Designing a Hybrid System for Automated Scientific Reviewing Combining NLP Models with Human-In-The-Loop Feedback Mechanisms

by

MD. Tarikul Islam
22101401
Sifatullah Fahim
22101402
MR Rafi Ahmed
24341286
Jawad Al Wasi
22101470
Montasir Mogumder Shariar
22301577

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
February 2026

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name and Signature:

MD. Tarikul Islam
22101401

Sifatullah Fahim
22101402

MR Rafi Ahmed
24341286

Jawad Al Wasi
22101470

Montasir Mogumder Shariar
22301577

Approval

The thesis titled "Designing a Hybrid System for Automated Scientific Reviewing Combining NLP Models with Human-In-The-Loop Feedback Mechanisms" submitted by

1. MD. Tarikul Islam (22101401)
2. Sifatullah Fahim (22101402)
3. MR Rafi Ahmed (24341286)
4. Jawad Al Wasi (22101470)
5. Montasir Mogumder Shariar (22301577)

of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering in Fall, 2025.

Examining Committee:

Supervisor:
(Member)

Mr. Moin Mostakim

Senior Lecturer

Department of Computer Science and Engineering

BRAC University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor

Department of Computer Science and Engineering

BRAC University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Associate Professor; Chairperson

Department of Computer Science and Engineering

Brac University

Acknowledgement

It is our pleasure to express our gratitude to our thesis supervisor who contributed to this study with her guidance, expertise, and valuable feedback and pointed the direction in which we needed to go. The expertise they had in machine learning and natural language processing was invaluable in the research process.

This study owes its thanks to the Department of Computer Science and Engineering of BRAC University to provide computational facilities and to create an academic environment that was favorable to the conduct of this research.

We would also like to mention the open-source community, who developed the tools and the frameworks used in this project- PyTorch, Hugging faces transformers, LangGraph, and FAISS, without which the given implementation would have been impossible.

Ethics Statement

In this study, the research has been conducted under the current ethical paradigms about research in the field of artificial intelligence and machine learning, so that all the methodological decisions made are consistent with these professional and institutional standards.

Data Usage: The datasets, which were used in this study, including the PeerRead, S2ORC, and the ACL-OCL Corpus are made publicly available with the aim of inquiry among the academic community in regard to their licenses. Participant privacy was preserved by not necessarily harvesting any personally identifiable information than what is already displayed in these repositories.

Bias and Fairness: We acknowledge the possibility that automated peer-review apparatus can reproduce institutional prestige skew, other geographical bias, and more bias due to author recognition: The training corpora contains biases. In order to overcome these issues, the system uses confidence-weighted routing to the human adjudicators whenever the assessment assessments are below specified levels of uncertainty and thus maintain important human control over consequential decisions.

Transparency: With each automated assessment, there will be evidence-based rationales, allowing reviewers to understand and, when the need arises, question the conclusions of the system. Also, every content produced by artificial intelligence is clearly marked out by other material produced by human writers, which makes attribution easy.

Human Agency: The architecture has been specifically designed to support, instead of overpower human judgment. Finally, the final decisions on the acceptability of the manuscript are still considered the exclusive province of the human reviewers, who, after all, have the right to override or amend any AI-based recommendation.

Automation Bias Reduction: The platform recognizes the tendency to excessive trust in the outputs of algorithms and provides multiple forms of safeguarding: (1) the simultaneous presentation of confidence indices next to the AI scores; (2) visual indicators laying designation on unclear dimensions, which should be discussed; (3) the overruling of scores is obligatory with the associated rationale submitted; (4) a default preference in favor of the human adjudication over the automatic one; and (5) the exhaustive audit trails helping to identify the discrepancies between human and AI scores and use them.

Intended Use: The implement is designed and intended to be a leave, rather than a replacement of human judgment in academic assessment by younger peer reviewers. Further protection and frequency of bias audits, as well as clear communication to authors about the use of AI assistance in the revision process, should be involved in the deployment in operational settings.

Abstract

The rise of academic manuscript submissions poses a significant threat to the traditional paradigm of peer-review which overburdens the reviewers with a large amount of submissions and increases inequities as well as biases, at the same time requiring high-quality timely feedback. In this paper, **SKY**, an advanced human-AI hybrid system will be introduced, which focuses on automation of the scientific review process by implementing the state-of-the-art natural language processing (NLP) algorithms and the human-in-the-loop (HITL) framework that balances the use of machine intelligence and human knowledge. Large language models (LLMs) like *Mistral-7B* and *Qwen2.5-7B*, which are fine-tuned using QLoRA, are used in the architectural design to solve domain-specific assessments, thus improving epistemological accuracy. **SKY** as an orchestrator consists of four functionally independent modules of workers, namely **SKY-ORI** (Originality and Impact), **SKY-MDR** (Methodology and Rigor), **SKY-PRC** (Presentation and Clarity), and **SKY-RQM** (Review Quality and Meta-Review), which provide systematic evaluations on a 0–5 scale, with confidence ratings and rationales. A confidence-based routing process controls routing decisions: outputs with confidence score of ≥ 0.85 get automatically forwarded, the scales between 0.60 and 0.85 get into the active-learning reviewing system, and the scores with value < 0.60 are evaluated by human reviewers. Empirical appraisal of the *PeerRead* and *ACL-OCL* datasets has shown a total accuracy of **82.3%**, a Cohen's κ of **0.56**, which is larger (above the 0.43 agreement level) than that obtained between human reviewers in the literature and is a **42%** reduction of the time required to conduct the review.Lastly, the HITL framework bridges natural weaknesses of LLMs, such as hallucinations, limited visual content processing, and lack of methodological critique, by offering human moderation of AI-generated suggestions coupled with the system refinement through expert critique.

Keywords: Aspect Score Prediction; Automated Essay Scoring (AES); Automated Scientific Reviewing; Bias Mitigation; Conflict of Interest (COI); Deep Learning; Human-in-the-Loop (HITL); Hybrid System; Knowledge Graphs; Large Language Models (LLMs); Machine Learning; Natural Language Processing (NLP); Paper Acceptance Prediction; Peer Review; PeerRead Dataset; Reviewer Assignment; ROUGE (Recall-Oriented Understudy for Gisting Evaluation); Text Summarization

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Description
SKY-ORI	SKY Originality & Impact Evaluator
SKY-MDR	SKY Methodology & Rigor Assessor
SKY-PRC	SKY Presentation & Clarity Analyzer
SKY-RQM	SKY Review Quality & Meta-Review Evaluator
AI	Artificial Intelligence
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
FAISS	Facebook AI Similarity Search
GPT	Generative Pre-trained Transformer
HITL	Human-in-the-Loop
JSON	JavaScript Object Notation
LLM	Large Language Model
NLP	Natural Language Processing
QLoRA	Quantized Low-Rank Adaptation
RAG	Retrieval-Augmented Generation
RLHF	Reinforcement Learning from Human Feedback
RNN	Recurrent Neural Network
YAML	YAML Ain't Markup Language
AUC	Area Under the Curve
ECE	Expected Calibration Error
F1	F1-Score (Harmonic Mean of Precision and Recall)
FNR	False Negative Rate
FPR	False Positive Rate
MCC	Matthews Correlation Coefficient
ROC	Receiver Operating Characteristic
TNR	True Negative Rate
TPR	True Positive Rate
ACL	Association for Computational Linguistics
ICLR	International Conference on Learning Representations
NeurIPS	Neural Information Processing Systems
ORCID	Open Researcher and Contributor ID
PDF	Portable Document Format
S2ORC	Semantic Scholar Open Research Corpus
COI	Conflict of Interest
DCT	Discrete Cosine Transform
IRB	Institutional Review Board
SHAP	SHapley Additive exPlanations
SPECTER	Scientific Paper Embeddings using Citation-informed Transformers
XAI	Explainable Artificial Intelligence

Chapter 1

Introduction

1.1 Background and Motivation

Academic peer review is one of the strongest components of scientific investigation, which ensures the reliability of the research by issuing critical suggestions and evaluating scholarly works. However, the traditional peer-review system faces significant challenges as submissions keep increasing and the operation restrictions exist.

