FRAME: Feedback-Refined Agent Methodology for Enhancing Medical Research Insights

Chengzhang Yu [1*], **Yiming Zhang**[2,3*], **Zhixin Liu**[4],
Zenghui Ding[2†], **Yining Sun**[2,3], **Zhanpeng Jin**[1†],

[1]South China University of Technology,
[2]HFIPS, Chinese Academy of Sciences,
[3]University of Science and Technology of China,

Abstract

The automation of scientific research through large language models (LLMs) presents significant opportunities but faces critical challenges in knowledge synthesis and quality assurance. We introduce Feedback-Refined Agent Methodology (**FRAME**), a novel framework that enhances medical paper generation through iterative refinement and structured feedback. Our approach comprises three key innovations: (1) A structured dataset construction method that decomposes 4,287 medical papers into essential research components through iterative refinement; (2) A tripartite architecture integrating Generator, Evaluator, and Reflector agents that progressively improve content quality through metric-driven feedback; and (3) A comprehensive evaluation framework that combines statistical metrics with human-grounded benchmarks. Experimental results demonstrate **FRAME**'s effectiveness, achieving significant improvements over conventional approaches across multiple models (9.91% average gain with DeepSeek V3, comparable improvements with GPT-4o Mini) and evaluation dimensions. Human evaluation confirms that **FRAME**-generated papers achieve quality comparable to human-authored works, with particular strength in synthesizing future research directions. The results demonstrated our work could efficiently assist medical research by building a robust foundation for automated medical research paper generation while maintaining rigorous academic standards.

1 Introduction

The traditional academic research paradigm relies on human researchers to gather knowledge, formulate hypotheses, and evaluate findings through peer review. While this process has driven significant technological advances, it is inherently limited by human cognitive constraints and time-intensive workflows, with studies showing an average publication cycle of 21.9 months (Smart et al., 2013). The emergence of large language models (LLMs), particularly since GPT-3.5 (Wu et al., 2023), has introduced unprecedented capabilities in natural language processing, from text generation to complex reasoning tasks (Zhao et al., 2023; Gómez et al., 2024; Wang et al., 2024a; Li et al., 2024). Advanced techniques like Chain-of-Thought (CoT) (Wei et al., 2022) and frameworks such as LangChain (LangChain, 2023) and MetaGPT (Hong et al., 2024) have further enhanced LLMs' ability to handle sophisticated multi-agent collaboration.

Current applications of LLMs in academic research fall into two categories. The first focuses on specific subtasks like code generation (OpenAI, 2021; cursor, 2023), idea formulation (Hu et al., 2024), and paper review (Jin et al., 2024; Sun et al., 2024). While effective within their domains, these applications cannot encompass the entire research process. The second category employs collaborative multi-agent systems for comprehensive research tasks, exemplified by AutoSurvey's automated literature review pipeline (Wang et al., 2024b). However, these applications have primarily focused on computational domains where experiments can be simulated.

The application of LLMs to support end-to-end academic research, particularly in medical research, faces two significant challenges. First, current large language models primarily rely on factual knowledge (e.g., understanding that colds can be caused by either bacteria or viruses) rather than learning from previous generation failures (e.g., recognizing that a previously unsuccessful attempt at generating a cold-related paper failed to adequately consider bacterial influences). This limitation hinders the models' ability to iteratively improve their output quality through experience-based learning (Wang

*Equal contribution
†Correspondence to Zhanpeng Jin, Zenghui Ding: zjin@scut.edu.cn, dingzenghui@iim.ac.cn

Findings of the Association for Computational Linguistics: ACL 2025, pages 7690–7704

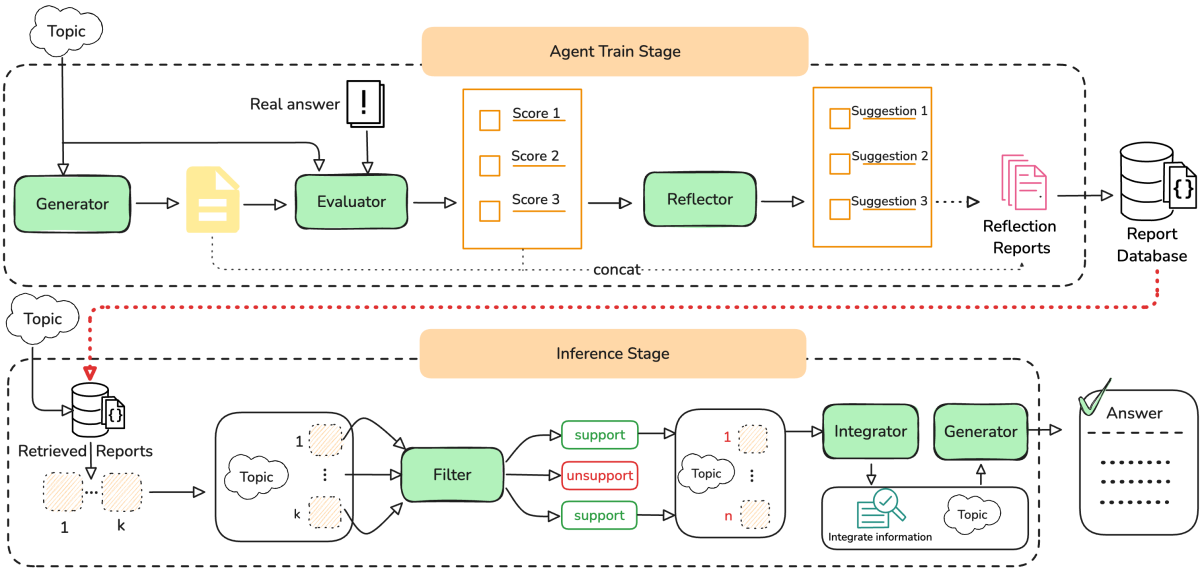

Figure 1: Architecture of the Feedback-Refined Agent Methodology (**FRAME**). During the training phase, the system generates and accumulates Reflection Reports in a dedicated database, which subsequently guides the formal paper generation process. This iterative training paradigm enables continuous refinement of the generation capabilities through structured feedback mechanisms.

et al., 2024b; Lu et al., 2024). Second, current paper generation models almost lack robust evaluation mechanisms, depending primarily on subjective agent assessments without rigorous benchmarks against human-authored papers. These limitations underscore the need for methods that ensure scientific validity while maintaining the academic standards characteristic of peer-reviewed publications.

