	10.25	11.26
− Suggestions	13.25	10.69	10.39	12.17	10.83	10.08	13.31	10.74	10.42	11.32
− Collaboration	13.50	11.01	10.71	12.47	10.99	10.41	13.65	10.91	10.70	11.59
LLM→Qwen-2.5	12.51	10.02	**9.52**	11.42	**9.89**	9.23	12.53	9.81	**9.53**	10.50
LLM→LLaMA-3.2	**12.37**	**9.88**	9.56	**11.21**	9.95	**9.01**	**12.42**	**9.64**	9.61	**10.41**

Table 9: Results of the ablation study.

Here is the original paper.

{{Original Paper}}

Here is the rewritten article.

{{Rewritten Article}}

You are assigned to assess the rewritten article. Please select the choice that matches your thought:

1. How well do you know about the topic of this content?
   [1] I have never heard about this topic before
   [2] I have limited knowledge about this topic
   [3] I am somewhat familiar with this topic
   [4] I have good knowledge and understanding of this topic
   [5] I have written research papers on this topic
2. How well does this article convey the information of the original paper correctly?
   [1] Very poorly
   [2] Poorly
   [3] Moderately
   [4] Well
   [5] Very well
3. How accurate is this article (without factual or common sense errors)?
   [1] Very inaccurate
   [2] Inaccurate
   [3] Neutral
   [4] Accurate
   [5] Very accurate
4. How easy or difficult was it for you to read the article?
   [1] Very difficult to read
   [2] Somewhat difficult to read
   [3] Neutral/Neither easy nor difficult to read
   [4] Somewhat easy to read
   [5] Very easy to read
5. How interesting did you find the article?
   [1] Not interesting at all
   [2] Slightly interesting
   [3] Moderately interesting
   [4] Quite interesting
   [5] Very interesting

Figure 5: The questionnaire for participants to evaluate articles.

Journalist. You are a science journalist for general audiences. Given a paper's summary, you are assigned to rewrite it into a short understandable article for general audiences.
Follow the rules strictly:
- Keep short yet informative.
- The output format:
Article
...

Reader. You are a general reader. Given a popular science article, please read it carefully and take some notes.
Please take the following steps:
1. First, extract all technical terms with their context from the article.
2. Then, explain the technical terms based on their context.
Follow the rules strictly:
- Extraction should mention the specific location of each technical term in the article.
- Explanation should be first extracted from the article; if not found, it can be some common-sense or specialized knowledge.
- Extraction and explanation should be in points, like "1...2...3...".
- The output format:
Extraction
1. ...
2. ...
...
Explanation
1. ...
2. ...

Editor. You are a senior editor. Here are a scientific paper summary, and a short popular science article. A general reader has read the science article and takes some notes.
Please take the following steps:
1. First, evaluate the **reader's notes** based on these factors: content accuracy, lexical and technical complexity, and information conveyance (from the original content).
2. Then, based on the above evaluation, list some brief yet informative writing advice that may benefit the popular science article, to make the article easier for general readers without specialized knowledge to read and understand. Specifically, the advice should benefit these factors of the article:
a) Content Accuracy: The factual correctness, scientific validity, and absence of errors in the general popular science article.
b) Accessibility: Higher accessibility means less technical, more readable and interesting, etc.
c) Information Conveyance: How effectively key information from the original paper is transferred to the popular science article.
Follow the rules strictly:
- Evaluation and advice sections should be in points, like "1...2...3...".
- Each advice should not go beyond the fact of original paper, but can be some common-sense or specialized knowledge.
- Each advice should be targeted at one specific aspect of the article.
- Don't suggest visualization, references or links.
- Suggest explanations rather than content additions.
- The output format:
Evaluation for reader's notes
- Content accuracy of reader's notes: ...
- Lexical and technical complexity of reader's notes: ...
- Information conveyance of reader's notes: ...
Advice
1. ...
2. ...

Revision You are a science journalist for general audiences. Given the paper summary and a short summary of the popular science article, you are assigned to rewrite the popular science summary for general audiences, who have no specialized knowledge. You are given some writing advice.
Please take the following steps:
1. Choose and refine the most relevant and suitable advice for writing improvement.
2. Then, based on the refined advice and the paper summary, rewrite the popular science article.
Follow the rules strictly:
- Keep your article short yet informative.
- Don't include visualization, references or links.
- Revision must not go beyond the original paper, but can be with some additional common-sense or professional knowledge for explanation.
- The output format:
Improvement
...
Revised Article
...

Table 10: Prompts for each LLM agent.