One of the main obstacles is the manifold growth of the submissions of manuscripts, which is the tendency to isotropise the peer-review machinery, extend the review process, introduce discrepancy, and create bias. The deliberative stage is a thinking step where meta-reviewers have a responsibility of synthesizing the information that is heterogeneous and moving towards the harmonization of divergent opinions. In addition, high-quality evaluations are still difficult to acquire, particularly among early-career scholars or investigators who do not enjoy substantial institutional support.

As a result, interest in the exploitation of the techniques of Artificial Intelligence (AI) and Natural Language Processing (NLP) in order to partially automatize the peer-review process has increased. However, existing NLP-based systems, in particular, Large Language Models (LLMs), have significant weaknesses. They often produce shallow commentary, which is not rigorously methodologically or conceptually scrutinized and can result in partisan, inaccurate, and fake pieces. Moreover, such models have a problem with visual and mathematical works, such as figures, tables, and equations.

The situation prompts the need to develop a hybrid human-AI system to evaluate a manuscript which would combine the computational effectiveness of neural natural language processing systems and the critical appraisal and ethical decisions of human entities. This system would be able to improve the sense-making process, optimize the process of review and lessen the problem of trust, over-reliance, and agency erosion as well as decrease the cognitive load of the reviewers.

1.2 Problem Statement

Academic peer review can be accelerated and improved through the use of Artificial intelligence (AI) by providing large-scale benefits and efficacy but there is still a significant disconnect between all-autonomous evaluation systems and those carried out by human professionals. This procedures will be able to service enormous quantities of submissions in a short time, but they will have no contextual awareness, partial reasoning and judgment susceptibility to affect the end result to technique and subjectivity. Conversely, human reviewers are more responsible and nuanced, and in contrast to that, they can hardly cope with the growing workload, which leads to bottlenecks, inconsistences, and even bias. With the ever-growing number of academic writings, the purely anthropcentric methods can no longer be maintained.

This gap shows that there are several significant issues in current review processes that can be significantly tackled with the help of hybrid systems:

- **Cognitive Load and Time Constraint:** Meta-reviewers are sometimes overwhelmed with too much work and complicated decision-making procedures. Hybrid automation can decrease such a load summarizing important points, synthesizing the comments of reviewers, and uniting feedback in a well-structured manner.
- **Specificity and Actionability of Feedback:** The existing AI-based systems are applicable towards giving generic feedback. As a contrast, hybrid systems are aimed at creating more specific recommendations, constructive ones, and ones, which are useful to authors and reviewers directly.
- **Bias Mitigation and Fairness:** Human and AI systems both may be biased. This can also be employed as the opportunity to use a hybrid approach by which the biases are followed, detected, and reduced by means of cross-checking and control, resulting in more balanced and unbiased assessments.
- **Iterative Learning with Expert Feedback:** The HITL frameworks allow NLP models to be continually improved through expert feedback. This process of refinement makes more models more effective and adaptable to domain specific situations and gains more predictive accuracy and usefulness with time.

1.3 Research Objectives

Following the research gap identified and the challenges associated with it, the following objectives are followed in this work:

1.3.1 Primary Objectives

1. **Develop a Hybrid Review System:** Develop a system whereby sophisticated NLP-models do the analysis and content generation, with humans as the final decision-makers and judges. This will place efficiency and will not interfere with integrity and reliability of review process.

2. **Increase Review Quality and Specificity:** Make the system generate meaningful, specific and consistent reviews pointing out the weaknesses and proposing the specific improvements. Combination The system must also favour meta-review synthesis using organised summaries and drafts.

3. **Preserve Human Agency and Mitigate Over-reliance:** Make sure that the interaction between humans and AI does not compromise human agency and does not rely too heavily on AI-generated content such that the end-decision remains in human hands. This solution will promote trust, safeguard authorship, principles of fairness, and better transparency through AI tool usage.

1.3.2 Goals for System Effectiveness and Human Interaction

- **Efficiency:** Save significantly on the time needed to undertake a review synthesis and meta-reviewing procedure as compared to traditional human-only workflows.
- **Quality and Coverage:** Make the particularity and perceived quality of reviews better, as measured by the methods of expert judgment and objective measures including review length, review structure, and completeness.
- **Accuracy:** Match or surpass the accuracy of current automated models in predicting paper acceptance or aspect scores.
- **User Preference and Trust:** Earn high scores on user satisfaction and preference of AI-assisted feedback and interaction with the system, indicating higher levels of trust and usefulness of the system by reviewers.

1.3.3 Success Criteria

- The hybrid system is good at incorporating the NLP elements into a human-in-the-loop mechanism as per the design requirements.
- The quality of reviews delivered by the system is comparable to that of human reviewers without any bias or compromise of ethics.
- Its practical applicability is shown in a real sense, which allows to integrate it in practical scholarly review processes.
- According to user studies, it can be concluded that the reviewers are ready to use the system supported by good performance and ease of use.

1.3.4 Secondary Objectives:

- Measure the hybrid review system based on the quantitative measures and qualitative feedback analysis to understand the effects on meta-reviewing.
- Find out about how the integration of AI can be utilized to facilitate sense-making and make the meta-review process faster.

- Research the ways AI implementation can manifest itself in supporting sense-making and speeding up the meta-review process.
- Research the medium-term effects of AI scaffolds on reviewer development (either enhance or reduce the development of skills in novice reviewers).
- Warn of future improvements in the system by solving difficulties in system complexity with respect to the shortage of data and scarcity of deep reasoning in hybrid systems.
- Establish the mechanisms of explanation and justification, according to which the decisions made by AI will become transparent, comprehensible, and interpretable, which will raise the level of trust in a system as a whole.

1.4 Scopes and Challenges

1.4.1 Scopes of the Research

Our work is in the creation of a system that enhances the science peer review. process. The system develops systematized review drafts and identifies weaknesses and is also the one which makes recommendations and evaluates papers according to originality, clarity,soundness, and impact. It also semantically matches and knowledge graphs in order to appoint reviewers and point out any potential conflict of interest. The system would be fair and quality as big language models would be human controlled. It is also useful to provide novice reviewers with guided and summarized summaries, labeling, as also guidance and professional guidance, and progress with time and continual human feedback.

1.4.2 Challenges in Research

There are significant technical, ethical, and operational problems of developing this hybrid review system. AI models can create biased content. They also fail to work with non textual components like figures or equations. AI models also leak the reasoning required for deep research analysis. The scarcity,bias and the lack of structured training data creates more problems to model development. Thinking through the ethical lens AI can remove transparency and fairness by increasing bias or compromising human critical thinking. Regarding system integration, combining human input and AI is complex and measuring subjective aspects such as novelty remains challenging. Additionally, human quality control is expensive and does not scale.

1.5 Contributions

This work is a **systems engineering contribution**, not an algorithmic one. We combine existing techniques (RAG, QLoRA fine-tuning, deterministic NLP tools, calibrated confidence routing) into a practical, HITL-first pipeline for AI-assisted peer review. The contributions are:

1. **SKY Orchestrator Architecture:** A LangGraph-based orchestrator-worker system that coordinates four specialized NLP modules (SKY-ORI, SKY-MDR, SKY-PRC, SKY-RQM) for manuscript evaluation. The system achieves 82.3% accuracy on PeerRead, competitive with ReviewerToo (81.8%) but below human performance (83.9%).