Drawing inspiration from adversarial learning principles, we propose a novel approach to enhance the quality of LLM-generated academic papers, as shown in Figure 1. Our Feedback-Refined Agent Methodology (**FRAME**) reimagines the dynamic interplay between content generation and quality assessment through a feedback-driven iterative process. Unlike traditional neural network approaches that rely on gradient-based parameter updates, **FRAME** implements a structured refinement cycle where specialized agents work in concert to progressively improve content quality. This methodology addresses the fundamental challenge of continuous self-improvement in LLM-based research systems through three key mechanisms: (1) Specialized agents assume distinct roles in content generation and quality evaluation; (2) The feedback process targets logical coherence and academic rigor through structured assessment; and (3) Knowledge accumulation is achieved through organized reflection reports that guide subsequent reasoning iterations.

Our contributions are summarized as follows:

1. We construct a comprehensive dataset of 4,287 medical research papers, covering diverse topics and methodologies, providing a robust knowledge base for training and evaluation.
2. We propose a Feedback-Refined Agent Methodology (**FRAME**) that effectively leverages prior research, demonstrating superior performance in generating high-quality medical research papers.
3. We introduce a novel evaluation method using human-authored papers as the gold standard, enabling objective assessment of generated content quality in medical research.

2 Related Works

2.1 LLMs for applications in the medical field

The integration of LLMs into medical practice has demonstrated transformative potential across documentation, education, and diagnostics while facing persistent challenges in reliability and ethical governance. In medical writing and research management, models like ChatGPT streamline manuscript drafting, literature synthesis, and multidisciplinary data integration—exemplified by applications in ophthalmology for surgical summaries and cross-disciplinary research coordination (Peng et al., 2023; Bernstein et al., 2023; Gu et al., 2023). However, limitations, including occasional reference fabrication and superficial contextual understanding, necessitate robust fact-checking mechanisms and integration with validated medical databases

(Nakaura and Naganawa, 2023; Thapa and Adhikari, 2023). Educational applications leverage LLMs for simulating patient interactions, generating practice questions, and automating assessments (Dave et al., 2023; Cascella et al., 2023), while diagnostic implementations assist in parsing unstructured clinical data and improving doctor-patient communication (Benary et al., 2023; Dias and Torkamani, 2019). Despite these advancements, critical gaps persist in causal reasoning and contextual interpretation, particularly for ambiguous clinical scenarios (Harris, 2023). The collective experience underscores the necessity for rigorous validation protocols, cultural contextualization, and ethical frameworks to ensure LLMs augment rather than compromise medical standards, serving as decision-support tools rather than autonomous agents (Shah et al., 2023; Peng et al., 2023).

2.2 AI-Powered Academic Research Methods

The evolution of AI in scientific paper writing has progressed from rudimentary language correction to sophisticated end-to-end manuscript generation. Early applications focused on foundational tasks such as grammar and domain-specific spelling checks, with specialized models trained to identify errors in technical fields like medical academic texts (Lai et al., 2015). As AI advanced, its role expanded to address broader linguistic challenges, particularly for non-native English speakers in the English-as-a-Foreign-Language (EFL) community. Tools like Wordtune emerged, enabling writers to overcome language barriers by dynamically rephrasing content across tones (e.g., formal vs. casual) and lengths (e.g., concise summaries or expanded explanations), thereby enhancing both clarity and stylistic adaptability (Zhao, 2023). The advent of large language models (LLMs) marked a paradigm shift, empowering AI systems to automate complex scholarly workflows. For instance, AutoSurvey systematizes literature review creation in fast-evolving domains (e.g., AI research) through a structured pipeline: preliminary data retrieval and outline generation, subsection drafting via specialized LLMs, content integration, and iterative refinement (Wang et al., 2024b). Further pushing boundaries, AI-Scientist demonstrates end-to-end research automation—generating hypotheses, designing experiments, analyzing results, and drafting full manuscripts—while even simulating peer-review processes to evaluate scientific rigor (Lu et al., 2024). These advancements underscore AI's growing capacity to augment—though not yet fully replace—human ingenuity in academic writing, balancing efficiency gains with persistent demands for domain expertise and critical oversight.

3 Dataset Construction

3.1 Existing Dataset Challenges

The construction of datasets for automated paper generation has long followed a paradigm focused on content aggregation, where researchers primarily harvest publicly available academic papers to build monolithic document repositories. For instance, AutoSurvey (Wang et al., 2024b) simply vectorized and stored $530,000$ arXiv computer science papers as its retrieval database without sophisticated preprocessing. Similarly, AI-Scientist(Lu et al., 2024) utilized 500 papers from ICLR 2022 to evaluate the reliability of its proposed paper assessment model. Another example is Nova (Hu et al., 2024), which leveraged a total of 170 papers from CVPR 2024, ACL 2024, and ICLR 2024 to generate initial idea seeds.

However, these existing approaches exhibit two significant limitations in their utilization of academic papers. First, they primarily treat the papers as a static knowledge repository, failing to deeply extract and analyze the structural and logical frameworks inherent in the papers. Second, the overall processing pipeline is often overly simplistic, lacking tailored extraction methods for different sections of the papers. These limitations hinder the ability to fully leverage the rich information embedded in academic papers, thereby constraining the potential of automated paper generation systems.