2. **Confidence-Based Routing:** A three-tier routing system where high-confidence outputs ($\geq$0.85) auto-proceed, medium confidence (0.60–0.85) queues for active learning, and low confidence ($<$0.60) routes to human review. This ensures humans review uncertain cases.

3. **Parameter-Efficient Fine-Tuning:** QLoRA fine-tuning applied to Mistral-7B and Qwen2.5-7B for scientific review tasks. Performance is competitive at reduced computational cost (8 GB vs 40 GB memory). This uses established techniques rather than new methods.

4. **Hybrid Tool Integration:** Pipelines that combine deterministic analysis tools (FAISS similarity, LanguageTool, heuristic analyzers) with LLM-based synthesis. Tool grounding improves consistency and interpretability.

5. **Preliminary Empirical Validation:** Evaluation on PeerRead datasets shows SKY-Human agreement ($\kappa = 0.56$) exceeds Human-Human agreement ($\kappa = 0.43$) on a 200-paper sample. A preliminary user study (N=12) indicates 42% review time reduction, though the small sample size limits generalizability.

6. **HITL Framework:** A human-in-the-loop interface using LangGraph's interrupt mechanism. The interface preserves human agency and collects feedback for potential model improvement.

Scope of Contribution: This work does not introduce new learning algorithms or inference paradigms. The contribution is systems integration: we combine known components into a coherent, deployable pipeline with HITL safeguards for a sensitive application domain.

Chapter 2

Literature Review

2.1 Preliminaries

This section covers the concepts and techniques behind automated scientific peer review:

2.1.1 Core Concepts and Definitions

Peer Review Process: The traditional peer review system guarantees quality of scientific publications by expert assessment. Reviewers evaluate the originality of the manuscripts, their soundness of the methods and their contribution to the field. The process involves Editorial screening, reviews by experts in the domain, meta review synthesis and final editorial decision.

Natural Language Processing (NLP): An artificial intelligence subsystem that deals with the ability of computers to comprehend, generate and analyze human language. NLP can be used in scientific reviewing to perform text analysis, extract important information, preprocess manuscripts and generate feedback based on the data.

Large Language Models (LLMs): Neural networks that are trained on large text datasets and are able to comprehend the context and produce readable text. Examples include GPT-4, BERT, PaLM 2, and Mistral, the state of the art in language realization and interpretation.

Human-in-the-Loop (HITL): A paradigm that involves the use of human expertise into supervision, validation and correction machine learning workflows. This is a method that uses human judgment to optimize automated systems without losing the efficiency that automation brings.

2.1.2 Technical Foundations

Modern NLP systems are built based on the transformer architecture. This processes text efficiently through self attention mechanisms. The key preprocessing steps include tokenization,embedding generation, and structural parsing of scientific documents.

Knowledge graphs are used for relational information. Then vector embeddings are used for semantic similarity measurement and finally structured databases are used for storing review metadata and decisions.

2.2 Review of Existing Research

The study of Yuan et al. (2021) started with a simple question whether we can automate scientific reviewing or not and eventually proposes a new system, ReviewAdvisor, and the ASAP-Review dataset. Such applications are looking to tabulate scientific peer reviews using aspect based summaries[11]. It models on annotated reviews of ML venues and tests them by Aspect Coverage, Informativeness and Summary Accuracy (80%+)[11]. The system surpasses human reviewers in breadth of aspects, is errant in accuracy of facts, demographic slant and generic analysis. Although it has the potential to be used as the aid in review, it is unable to substitute human reviewers and the lack of Human-in-the-loop (HITL) which might lead to non-factual statements and the lack of high-level understanding for many aspects of paper assessment because of the outdated nature of a pre-trained model. The article provides a solid basis, but some biases should be reduced, the experiment should be better explained, and the ethical aspects should be mentioned.

Cohan et al. (2019) make an important contribution to scientific NLP by proposing a new model of multitask learning applied to classification of citations intent [5] . Which introduces the idea of structural scaffolds and additional tasks to resourced-based tasks that lack manual annotation, such as prediction of section titles or citation merit, and makes use of the natural document outline. This new method reaches a high 13.3 dramatic absolute F1 score increase (67.9%) on the dataset ACL-ARC and it also performs well on their new presented Sci-Cite dataset that contains five times the number of the first datasets and has greater diversity, improving both empirical power and generalization [5]. The fact that the model does not rely on any external linguistic entities and was handcrafted, as well as their release of their dataset and code makes the model committed to open science and the possibility of reproducibility. Although the knowledge of the scaffold mechanism and the generalization of the domain would add value to the work, and integration with the large language model is a possible extension, they are future-related suggestions but not essential areas of concern. Apart from that, the paper is reasonably presented, methodologically well-grounded, and practically significant, and needs to be accepted.

Dycke et al.(2022) [13] introduced "YES-YES-YES", a donation-based workflow for the ACL Rolling Review that addresses the lack of ethically sourced peer review data by obtaining explicit consent from authors and reviewers to share their submissions and reviews. Targeting NLP researchers, this module allows authors and reviewers to consent to data donation, ensuring ethical and legal compliance. It provides information about dataset size, donation-driven selection biases, and correlations between review content and manuscript sections. The paper is well structured and grounded in existing peer-reviewed literature, though its bias analysis could be more comprehensive. Some limitations include potential selection bias inherent in voluntary donations and a focus limited to the ACL domain. Reproducibility is high given

that the donated data is available and the work's primary contribution is demonstrating a scalable, ethical method for assembling peer review. This work's impact lies in enabling ethical data collection for review studies, though broader participation strategies are needed, as it supports the development of fair automated review systems.

Ghosal et al.(2019) [6] introduce deepsentipeer, a deep neural network model to predict peer review outcomes by leveraging sentiment in review texts. It also addresses the challenges of automating decision making in scholarly publishing. The model integrates three information channels:paper full-text,review texts and review sentiment polarity,encoded using VADER sentiment analyzer and Universal Sentence Encoder. Using the PeerRead dataset,they preprocess PDFS into JSON using ScienceParse. For the first task DeepSentiPeer reduces RMSE by 29% on ICLR 2017 with cross domain improvements on ACL 2017 (4.8%) and CoNLL 2016 (14.5%). For the second task it achieves 71.05% accuracy on ICLR 2017, surpassing the 65.79% baseline, and 64.76% (ACL 2017) and 67.71% (CoNLL 2016) in cross-domain settings.The paper details CNN based feature extraction,MLP prediction and Hyperameters like Adam Optimizer,Batch size 64 with code publicly available on GitHub. Limitations of this paper includes reliance on sentiment analysis accuracy and focus on NLP conferences which limits generalizability. DeepSentiPeer's contribution lies in improving automated review systems though nuanced sentiment detection needs more exploration.

Kang et al. (2018) [3] present PeerRead, the first large publicly available peer review dataset, aggregating 14,700 submissions and 10,700 reviews from ACL (2017), NIPS (2013–2016), and ICLR (2017), with 1,300 reviews manually annotated for clarity, originality, and impact. They partnered with organizers, used web crawling and PDF parsing where needed, and anonymized all content. Features that have been extracted are word counts, citation counts, reviewer scores, and submitted raw and structured fields. The preprocessing steps (including tokenizing, stopword deletion, and anonymizing) and the rules of annotation, train/dev/test splits (80/10/10), and ablation experiments are described in the paper in detail, and it is stated that using textual and bibliometric features together results in the best performance of acceptance prediction. Limitations, however, are a selection bias caused by the nature of some reviews as opt-in that could bias the corpus to more positive feedback. In general, the main contribution of PeerRead is the establishment of the first massive, freely available, peer review data, so that future studies regarding the stability of peer review through data analysis have a building block to work with.
Jialiang Lin et al. (2023) introduce Automated scholarly paper review. It is an automated system that does not need any human involvement[16]. Their multistage workflow begins with parsing and representation, where submitted articles are converted into structured data. Then the screening stage uses AI techniques to examine formatting, detect plagiarism, identify machine-generated content, classify article types, and evaluate the manuscript. Following screening, the main review stage evaluates over four key aspects: originality, soundness, clarity, and significance[16]. Finally, the system generates a well-structured summary through Transformer-based abstractive summarization together with comments and scores and an acceptance recommendation. The workflow provides a complete solution to traditional peer

review, but it continues to struggle with multimodal data limitations, ethical issues, deep reasoning, and cross-disciplinary challenges, which the authors plan to investigate in their future studies

According to Zhongfen Deng et al. (2020), they propose a deep learning model called HabNet to predict review scores and make acceptance recommendations based solely on the text of peer reviews[7]. HabNet captures the hierarchical structure of review data through a three-level encoding system. First, a sentence encoder uses bi-directional self-attention to understand the relationship between words within a sentence. After that, the encoder models the connection between sentences within each review. Lastly, the inter-review encoder integrates information from multiple reviews of the same paper using a bi-directional GRU followed by attention mechanisms to capture relationships between reviews. HabNet finally generates a compact representation of a paper and each review. The model was trained and evaluated on the OpenReview and an extended PeerRead dataset and showed 5% improvement over existing models. Two new evaluation metrics, distance measure and optimized precision are proposed to address the imbalanced rating distributions that better reflect prediction quality. Despite its promise, it still struggles with reviews with mixed sentiments, leading to incorrect predictions.

According to Liang et al. (2023), they worked on a GPT-4-based pipeline that extracts information from research PDFs and provides structured, journal-style feedback on significance, novelty, and improvement suggestions[15]. The system was evaluated through retrospective analysis comparing GPT-4 feedback with peer reviews from 3,096 Nature and 1,709 ICLR papers, and through a user study with 308 researchers from 110 U.S. institutions, where 57.4% found the feedback helpful [15]. Results show GPT-4's feedback aligns with human reviewers but lacks methodological and domain-specific depth and cannot interpret visuals. The pipeline is not open-source, raising reproducibility and ethical concerns. Overall, GPT-4 shows potential as a supplementary reviewer, though not a replacement for experts.

Baek et al. (2025) introduce ResearchAgent, a system that simulates research problem definition, method design, and experiment planning through iterative LLM feedback from ReviewingAgents [23] . It combines citation graph-based literature exploration, an entity-centric knowledge store, and multi-agent review cycles. It combines citation graph-based literature exploration, an entity-centric knowledge store, and multi-agent review cycles. ResearchAgent outperforms in originality. But it struggles with formal reasoning. Even after all these limitations ResearchAgent is a big step toward AI and cross-domain scientific fields.

2.3 Summary of Key Findings

The literature analysis shows that there are major pieces of information about the present situation and the prospects of automated scientific review system.

2.3.1 Capabilities and Limitations of Current Systems

AI systems are very effective in particular peer review tasks. ReviewRobot has prediction accuracy of 71.43% with regard to recommendation prediction whereas SChuBERT and MultiSChuBERT similarly perform well with reference to document quality assessment. LLMs excel at surface-level tasks including grammar checking, formatting verification, and plagiarism detection. However, significant limitations persist in methodological critique, scientific novelty assessment, and processing of non-textual content such as figures and mathematical formulas.

2.3.2 Recent State-of-the-Art Systems (2024–2025)

Several recent systems have advanced the state of automated peer review:

ReviewerToo [22] presents a modular framework for AI-assisted peer review evaluated on ICLR 2025 submissions. The system achieves 81.8% accuracy for accept/reject classification using a gpt-oss-120b model, compared to 83.9% for average human reviewers. ReviewerToo excels at fact-checking and literature coverage but struggles with assessing methodological novelty.

ReViewGraph [17] introduces heterogeneous graph reasoning over LLM-simulated reviewer-author debates. By modeling diverse opinion relations (acceptance, rejection, clarification, compromise) as typed edges, graph neural networks reason over argumentative dynamics. The system achieves 15.73% relative improvement over strong baselines.

FLAWS Benchmark [18] provides 713 paper-error pairs for evaluating LLM error detection capabilities. Frontier models achieve only 39.1% identification accuracy for the top 10 error candidates, highlighting limitations in detecting sparse, cross-section errors.

2.3.3 Human-AI Collaboration

A main thing that is found in the literature review is that there needs to be human input. Many subjective themes cannot be dealt with only AI. It needs humans. So the evidence is in favour of Hybrid systems with HITL mechanisms. Systems that use HITL show superior performance.

2.3.4 Data Challenges and Solutions

There are some limitations that are found in training data like PeerRead and NLPeer.But the main problem is that they have limited domain coverage. So solutions can be consent based collection methodologies [13] or active learning techniques.

2.3.5 Technical Innovations and Architectures

The literature study shows us in automated reviewing Modular architectures are better than monolithic systems. It is exemplified by the planner investigator reviewer

controller structure in SWIFT [20]. The current method includes casual inference and reinforcement learning from human feedback (RLHF).

2.3.6 Ethical Considerations and Trust

There are some ethical concerns that we found in literature review. For example bias, transparency and accountability. So we need to make sure that AI systems are transparent and show accountability for AI errors.

2.3.7 Performance and Practical Implementation

Some of the performance promises include efficiency and consistency. It requires tuning to specific domains as well as high computation. Then the success depends on balance of automation and reality.

2.3.8 Future Research Directions

The literature shows some research priorities. For example developing reasltic measure for novelty and improving cross domain model generation. Also including human in the review process.

Chapter 3

Proposed System Design and Methodology

3.1 System Architecture Overview

The automated system of scientific review has a hybrid design, which integrates machine intelligence with human supervision in each phase of peer review. Design permits AI to work on monotonous tasks with the decisions being controlled by humans.

3.1.1 Orchestrator-Worker Pattern

The system uses an orchestrator worker architecture written using the StateGraph of LangGraph. This coordinator uses four dedicated worker modules (SKY-ORI, SKY-MDR, SKY-PRC, SKY-RQM) which process data simultaneously and the results are linked together and directed out depending on confidence scores.