To address these shortcomings, we propose a more refined approach to dataset construction. Specifically, we decompose each paper into six distinct sections: Topic, Background, Related Work, Method, Result, and Conclusion. The Topic section captures the specific research question or problem the paper aims to address. The Background section provides the contextual foundation and motivation for the study. The Related Work section identifies and discusses the connections between the paper and prior research. The Method section details the research methodology employed, while the Result section presents experimental data, including specific numerical findings. Finally, the Conclusion section summarizes the key takeaways and implications of the study. By extracting and ana-

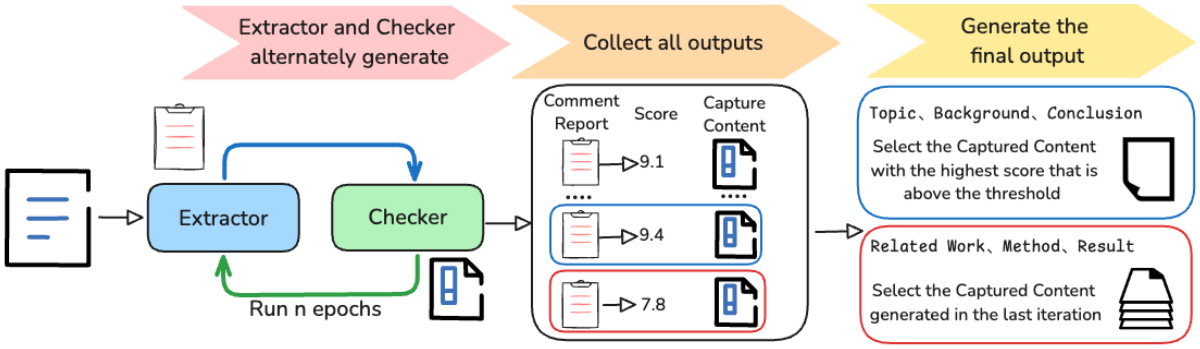

Figure 2: Overview of the dataset construction process. Core information from academic papers is iteratively extracted and refined through N rounds of *Extractor-Checker* cycles ($N = 3$ in our implementation), resulting in a more structured and concise representation of the content.

lyzing these sections individually, we aim to create a more robust and granular dataset that can serve as a stronger foundation for subsequent paper generation tasks. This approach not only enhances the depth of information extraction but also enables more targeted and context-aware generation processes.

3.2 Dataset Construction Process

Our dataset construction begins with an initial corpus of 10,000 research articles harvested from the medRxiv platform, covering 51 medical disciplines (e.g., Allergy and Immunology, HIV/AIDS, Emergency Medicine) and published between 2023 and 2024. To ensure the quality and relevance of the data, we employed a two-tier selection protocol: (1) preference was given to articles subsequently accepted by peer-reviewed journals or cited in other scholarly works; (2) for the remaining articles, we utilized LLM to automatically screen for methodological rigor and completeness, discarding submissions that failed to meet these criteria.

We then implemented a systematic *chapter-to-section mapping* protocol to align the structural components of the research articles with our dataset schema. This mapping ensures that each article adheres to conventional academic organization, thereby validating structural integrity while standardizing content extraction. Specifically, key components are mapped as follows:

- **Topic (Top.)**:Focused on the research question or problem statement.
- **Background (Bkgd.)**: Derived from introductory material establishing the study context and motivation.
- **Related Work (RelW.)**: Sourced directly from dedicated literature review sections.
- **Method (Meth.)**: Extracted from the methodology sections detailing research designs and procedures.
- **Result (Res.)**: Captured from empirical data and analyses presented in the results section.
- **Conclusion (Conc.)**: Synthesized from discussions of findings and implications in the conclusion.

To further ensure data quality, we employed a three-stage filtration process. First, the initial selection criteria (journal acceptance, citation status, and LLM screening) eliminated fundamentally flawed submissions. Second, we excluded articles with non-standard section titles that did not align with our predefined aliases (see Table 1), effectively removing those lacking essential structural components. Finally, during the agent-based extraction process (detailed in Section 3.3), we discarded samples with failed extraction attempts. This multi-stage filtering progressively enforced content quality, structural completeness, and technical consistency, resulting in a final curated dataset of 4,287 high-quality medical research articles.

3.3 Data Extraction Agent

To ensure the accuracy and reliability of information extraction from academic papers, we design a dual-agent framework consisting of an *Extractor* and a *Checker* for each section of the paper. As shown in Figure 2, this iterative refinement process enables progressive quality improvement through structured feedback loops.

The workflow operates through three key phases:

1. The *Extractor* first retrieves key information from designated chapters using the mapping defined in Table 1, yielding the Captured Content that will serve as the foundation for all subsequent processing steps.
2. The *Checker* then evaluates the Captured Con-

tent using the evaluation metrics specified in Table 1, assigning quantitative scores (1-5 scale) for each metric

3. Based on the evaluation results, the *Checker* generates targeted improvement suggestions that are fed back to the *Extractor* for the next iteration

Section	Capture	Candidate Alias
Top.	Introduction	introduction
Bkgd.	Introduction	introduction
RelW.	Related Work	related work, literature review, previous work
Meth.	Methodology	methods, methodology
Res.	Experiment	results, findings, outcomes
Conc.	Conclusion or Discussion	conclusions, summary, discussion

Table 1: Mapping Between Sections and Captured Chapters with Candidate Aliases, indicating acceptable chapter titles for content extraction

This cyclic process continues for n rounds, with each iteration producing progressively refined outputs. We implement different strategies for different sections based on their typical complexity and length characteristics.

For the Related Work, Method, and Result sections, which are typically longer and more structured, we adopt a progressive refinement approach. The output from the final iteration of the Extractor is selected by default.