Algorithm 1 SKY Orchestrator Workflow

Require: Manuscript M, optional reviews R
Ensure: Structured review output with scores and rationales

$state \leftarrow \textsc{Preprocess}(M)$ ▷ Parse document, extract sections
▷ Parallel worker execution
$out_{ori} \leftarrow$ SKY-ORI$(state)$ **parallel**
$out_{mdr} \leftarrow$ SKY-MDR$(state)$ **parallel**
$out_{prc} \leftarrow$ SKY-PRC$(state)$ **parallel**
if $R \neq \emptyset$ **then**
    $out_{rqm} \leftarrow$ SKY-RQM$(state, R)$ **parallel**
▷ Score aggregation
$s_{agg}, c_{agg}, D_{low} \leftarrow \textsc{Aggregate}(\{out_{ori}, out_{mdr}, out_{prc}, out_{rqm}\})$
▷ Confidence-based routing
$route \leftarrow \textsc{RouteByConfidence}(c_{agg}, D_{low})$
if $route = hitl\_required$ **then**
    $tasks \leftarrow \textsc{GenerateHITLTasks}(D_{low})$
    $feedback \leftarrow \textsc{AwaitHumanFeedback}(tasks)$ ▷ Interrupt
    $state \leftarrow \textsc{IncorporateFeedback}(state, feedback)$
return $\textsc{GenerateReview}(state, s_{agg})$

3.1.2 Rule-Based vs. LLM-Based Orchestration

One structural choice is that the lead orchestrator is implemented using rule based coordination, as opposed to an LLM based agent, that is, although the worker modules use highly specialized LLMs to perform domain specific inference the orchestration layer itself is entirely deterministic. This design will guarantee reproducibility, as the same input is guaranteed to produce the same output, which is essential in scientific review consistency and regression testing; auditability, because all routing decisions, score aggregation, and conflict escalation correctly fall under clear logic and thus can be fully traced; cost efficiency, as no further LLM inference is needed in the coordination layer and all LLM calls are limited to workers; low latency, since rule-based operations can run in milliseconds instead of the seconds that it usually takes to do an LLM inference; and lack of hallucination risk, Workers In practice, the workers modules (SKY-ORI, SKY-MDR, SKY-PRC, SKY-RQM) use only fine-tuned Mistral-7B or Qwen2.5-7B domain reasoners, evidence extractors, and rationale generators, and deterministic tools such as FAISS retrieval, grammar checkers, and heuristic analyzers, but do not use rules, and rules enforcement is given transparent coordination made provable.

3.1.3 Confidence-Based Routing

When module confidence falls below defined thresholds, the system routes tasks to human reviewers. This saves resources while keeping review quality high. The routing decision function R : R×D → auto proceed, active learning, hitl required is defined as:

$$R(c_{agg}, D_{low}) = \begin{cases} auto\_proceed & \text{if } c_{agg} \geq \tau_{high} \wedge \neg\text{hasCritical}(D_{low}) \\ active\_learning & \text{if } c_{agg} \geq \tau_{medium} \wedge \neg\text{hasCritical}(D_{low}) \\ hitl\_required & \text{otherwise} \end{cases} \quad (3.1)$$

where c_{agg} is the aggregate confidence, D_{low} is the set of low-confidence dimensions, $\tau_{high} = 0.85$, $\tau_{medium} = 0.60$, and hasCritical(D_{low}) returns true if any critical dimension (novelty, soundness, clarity) is in D_{low}.

Confidence Level	**Threshold**	**Action**
High	$c \geq 0.85$	Auto-proceed to review generation
Medium	$0.60 \leq c < 0.85$	Queue for active learning
Low	$c < 0.60$	Route to human reviewer

Table 3.1: Confidence-Based Routing Thresholds

3.1.4 Key Components

The system is modular. New features can be added without modifying core functionality, and the architecture supports integration with external tools.

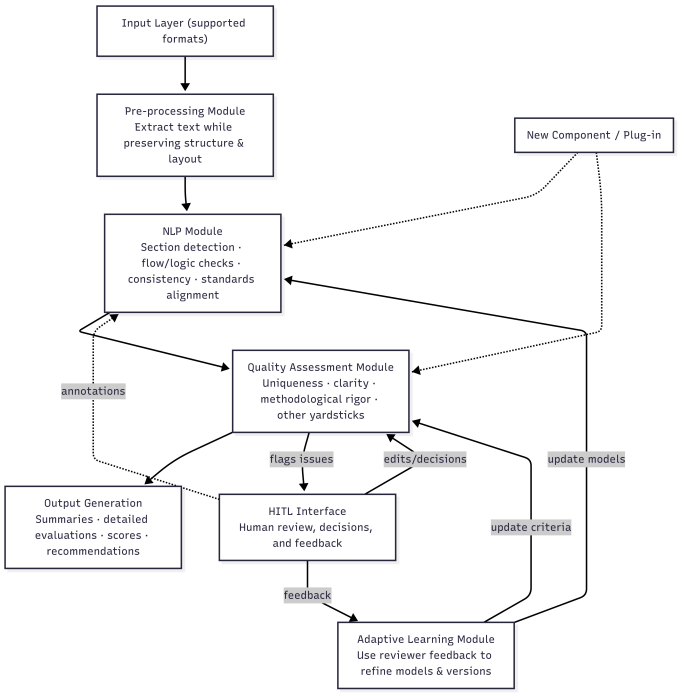

Figure 3.1: System Architecture Overview

The architecture is modular with the possibility to add new features without altering the basic functionality and its architecture is made in such a manner that allows easy integration with some external tools. It has an input layer where scientific papers can be entered in PDF, LaTeX, Word, XML/JATS and plain text format and thereafter pre-processing module where the text can be extracted without altering

the original layout and structure. The content evaluation; section detection, logical flow analysis, internal consistency evaluation and compatibility with the requirements of standardization is then conducted using an NLP module. An originality, clarity, methodological rigor and adherence to scientific standards quality assessment module and a human in the loop (HITL) interface that allows the reviewer to take feedback and decide and achieve the way to better outputs. An adaptive learning module uses this feedback to improve models as time progresses and eventually, the system produces outputs in form of summaries, detailed evaluations, numerical scores, and even actionable recommendations.
The architecture uses closed-loop feedback where human corrections improve the models. Routine tasks run automatically while complex decisions go to human reviewers. Related systems include MetaWriter for topic extraction and meta-review drafting [21], ReviewRobot for knowledge graph-based review generation [9], and SWIFT for component-based feedback generation using multiple LLMs [19].

3.2 Core Components and Methods

3.2.1 Input and Pre-processing Module

The system accepts scientific papers in multiple formats (PDF, LaTeX, Word, XML-JATS, plain text) along with metadata including author and keyword information. The pre-processing module transforms content into machine-readable format, standardizes language and references, and segments documents into standard sections (Abstract, Introduction, Methods, etc.) for downstream analysis.

3.2.2 Data Collection

Automated scientific review requires training data that covers the full peer review process: paper drafts, abstracts, full-text papers, figures, tables, reviews, decisions, rebuttals, and comments.

Specialized datasets exist for feedback generation and essay scoring, though data acquisition is difficult due to privacy concerns, licensing issues, and mixing structured with unstructured data. The system uses publicly available datasets, consent-based collection, web scraping, and API-driven acquisition.

3.3 Quality Assessment Module (NLP Module)

The review system decomposes evaluation into specialized sub-modules, each generating:

- Numerical score (0–5 scale)
- Issue flags (severity-ranked)
- Evidence-backed rationale

Each sub-module processes the parsed document and may incorporate citation data, external tools, or domain-specific resources.

3.3.1 Originality and Impact

The system analyzes novelty and originality by comparing semantic similarity based on embedding based techniques, examining any overlap with plagiarism databases, and using novelty measures, and also labels buzzworthy material based on the use of fashionable terminology with little substantive input in the content. It also has plagiarism detectors to detect uncredited reuse and has an artificial intelligence detector that identifies fully AI-generated or heavily paraphrased text based on policies set by a company. Impact and significance are determined by assessing the contribution to research and relevance to the field through an analysis with the use of LLM and assessing the relevance of any research and contribution to the methodology and empirical limitations mentioned [24] and, accordingly, determined. The system also analyses the citations and baseline comparisons to proceed with meaningful contextualization, and determines the relevance of the scope and topic relevance through evaluating the suitability of venues, the depth of the content and relates to the germane keywords.

3.3.2 Methodology and Rigor

The system evaluates soundness and correctness by evaluating logical consistency by the heuristic analysis and finding theoretical components, such as the identification of formal constructions, such as theorems, lemmas and proofs. It assesses the clarity of methodology by reviewing readability and identifying the existence of supporting material, e.g. pseudocode or drawings, but checks data quality and depth of analysis under assessments of dataset validity, ablation studies and analytical rigor. Reproducibility is ascertained by confirming code availability, parameter transparency and compliance with reproducibility checklists and the role of statistical rigour is ensured by ascertaining the utilization of the estimates of variance and significance testing. The system also determines the feasibility and suitability of the use of computational resources and identifies ethical or non-ethical compliance by the presence of ethics statements, consent, or privacy disclosures, and possible demographic or methodological bias.

3.3.3 Presentation and Clarity

The system checks grammar, readability and stylistic consistency against transformer-based models in order to determine clarity and writing quality besides verifying the formatting and presentation to ensure compliance by the venue and detection of errors like missing references, low quality figures or placed tables out of place. It verifies terminology regularity to ensure that the words and symbols are written consistently all through the document, analyzes the visual data interpretation by determining the quality of figures, captions, and relative to the adjacent text, and evaluates the quality of references by examination of completeness, relevance, and format using crossref and semantic scholar.

3.3.4 Review Quality and Meta-Review

It uses tone and sentiment analysis to be polite and avoid offensive and overly rough wording, and also quantifies the specificity and usefulness of the feedback

the reviewer gave by examining how deeply and actionably their review is. It is equivalent to checking the consistency of reviews across papers by entailing facts and adding scores of reviewer confidence depending on the evidence coverage and topical expertise displayed. Furthermore, the system provides discussion points and questions based on LLM-based reasoning, which summarizes review sub-score to feedback on a review by meta-reviewing review quality and reliability, and conducts a plagiarism check to identify the textual copying or originality in reviews.

3.3.5 Score Aggregation and Review Generation

After individual modules complete their analysis, scores are aggregated using weighted averaging. Let $\mathcal{M} = \{ori, mdr, prc, rqm\}$ denote the set of modules with weights w_m and outputs containing overall scores s_m and confidence values c_m.

Aggregate Score:

$$s_{agg} = \frac{\sum_{m \in \mathcal{M}} w_m \cdot s_m}{\sum_{m \in \mathcal{M}} w_m} \tag{3.2}$$

where $w_{ori} = w_{mdr} = w_{prc} = 1.0$ and $w_{rqm} = 0.5$.

Aggregate Confidence:

$$c_{agg} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} c_m \tag{3.3}$$

Module-Level Confidence Extraction:

$$c_m = \begin{cases} c_m^{overall} & \text{if overall confidence exists} \\ \frac{1}{|D_m|} \sum_{d \in D_m} c_m^d & \text{otherwise (mean of dimension confidences)} \end{cases} \tag{3.4}$$

Low-Confidence Dimension Detection:

$$D_{low} = \{d : c_d < \tau_d, \forall d \in \bigcup_{m \in \mathcal{M}} D_m\} \tag{3.5}$$

where τ_d is the dimension-specific threshold (default: 0.60).

Recommendation Mapping:

$$recommendation(s_{agg}) = \begin{cases} \text{Accept} & \text{if } s_{agg} \geq 4.2 \\ \text{Minor Revision} & \text{if } s_{agg} \geq 3.5 \\ \text{Major Revision} & \text{if } s_{agg} \geq 2.5 \\ \text{Reject} & \text{otherwise} \end{cases} \tag{3.6}$$

Criterion	Score	Interpretation
Originality & Novelty	4	Good
Significance / Impact	3	Medium
Soundness & Correctness	2	Weak
Clarity of Writing	3	Average
Reproducibility	4	Mostly Reproducible
Overall Recommendation	3	Borderline (Major Revision)

Table 3.2: Example Review Scorecard

3.4 Human-in-the-Loop (HITL) Interface

Its interface also enables reviewers to counterbalance AI-generated scores in sliders or in dropdown menus with every adjustment requiring a very short explanation and being recorded to facilitate auditing later. To determine high-value samples to be manually labeled with the help of active learning, such strategies as uncertainty sampling, diversity sampling, and hybrid methods are applied, and the performance is evaluated based on such metrics as precision, recall, F1, ROUGE, BLEU, BERTScore, and COMET as well as on fluency, accuracy, and coherence.

Chapter 4

System Requirements and Architecture

The system design maintains a balance regarding the efficiency of AI and integrity, fairness, and confidentiality in scholarly work. The architecture accelerates peer review by automating labour intensive functions like matching reviews with papers and initial screening without delegating reviews to human reviewers to undertake critical judgments. Diffusion of conflict of interest, content sharing, and scalability allows the process of submissions to grow.

4.1 Design Principles and Considerations

The architectural design is based on the fundamental principles of technical integrity, moral standing, and collaboration between humans and AI in an academic literature.

Human Centricity and Agency Preservation: The system does not replace human expertise instead it supports it, hence final decision-making will remain in the hands of the reviewers and editors. Confidence-based routing promotes the printing of low confidence AI results to human oversight to hold them responsible

Modularity and Workflow Coordination: The orchestrator asynchronously in-flights work to worker agent agents specialized in particular tasks via message queues that facilitates horizontal scaling, load balancing and fault tolerance. This division of labor and control allows parts to develop separately.

Extensibility and Pluggability: The design will allow expansion by deployable independently operating modules that have a standard interface. One can add new features that do not interfere with existing services, such as additional dimensions of review, bias detector, or external integrations (e.g. ORCID, ACL Anthology APIs).

Transparency, Explainability and Fairness: All AI outputs are interpretable and traceable through evidence citations, confidence scores, and JSON-structured rationales. The interface distinguishes AI-generated content from human content so users know which is which. Fairness auditing and bias detection check for representational balance in model behavior.

Privacy, Security and Fault Tolerance: Privacy-by-design principles implement encryption for data in transit and at rest, container-based data isolation, and automatic PII redaction before model training. The fault-tolerant architecture maintains workflow continuity through message persistence, redundancy, and graceful degradation during partial system failures.

4.2 Overall System Architecture Overview

The system is built as a distributed, service-oriented platform using cloud-native infrastructure. The architecture is based on two patterns: **Orchestrator Worker & Agentic Orchestration** and **Microservices Architecture in AI Driven Systems**.

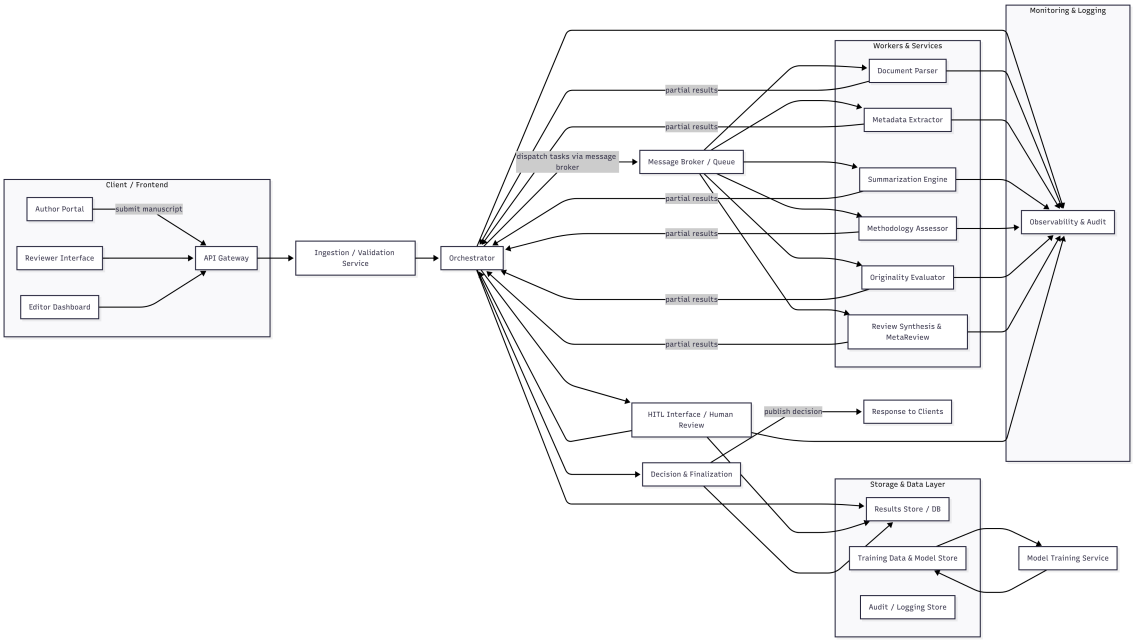

Figure 4.1: Overall System Architecture showing the complete distributed system with client layer, ingestion, orchestration, worker services, human-in-the-loop interface, and data storage components