For the typically shorter but often conceptually dense Topic, Background, and Conclusion sections, we implement a quality-gated selection mechanism. Due to their higher complexity-to-length ratio, the refinement process may not always be monotonic. Therefore, only iterations exceeding predefined quality thresholds across all relevant metrics are considered, and the highest-scoring iteration among those that meet the threshold requirements is selected. If none of the iterations surpass the threshold, the extraction is deemed unsuccessful.

To further enhance the extraction accuracy for the Result section, we employ a preprocessing step using regular expressions to identify and extract all numerical values from the input text. Irrelevant numbers (e.g., page numbers, section numbers) are filtered out through pattern matching and contextual analysis, with the remaining numerical data provided as additional context to the *Extractor*. This focused approach enables the system to prioritize experimentally significant results while maintaining methodological rigor.

Section	Evaluation Metric	Recommendation Dimensions
Bkgd.	Completeness, Relevance, Organization	Content, Relevance, Structure, Style
RelW.	Diversity, Criticality, Connection	Coverage, Analysis, Connection, Synthesis
Meth.	Coherence, Necessity, Completeness	Completeness, Flow, Precision, Justification
Conc.	Comprehensiveness, Impact, Future Direction	Summary, Impact, Future, Synthesis

Table 2: Section-Specific Evaluation Metrics and Corresponding Recommendation Dimensions

4 Feedback-Refined Agent Methodology (FRAME)

To enable our Agent to craft high-quality medical papers on par with human standards, it is insufficient to merely furnish the Agent with a wealth of background knowledge. Instead, the Agent must be capable of explicitly learning from each deficiency and updating its generation strategy in subsequent iterations. Drawing inspiration from adversarial learning principles, we have developed a feedback-driven iterative methodology known as the Feedback-Refined Agent Methodology (**FRAME**), whose architecture is illustrated in Figure 1. This approach emphasizes continuous improvement through structured feedback loops, where specialized agents work in concert to progressively enhance content quality while maintaining rigorous academic standards.

4.1 Agent Training Stage

Our training framework employs a tripartite architecture that synergistically combines generation, evaluation, and reflection mechanisms. As depicted

in Figure 1, the system operates through three core components:

1. The *Generator* synthesizes manuscript sections (e.g., Background, Related Work, Method) conditioned on the research topic
2. The *Evaluator* conducts multi-dimensional quality assessments using the evaluation metrics defined in Table 2 (Column 2), producing quantitative scores (1-5 scale) for each metric
3. The *Reflector* translates *Evaluator* feedback into structured reflection reports by mapping criticism to specific suggestion dimensions from Table 2 (Column 3)

This triadic interaction creates a closed-loop learning system where each component informs the others' improvements. The Evaluator's metric-driven scoring (e.g., assessing Method section coherence through the "Coherence, Necessity, Completeness" metrics) provides objective performance measures, while the *Reflector*'s dimension-specific recommendations (e.g., "Improve Flow and Justification" for Methods) guide targeted revisions.

4.2 Inference Stage

In the phase of generating new papers, our objective is to ensure that the Agent effectively leverages the valuable insights gained during the training phase while disregarding irrelevant experiences. Moreover, it is crucial to manage the context length to optimize the model's generation performance.

We employ a *Retrieval-Augmented Generation (RAG)* approach to retrieve N Reflection Reports from the database, which were formulated during the training phase. This method focuses on extracting substantive insights rather than merely related information, ensuring a stronger foundation for the generation process.

Subsequently, we introduce a model known as the *Filter*, which acts as a gatekeeper by eliminating reports that appear proximal in the vector space but are not truly relevant. This filtering step effectively reduces the interference of unrelated experiences, allowing the Agent to focus on generating content that is more aligned with the intended objectives.

Directly inputting multiple reports into the *Generator* may result in excessive context length, which could impair the Agent's ability to focus on crucial information. Conversely, relying solely on a single report might lead to incomplete understanding, as individual reports typically contain only partial experiential insights. To address this challenge, we utilize an *Integrator* to consolidate and merge multiple pertinent reports. This process is akin to using a larger batch size in Neural Network rather than a batch size of one, facilitating the acquisition of more balanced and comprehensive information. This approach helps the Agent avoid misleading influence from extreme data points.

5 Experiment

Our experimental evaluation comprises four key components. First, we systematically evaluate the impact of FRAME on text generation quality by comparing outputs from two state-of-the-art language models (DeepSeek V3 and GPT-4o Mini) with and without our method. Second, we conduct a rigorous human evaluation where 20 randomly selected FRAME-enhanced papers generated by DeepSeek V3 are compared against human-authored counterparts through expert assessments by medical professionals. Third, we investigate the generalizability of FRAME by evaluating its performance on models released prior to our test set creation, ensuring no potential data contamination. Finally, we examine the influence of training dataset scale on model performance by conducting systematic experiments with varying dataset sizes, providing insights into the data efficiency of our approach.

5.1 Experimental Setup

All experiments were conducted on an A800 computer to ensure a consistent and controlled environment for evaluating model performance across various tasks and metrics. The models DeepSeek v3 and GPT-4o Mini were accessed via their official APIs, while other models were deployed using PyTorch and vLLM (Kwon et al., 2023). The dataset, comprising a total of $4,287$ samples, was carefully partitioned to avoid temporal data leakage. Specifically, we used September 1, 2024, as the cutoff point to divide the dataset into training and testing sets, ensuring no temporal overlap between the two. The training set consists of $4,119$ samples, while the testing set contains 168 samples. We employed FAISS as our database software, which allows for flexible data management, including the ability to freely add or remove entries, facilitating incremental training in future updates.

5.2 Evaluation Metrics

In this study, we employ two distinct sets of evaluation metrics to assess the performance of our

models. The first set comprises statistical metrics, specifically Soft Precision and Soft Recall, which measure the alignment between the model's output and the ground truth. Soft Precision is defined as the ratio of correctly predicted relevant elements to the total number of predicted relevant elements, while Soft Recall is the ratio of correctly predicted relevant elements to the total number of actual relevant elements. These metrics provide a nuanced understanding of the model's accuracy and coverage, particularly in scenarios where binary classification is insufficient.

$$\text{Prec.} = \frac{\sum_{i=1}^{n} \text{Sim}(P_i, G_i) \cdot \mathbb{I}(P_i \in \text{Relevant})}{\sum_{i=1}^{n} \mathbb{I}(G_i \in \text{Relevant})},$$

$$\text{Rec.} = \frac{\sum_{i=1}^{n} \text{Sim}(P_i, G_i) \cdot \mathbb{I}(G_i \in \text{Relevant})}{\sum_{i=1}^{n} \mathbb{I}(G_i \in \text{Relevant})},$$

where P_i represents the i-th predicted element, G_i denotes the corresponding ground truth element, and $\text{Sim}(P_i, G_i)$ quantifies the similarity between P_i and G_i. The function $\mathbb{I}(P_i \in \text{Relevant})$ is an indicator that returns 1 if P_i is deemed relevant and 0 otherwise, while $\mathbb{I}(G_i \in \text{Relevant})$ indicates whether the i-th ground truth element is relevant.

The second set of metrics involves LLM-based scoring, where the output is evaluated across multiple dimensions such as Background, Related Work, Method, and Conclusion. For each dimension, the model's output is compared against the topic and the ground truth, and a score ranging from 1 to 5, in increments of 0.1, is assigned. To mitigate the randomness in LLM-based evaluations, we conduct three independent assessments for the same content and dimension, and the final score is computed as the average of the three evaluations. This scoring system allows for a detailed and robust assessment of the model's performance in generating coherent and contextually relevant content. The specific evaluation dimensions are detailed in Table 2.

5.3 Model Comparisons

As demonstrated in Tables 3 and 4, our method exhibits consistent superiority across both mainstream models. The comprehensive evaluation encompassing 40 comparative dimensions (4 sections × [5 metrics (soft precision, soft recall, S1, S2, S3)] × 2 models) reveals two key findings when compared against three clearly defined baselines: (1) No-RAG, which uses direct model inference without retrieval augmentation; (2) standard RAG, which retrieves relevant content from a FAISS vector database and concatenates it with model input; and (3) Filter, an ablation of our GIA framework that evaluates retrieved content using an LLM. Our approach outperforms all these baselines in every experimental configuration, with an average performance improvement of 9.91% across all sections compared to the strongest baseline (calculated as the mean of five evaluation metrics per section). These results substantiate the effectiveness of our method in enhancing scientific paper generation across different model architectures and document sections.

Sect.	Method	Metrics (%)		LLM Scores			
		Prec.	Rec.	S1	S2	S3	Total
Bkgd.	**Ours**	**90.98**	**89.50**	**74.67**	**83.67**	**82.82**	**84.33**
	Filter	87.33	86.28	59.01	65.30	64.78	72.54
	RAG	87.12	85.85	57.42	63.39	62.70	71.30
	No-RAG	87.64	86.15	55.94	61.91	63.24	70.98
RelW.	**Ours**	**100.0**	**98.46**	**86.98**	**90.90**	**92.97**	**93.86**
	Filter	99.10	97.36	83.04	86.39	88.93	90.96
	RAG	95.95	94.49	73.94	77.26	78.77	84.08
	No-RAG	95.32	94.22	74.15	77.74	79.38	84.16
Meth.	**Ours**	**98.87**	**94.08**	**83.29**	**85.96**	**78.96**	**88.23**
	Filter	98.68	92.41	79.54	83.12	73.82	85.52
	RAG	94.58	89.96	75.89	80.11	70.98	82.31
	No-RAG	95.50	91.52	79.05	82.59	74.62	84.66
Conc.	**Ours**	**93.74**	**93.74**	**73.88**	**76.59**	**77.03**	**83.00**
	Filter	92.98	92.98	69.74	71.44	72.73	79.97
	RAG	89.79	89.79	55.94	57.30	58.30	70.22
	No-RAG	90.07	90.07	58.54	60.14	60.29	71.82

Table 3: Performance comparison of DeepSeek V3 with different methods. S1-S3 represent distinct evaluation dimensions across different sections (e.g., background, methodology), with detailed descriptions provided in Table 2.

Sect.	Method	Metrics (%)		LLM Scores			
		Prec.	Rec.	S1	S2	S3	Total
Bkgd.	**Ours**	**90.48**	**90.07**	**73.44**	**82.08**	**80.07**	**83.23**
	Filter	89.24	88.34	65.35	72.86	72.26	77.61
	RAG	89.21	88.33	66.03	73.37	72.65	77.92
	No-RAG	89.37	88.52	65.66	72.92	72.83	77.86
RelW.	**Ours**	**99.77**	**97.44**	**80.79**	**84.97**	**87.96**	**90.19**
	Filter	97.17	95.08	74.36	78.86	81.29	85.35
	RAG	97.20	95.07	74.06	78.43	81.15	85.18
	No-RAG	97.11	95.12	74.05	78.59	81.17	85.21
Meth.	**Ours**	**98.78**	**91.50**	**77.43**	**81.37**	**71.71**	**84.16**
	Filter	97.05	89.86	74.90	79.38	67.86	81.81
	RAG	96.51	89.57	74.73	79.04	68.33	81.64
	No-RAG	96.52	89.64	74.86	79.19	68.04	81.65
Conc.	**Ours**	**93.27**	**93.27**	**74.23**	**77.27**	**78.25**	**83.26**
	Filter	91.77	91.77	68.37	71.40	72.38	79.14
	RAG	91.69	91.69	67.64	70.36	71.47	78.57
	No-RAG	92.24	92.24	70.90	73.81	74.89	80.81

Table 4: Performance comparison of GPT-4o Mini with different methods. S1-S3 represent distinct evaluation dimensions across different sections (e.g., background, methodology), with detailed descriptions provided in Table 2.