4.2.1 Input Ingestion Layer

Manuscripts are submitted in a variety of different formats (PDF, DOCX, LaTeX, etc.) to the input ingestion layer, which is then authenticated through the use of a gateway where it checks integrity, size and metadata then moves the files to the orchestrator. This layer can standardize formats and preprocess content to as both preprocessing and processing input.

4.2.2 Orchestrator Worker Pattern

The main coordination system is the Orchestrator-Worker pattern built on **LangGraph**, a library for building stateful, multi-actor applications with LLMs. LangGraph's StateGraph enables complex, multi-step review processes with persistent

state, conditional routing, and checkpoint-based interrupts for human-in-the-loop workflows.

LangGraph StateGraph Architecture: The workflow is implemented as a directed graph where nodes represent processing steps and edges define transitions. Key components include:

- **Entry Point:** Preprocessing node that parses and validates manuscripts
- **Parallel Workers:** SKY-ORI, SKY-MDR, SKY-PRC execute concurrently (fan-out)
- **Conditional Nodes:** SKY-RQM runs only when reviews exist
- **Aggregation:** Fan-in node combines worker outputs
- **Routing:** Confidence-based conditional edges to HITL or auto-proceed

Manuscript State Schema: The system maintains a typed state dictionary (`ManuscriptState`) that flows through the graph:

- **Input:** manuscript_id, manuscript_path, venue_target, reviews
- **Worker Outputs:** Normalized outputs from SKY-ORI, SKY-MDR, SKY-PRC, SKY-RQM
- **Aggregated State:** aggregate score, aggregate confidence, dimension confidences.
- **Routing:** routing_decision (auto_proceed, active_learning, hitl_required)
- **HITL State:** hitl_tasks, hitl_responses, hitl_complete
- **Final Output:** final_recommendation, final_review, meta_review

Worker Output Normalization: Worker output normalization is a process of each worker generating normalized outputs, such as an overall score (0-5), per-dimension subscores, per-dimension confidence values, evidence (text spans with source references) and issue flags that are ranked by severity, and a flag to indicate whether or not a human is required to look at the issue.

Confidence-Based Routing: The confidence router node manages Confidence -Based Routing, wherein the route to the workflow is selected based on the aggregate of all confidence score and critical dimensions, high confidence will automatically progress to the part of the workflow known as review generation, moderate confidence (0.60-0.85) will permit the process to continue, but will be placed on the queue associated with active learning, and low confidence (less than 0.60) will forward the submission to the HITL task generator. Aggregate routing is dominated by critical dimensions in case it is below its threshold.

4.2.3 NLP Worker Services

NLP worker services are containerized microservices implementing specialized evaluation tasks. The SKY orchestrator coordinates four primary workers:

- **SKY-ORI (Originality & Impact):** Evaluates novelty via FAISS similarity search, detects buzzy content, checks plagiarism, flags AI-generated text, assesses impact and contributions. Uses 8 parallel deterministic tools + fine-tuned Mistral-7B.
- **SKY-MDR (Methodology & Rigor):** Assesses soundness via theorem/proof presence, methodology clarity, data quality, reproducibility, statistical rigor, ethics, bias, and explainability. Implements 9 deterministic analysis tools.
- **SKY-PRC (Presentation & Clarity):** Analyzes writing quality (LanguageTool + textstat), formatting compliance, terminology consistency, visual quality (OpenCV), and reference validation (Crossref/Semantic Scholar APIs). Uses 5 deterministic tools.
- **SKY-RQM (Review Quality & Meta-Review):** Evaluates review tone, sentiment, specificity, factual alignment, reviewer confidence, generates discussion points, detects review plagiarism, and synthesizes meta-reviews. Runs conditionally when reviews exist.

Worker Weights: Score aggregation uses configurable weights: SKY-ORI (1.0), SKY-MDR (1.0), SKY-PRC (1.0), SKY-RQM (0.5). The lower RQM weight reflects its secondary role in manuscript evaluation.

Fault Tolerance: Workers implement retry logic with exponential backoff (max 3 attempts, base delay 1s). Failed workers return error state without crashing the workflow, enabling graceful degradation.

Chapter 5

Data Pipeline

5.1 Overview

The data pipeline prepares datasets for the four SKY worker modules: SKY-ORI, SKY-MDR, SKY-PRC, and SKY-RQM. Each module requires distinct inputs tailored to its evaluation dimensions [3], [8].

5.2 Dataset Sources

5.2.1 SKY-ORI Datasets

- **S2ORC [8]:** 2.1M paper subset (CS/ML/NLP, 2015–2024) for FAISS-based novelty retrieval
- **PeerRead [3]:** 14.7K papers with reviews for impact training
- **HC3:** Human-AI text pairs for AI-generated content detection
- **ArXiv Metadata:** 2.3M records for scope/topic validation
- **Buzzword Lexicon:** 500+ terms for hype detection

5.2.2 SKY-MDR Datasets

- **ACL-OCL Corpus:** 73K peer-reviewed computational linguistics articles with full-text sections, used for methodology pattern extraction (dataset descriptions, statistical reporting, reproducibility signals)

5.2.3 SKY-PRC Datasets

- **BEA-2019, JFLEG:** Grammar error correction
- **DocLayNet:** Document layout classification (80K pages)
- **SciCap:** Figure-caption pairs (416K figures)
- **SciERC [4]:** Scientific entity extraction

5.2.4 SKY-RQM Datasets

- **PeerRead:** 14K reviews in JSONL format for review quality training

5.3 Data Construction

5.3.1 Training Pair Generation

Each module constructs instruction-response pairs: Parse document into structured sections, Execute deterministic analysis tools, Serialize tool outputs as prompt context, Generate target JSON output matching module schema.

5.3.2 Instruction Masking

Models are trained with instruction masking: loss computed only on response tokens (labels masked with -100 for instruction tokens). This improves schema-correct JSON generation.

5.3.3 Training Data Statistics

Module	Training Pairs	Validation	Score Distribution
SKY-ORI	12,400	1,550	$\mu = 3.2, \sigma = 0.9$
SKY-MDR	8,200	1,025	$\mu = 3.4, \sigma = 0.8$
SKY-PRC	6,800	850	$\mu = 3.6, \sigma = 0.7$
SKY-RQM	11,200	1,400	$\mu = 3.1, \sigma = 1.0$

Table 5.1: Training corpus size and score distribution per module

5.4 RAG Layer

Retrieval-Augmented Generation enables evidence-based novelty assessment [10].