Sect.	Method	Metrics (%)		LLM Scores			
		Prec.	Rec.	S1	S2	S3	Total
Bkgd.	**Ours**	**90.48**	**90.07**	**73.44**	**82.08**	**80.07**	**83.23**
	filter	89.24	88.34	65.35	72.86	72.26	77.61
	rag	89.21	88.33	66.03	73.37	72.65	77.92
	No-RAG	89.37	88.52	65.66	72.92	72.83	77.86
RelW.	**Ours**	**99.77**	**97.44**	**80.79**	**84.97**	**87.96**	**90.19**
	filter	97.17	95.08	74.36	78.86	81.29	85.35
	rag	97.20	95.07	74.06	78.43	81.15	85.18
	No-RAG	97.11	95.12	74.05	78.59	81.17	85.21
Meth.	**Ours**	**98.78**	**91.50**	**77.43**	**81.37**	**71.71**	**84.16**
	filter	97.05	89.86	74.90	79.38	67.86	81.81
	rag	96.51	89.57	74.73	79.04	68.33	81.64
	No-RAG	96.52	89.64	74.86	79.19	68.04	81.65
Conc.	**Ours**	**93.27**	**93.27**	**74.23**	**77.27**	**78.25**	**83.26**
	filter	91.77	91.77	68.37	71.40	72.38	79.14
	rag	91.69	91.69	67.64	70.36	71.47	78.57
	No-RAG	92.24	92.24	70.90	73.81	74.89	80.81

Table 5: Performance comparison of Qwen 1.5 32B with different methods. S1-S3 represent distinct evaluation dimensions across different sections (e.g., background, methodology), with detailed descriptions provided in Table 2.

5.4 LLM Knowledge Cutoff and Paper Generation

To validate the effectiveness of our proposed FRAME method across smaller-scale language models and to mitigate potential biases arising from the inclusion of this paper in LLM pretraining datasets, we conducted experiments using Qwen 1.5 32B (released on February 6, 2024) on a test set comprising papers published on or after September 1, 2024. The experimental results demonstrate that Qwen 1.5 32B achieves comparable performance to DeepSeek V3 without our method, indicating that long-chain paper reasoning does not significantly differ between non-reasoning-focused models. However, as demonstrated in Table 5, the implementation of FRAME yields a substantial performance improvement of 3.8% (81.38% vs 85.21%), thereby substantiating the efficacy of our approach in enhancing the quality of generated scientific papers across different model architectures and scales.

5.5 Human Assessment

To objectively evaluate the effectiveness of FRAME, we conducted a human assessment study comparing 20 papers generated by FRAME-enhanced DeepSeek V3 with 20 human-authored papers. A panel of medical professionals evaluated both sets of papers across key sections including Background, Methodology, Results, and Conclusion.

As shown in Figure 4, the evaluation results demonstrate that FRAME-generated papers achieve statistically equivalent quality to human-authored papers across most sections, with no significant difference in composite scores ($M_{\text{model}} = 92.80\%$ vs $M_{\text{human}} = 92.88\%$, $p = 0.746$, Cohen's $d = -0.098$). However, in synthesizing future research directions—a critical aspect of scientific writing—FRAME exhibits superior performance compared to human authors ($p < 0.001$, $d = 2.27$), suggesting particular effectiveness in forward-looking synthesis.

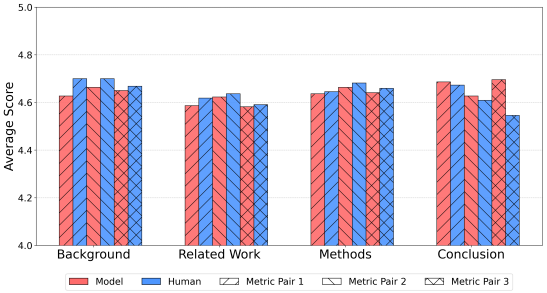

Figure 3: Human vs Model Writing Quality Comparison

5.6 Impact of Training Dataset Scale on Model Performance

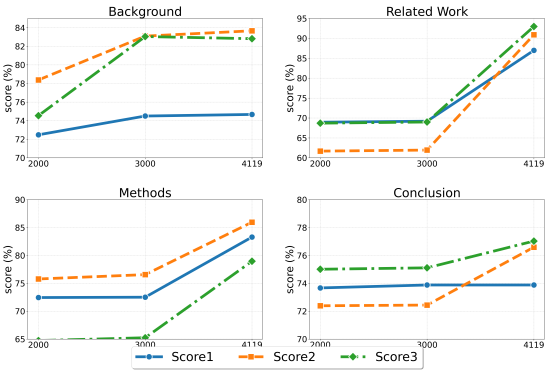

Figure 4: Impact of Training Sample Size on Multi-Dimensional Scoring Metrics

This experiment examines the relationship between training dataset size and model performance using three dataset scales: 2,000, 3,000, and 4,119 samples. As shown in Figure 4, model scores exhibit a non-negative correlation with dataset size, increasing or remaining stable as the dataset expands. This trend is attributed to the model's enhanced ability to generate more accurate reports, as larger datasets provide a broader range of patterns and linguistic structures. During inference, the model is more likely to identify highly similar reference reports, leading to improved performance with larger training datasets.

6 Conclusion

Our Feedback-Refined Agent Methodology (**FRAME**) demonstrates significant improvements in medical paper generation, with DeepSeek V3 and GPT-4o Mini showing average performance gains of 9.91% across 40 evaluation dimensions, while Qwen 1.5 32B achieves a 3.8% improvement. Human evaluations reveal that **FRAME**-generated papers achieve comparable quality to human-authored works (92.80% vs 92.88%), particularly excelling in conclusion synthesis. The proposed tripartite training architecture and structured dataset construction method effectively address key challenges in medical research automation, though limitations in retrieval dependency and offline dataset usage suggest potential for future enhancements through adaptive Generator-Evaluator-Reflector retrieval strategies and dynamic learning from external sources.

Limitations

While the method paradigm brings significant advancements and proves effective, it is also subject to certain limitations that merit discussion.

Retrieval Dependency Our study focused exclusively on a single retrieval step conducted prior to each section's generation. However, suboptimal retrieval quality may indirectly compromise the performance of our core modules. To address this, we propose that future work explore an adaptive, multi-round retrieval strategy that dynamically interacts with these core components, offering a promising direction for enhancing overall effectiveness.

Offline Dataset We build our training dataset by downloading the paper from the website. This offline method could limit our method engage with the latest paper. A promising avenue for further research lies in developing models that actively learn to search the information from the external search engine to improve information retrieval.

Experimental Constraints As an auxiliary tool for scientific writing, our method is designed to assist researchers in rapidly drafting papers based on existing topics and experimental results. However, due to the inherently specialized nature of medical research, our Agent cannot directly assist in conducting experiments or verifying the accuracy of experimental outcomes. Consequently, all experiments and analyses in this study are predicated on the assumption that the underlying papers do not contain fabricated or flawed data. The use of incorrect or incomplete data could severely mislead the paper generation process, potentially compromising the reliability of the output. Future improvements should include mechanisms to validate experimental data integrity or incorporate human-in-the-loop verification for critical research components.

7 Acknowledge

This work was supported in part by the Guangdong Provincial Key Laboratory of Human Digital Twin (Grant No. 2022B1212010004), the Guangzhou Basic Research Program (Grant No. SL2023A04J00930), and the Shenzhen Holdfound Foundation Endowed Professorship. Additional support was provided by the Anhui Provincial Major Science and Technology Project (Grant Nos. 202303a07020006 and 202304a05020071), the Anhui Provincial Clinical Medical Research Transformation Project (Grant No. 202204295107020004), and the National Key Research and Development Program of China Engineering Science and Comprehensive Interdisciplinary Special Project (Grant No. 2024YFF0507603).

References

Manuela Benary, Xing David Wang, Max Schmidt, Dominik Soll, Georg Hilfenhaus, Mani Nassir, Christian Sigler, Maren Knödler, Ulrich Keller, Dieter Beule, et al. 2023. Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*, 6(11):e2343689–e2343689.

Isaac A Bernstein, Youchen Victor Zhang, Devendra Govil, Iyad Majid, Robert T Chang, Yang Sun, Ann Shue, Jonathan C Chou, Emily Schehlein, Karen L Christopher, et al. 2023. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA network open*, 6(8):e2330320–e2330320.

Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. 2023. Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios. *Journal of medical systems*, 47(1):33.

cursor. 2023. Cursor: The ai code editor.

Tirth Dave, Sai Anirudh Athaluri, and Satyam Singh. 2023. Chatgpt in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Frontiers in artificial intelligence*, 6:1169595.

Raquel Dias and Ali Torkamani. 2019. Artificial intelligence in clinical and genomic diagnostics. *Genome medicine*, 11(1):70.

Sergio Gómez, Miguel Domingo, and Francisco Casacuberta. 2024. Interactive machine translation with large language models in low resources languages. In *Proc. IberSPEECH 2024*, pages 66–70.

Yang Gu, Jian Cao, Yuan Guo, Shiyou Qian, and Wei Guan. 2023. Plan, generate and match: Scientific workflow recommendation with large language models. In *International Conference on Service-Oriented Computing*, pages 86–102. Springer.

Emily Harris. 2023. Large language models answer medical questions accurately, but can't match clinicians' knowledge. *JAMA*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. 2024. Nova: An iterative planning and search approach to enhance novelty and diversity of llm generated ideas. *arXiv*.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. Agentreview: Exploring peer review dynamics with llm agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1208–1226.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Kenneth H Lai, Maxim Topaz, Foster R Goss, and Li Zhou. 2015. Automated misspelling detection and correction in clinical free-text records. *Journal of biomedical informatics*, 55:188–195.

LangChain. 2023. Langchain documentation. https://docs.langchain.com/.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Takeshi Nakaura and Shinji Naganawa. 2023. Writing medical papers using large-scale language models: a perspective from the japanese journal of radiology. *Japanese Journal of Radiology*, 41(5):457–458.

OpenAI. 2021. Github copilot: The ai editor for everyone.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. 2023. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210.

Nigam H Shah, David Entwistle, and Michael A Pfeffer. 2023. Creation and adoption of large language models in medicine. *Jama*, 330(9):866–869.

Ryan J Smart, Srinivas M Susarla, Leonard B Kaban, and Thomas B Dodson. 2013. Factors associated with converting scientific abstracts to published manuscripts. *Journal of Craniofacial Surgery*, 24(1):66–70.

Lu Sun, Stone Tao, Junjie Hu, and Steven P Dow. 2024. Metawriter: Exploring the potential and perils of ai writing support in scientific peer review. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–32.

Surendrabikram Thapa and Surabhi Adhikari. 2023. Chatgpt, bard, and large language models for biomedical research: opportunities and pitfalls. *Annals of biomedical engineering*, 51(12):2647–2651.

Tian Wang, Junming Fan, and Pai Zheng. 2024a. An llm-based vision and language cobot navigation approach for human-centric smart manufacturing. *Journal of Manufacturing Systems*, 75:299–305.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, et al. 2024b. Autosurvey: Large language models can automatically write surveys. *arXiv preprint arXiv:2406.10252*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

Xin Zhao. 2023. Leveraging artificial intelligence (ai) technology for english writing: Introducing wordtune as a digital writing assistant for efl writers. *RELC Journal*, 54(3):890–894.

Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023. Investigating table-to-text generation capabilities of llms in real-world information seeking scenarios. *arXiv preprint arXiv:2305.14987*.

A Ethical Considerations

All personnel involved in the evaluation process participated voluntarily and received ample compensation. All data used in our experiment is sourced from arXiv and is allowed for non-commercial use. The core sections of the paper and all experiments were completed by humans, with AI only assisting in polishing the language and wording.

B Prompt used in FRAME

Due to the space limit, we released our prompt used during the agent train stage. The remaining details will be available in the code.

Extractor Prompt (Taking Conclusion as an example)

Role: Paper Analyst
Task: Identify and analyze the real-world problems addressed by the provided article paragraph.
Requirements:

- If the input only contains the paragraph, then extract the conclusion based on the paragraph. If the input includes historical records from previous instances, it is necessary to reference both the paragraph and the previous scores and reasons to provide a better conclusion.
- Output format: Describe the analysis results in natural language and output in JSON format, which should only contain a single key-value pair with the key "conclusion" and the value being the conclusion description.

Input Role:

- Current content: The article paragraph.
- Previous evaluations: Historical records or scores from previous evaluations.

Checker Prompt (Taking Conclusion as an example)

Role: Conclusion Evaluator
Task: Evaluate the conclusion extracted from the provided article paragraph.
Requirements:

- Input format: The input should consist of two parts: the article paragraph and the extracted conclusion.
- Output format: Return a JSON object containing two key-value pairs:
  - "score": A numerical score from 0 to 10 (incremented by 0.1) indicating the relevance of the conclusion to the article paragraph.
    - 0-2: Completely irrelevant
    - 2-4: Mostly irrelevant
    - 4-6: Partially relevant
    - 6-8: Mostly relevant
    - 8-10: Highly relevant
  - "reason": A textual explanation for the assigned score.

Generator Prompt (Taking Conclusion as an example)

Task: Based on the provided information from multiple related papers, we kindly request you to synthesize a comprehensive and well-structured conclusion section for a new research paper addressing this specific topic, ensuring that it integrates key findings and implications while maintaining academic rigor.
Input:

- Research question: [question]

- Background: [background]
- Related Work: [related works]
- Methods: [methods]
- Result: [result]

Output Format: Please return your response in the following JSON format.
Requirements:
- Summarize the key findings and their significance
- Connect back to the research question
- Discuss implications and potential impact
- Acknowledge any limitations
- Suggest future research directions

Reflector Prompt (Taking Conclusion as an example)

Please analyze the following conclusion section generation and provide improvement suggestions:
Topic: [topic]
1. Model Prediction (Generated Conclusion): [prediction]
2. True Answer (Reference Conclusion): [reference]
3. Evaluator's Comments: [evaluator comments]
4. Evaluation Scores:
- Comprehensiveness: [comprehensiveness score]/5.0
- Impact: [impact score]/5.0
- Future Direction: [future direction score]/5.0

Please analyze the gaps between the generated content and the reference, considering the evaluator's feedback and scores. Focus on identifying specific areas for improvement and providing actionable suggestions.
THOUGHT: Analyze the following aspects:
- Results Summary:
  - What key findings need better emphasis?
  - Which results need clearer interpretation?
  - How can the significance be better conveyed?
- Research Impact:
  - What theoretical contributions need better articulation?
  - Which practical implications should be highlighted?
  - How can the broader impact be better demonstrated?
- Future Directions:
  - What promising research directions should be added?
  - How can future work suggestions be more specific?
  - Which limitations need better connection to future work?
- Integration and Synthesis:
  - How can findings be better connected to research goals?
  - What overarching themes need better emphasis?
  - How can the conclusion be more compelling?

C Specific examples of comments generated by the reflector

Background Section Generation Report

Evaluation Results:
Scores:
- Completeness: 1.75/5
- Relevance: 2.0/5
- Organization: 2.0/5

Evaluator Comments:
- **Content Coverage Comparison:**
  - The depth and breadth of coverage in the content to evaluate are adequate for its specific topic but lack the detailed methodological and application context provided in the reference.
  - (Some comments have been omitted due to length constraints.)
- **Completeness Score and Rationale:** Based on the above analysis, the content receives a Completeness score of 1.75 because it covers the main concepts related to thymic involution and hyperplasia quantification but lacks detailed methodological discussion and specific application examples as seen in the reference.
- **Relevance Alignment:**
  - The content does not deviate from its research focus but does not expand its relevance to broader societal impacts as effectively as the reference.
  - (Some comments have been omitted due to length constraints.)
- **Relevance Score and Rationale:** The content scores 2.0 in Relevance for its strong alignment with the research question but lacks the broader societal impact demonstrated in the reference.
- **Organizational Comparison:**
  - The organization falls short of the reference's clarity and balance in presenting a detailed methodological approach and its validation.
  - (Some comments have been omitted due to length constraints.)
- **Organization Score and Rationale:** The content scores 2.0 in Organization for its logical flow and structure but falls short in methodological depth and balance compared to the reference.

Overall Gap Analysis:
- The key qualitative differences from the reference include the lack of detailed methodological discussion, specific application examples, and broader societal relevance.
- (Some comments have been omitted due to length constraints.)

Reflection Analysis: The content effectively outlines the thymus's immune function and CT imaging challenges but lacks methodological depth, specific applications, and societal relevance. While well-organized, it would benefit from clearer transitions and a more detailed methodological progression.

Improvement Suggestions:
- **Content Improvements:**
  - Include detailed methodological discussion, such as the specific image processing and machine learning techniques used for thymic quantification.
  - (Some comments have been omitted due to length constraints.)
- **Relevance Improvements:**
  - Ensure all content is directly aligned with the research question by removing or revising any information that does not contribute to the understanding of automatic thymic quantification.
  - (Some comments have been omitted due to length constraints.)
- **Structure Improvements:**
  - Add clearer transitions between topics to guide the reader through the methodological development and its validation.

  - (Some comments have been omitted due to length constraints.)
- **Style Improvements:**
  - Make the explanation clearer by defining technical terms and providing examples where necessary.
  - (Some comments have been omitted due to length constraints.)