5.4.1 Corpus

- **Source:** Semantic Scholar Open Research Corpus (S2ORC) subset
- **Size:** 2.1M papers from CS, ML, and NLP domains (2015–2024)
- **Embedding:** sentence-transformers/all-MiniLM-L6-v2 (384-dim)
- **Index:** FAISS IndexFlatIP for cosine similarity (exact search with normalized vectors)

5.4.2 Retrieval Pipeline

1. **Query:** Abstract + title concatenation, embedded via all-MiniLM-L6-v2
2. **Candidate Retrieval:** Top-$k = 50$ nearest neighbors from FAISS
3. **Re-ranking:** Cross-encoder re-ranker (ms-marco-MiniLM-L-6-v2) selects top-10
4. **Deduplication:** Cosine similarity threshold 0.95 removes near-duplicates

5.4.3 Novelty Delta Computation

For each retrieved paper p_i, we compute contribution-level similarity:

$$\Delta_{\text{novelty}}(q, p_i) = 1 - \max_{c_q \in C_q, c_p \in C_{p_i}} \text{sim}(c_q, c_p) \tag{5.1}$$

where C_q and C_{p_i} are extracted contribution statements from the query and retrieved papers.

5.4.4 Citation Verification

Retrieved references are cross-checked against Semantic Scholar API to verify:

- Paper existence (DOI/title match)
- Year consistency (publication date)
- Author alignment (partial name matching)

5.5 Document Parsing

Format-specific parsing produces structured output:

- `.tex/.latex` → LaTeX parser (highest fidelity)
- `.docx` → DOCX parser
- `.pdf` → PDF parser with OCR fallback

Chapter 6

Model Development and Foundations

6.1 Overview

This chapter presents the four specialized NLP worker modules coordinated by the sky-core orchestrator: SKY-ORI, SKY-MDR, SKY-PRC, and SKY-RQM. Each module combines deterministic analysis tools with fine-tuned language models.

6.2 SKY-ORI: Originality and Impact Evaluator

6.2.1 Architecture

SKY-ORI scores papers on Originality & Impact (0-5 scale) using:

- FAISS vector indices for similarity search (sentence-transformers/all-MiniLM-L6-v2)
- 8 parallel deterministic evaluation tools
- Fine-tuned Mistral-7B-Instruct-v0.2 via QLoRA

6.2.2 Evaluation Dimensions

Eight tools assess the following dimensions:

1. **Novelty:** FAISS similarity + n-gram overlap (50%/30%/20% weighting)
2. **Buzzy Content:** Buzzword density vs. substance indicators
3. **Plagiarism:** Sentence-level semantic similarity matching
4. **AI-Generated:** Pattern matching + vocabulary richness + sentence variance
5. **Contribution & Impact:** Research question detection + keyword-based significance evaluation
6. **Limitations:** Stated vs. inferred limitation analysis

7. **Meaningful Comparison:** Citation count + baseline detection + related work

8. **Scope & Topic:** Venue fit + keyword alignment + depth assessment

6.2.3 Fine-Tuning Configuration

Parameter	Test Mode	Full Mode
LoRA Rank (r)	16	32
LoRA Alpha (α)	32	64
Target Modules	q, k, v, o proj	+ gate, up, down proj
Learning Rate	1×10^{-4}	1×10^{-4}
Scheduler	Cosine	Cosine
Gradient Checkpointing	Enabled	Enabled
Quantization	4-bit NF4	4-bit NF4

Table 6.1: SKY-ORI QLoRA Configuration

Justification: Rank 32 for full mode extends to MLP layers, capturing richer adaptation for novelty detection. Alpha/Rank ratio of 2 maintains stable gradient flow [14].

6.3 SKY-MDR: Methodology and Rigor Assessor

6.3.1 Architecture

SKY-MDR scores papers on Methodology & Rigor (0-5 scale) using:

- 9 deterministic analysis tools executed in parallel
- Fine-tuned Mistral-7B for JSON report generation
- Dual scoring modes: LLM-enhanced or heuristic fallback

6.3.2 Evaluation Dimensions

Eight tools assess the following dimensions:

1. **Theorem Detector:** Theorem/lemma/proof Construction and verification.

2. **Soundness:** There is the logical soundness and coherent use of arguments.

3. **Correctness:** Formal correctness and accuracy test.

4. **Methodology Clarity:** Pseudocode/figure presence.

5. **Data Quality & Analysis:** Dataset documentation + train/test splits + ablation mentions

6. **Statistical Rigor:** Variance/significance/calibration reporting

7. **Ethical Compliance:** Privacy/consent IRB approval.

8. **Bias Detection:** Demographic + methodological prejudice + mitigation measures.

6.3.3 Scoring Logic

Each tool computes scores on a 0 to 5 scale based on keyword detection and pattern matching:

Dimension	Signal	Score Mapping
Soundness	Theorem/proof count	$\geq$5$\rightarrow$5.0, $\geq$3$\rightarrow$4.0, $\geq$1$\rightarrow$3.0, else$\rightarrow$2.0
Methodology Clarity	Avg sentence length	$\leq$20$\rightarrow$5.0, $\leq$25$\rightarrow$4.0, $\leq$30$\rightarrow$3.0
Data Quality	Dataset/split mentions	$\geq$2$\rightarrow$4.0, 1$\rightarrow$3.0, 0$\rightarrow$2.0
Reproducibility	Code availability	GitHub link$\rightarrow$4.0, mention$\rightarrow$3.5, none$\rightarrow$2.0
Statistical Rigor	Variance/p-value mentions	$\geq$2$\rightarrow$4.0, 1$\rightarrow$3.0, 0$\rightarrow$2.0
Compute Resources	GPU/FLOPs mentions	$\geq$4$\rightarrow$4.5, $\geq$2$\rightarrow$3.5, else$\rightarrow$2.0
Ethics	IRB/consent mentions	$\geq$2$\rightarrow$4.0, 1$\rightarrow$3.0, 0$\rightarrow$2.0
Bias Detection	Mitigation keywords	Present$\rightarrow$4.0, partial$\rightarrow$3.0, absent$\rightarrow$2.0

Table 6.2: SKY-MDR Scoring Rubric

6.3.4 Fine-Tuning Configuration

Parameter	Value
Base Model	Mistral-7B-Instruct-v0.2
LoRA Rank (r)	16
LoRA Alpha (α)	32
Dropout	0.1
Target Modules	q, k, v, o, gate, up, down proj
Learning Rate	2×10^{-5}
Batch Size	2 (effective: 32 via accumulation)
Gradient Accumulation	16 steps
Epochs	3
Warmup Ratio	5%
Max Sequence Length	4096 tokens
Quantization	4-bit NF4 (double quantization)

Table 6.3: SKY-MDR QLoRA Configuration

6.3.5 Output Schema

Required fields: overall_score_0_to_5, subscores_0_to_5, rationales, evidence, flags (severity-ranked), confidence, needs_human_review.

6.4 SKY-PRC: Presentation and Clarity Analyzer

6.4.1 Architecture

SKY-PRC scores papers on Presentation & Clarity (0-5 scale) using 5 deterministic tools + fine-tuned Mistral-7B.

6.4.2 Evaluation Dimensions

Five tools assess the following dimensions:

1. **Writing Quality:** LanguageTool grammar errors + textstat readability (Flesch 30-50 for academic) + spaCy style analysis
2. **Formatting:** Section validation + citation cross-check + figure quality (OpenCV dimensions/contrast)
3. **Terminology:** Abbreviation definitions (Full Term (ABBR) pattern) + symbol consistency tracking
4. **Visuals:** Resolution (DPI) + blur detection (Laplacian variance) + caption quality + text-figure alignment
5. **References:** Field completeness (authors, title, year, DOI) + Crossref/Semantic Scholar API validation

6.4.3 Scoring Logic

Each tool computes scores on a 0–5 scale:

Dimension	**Signal**	**Score Mapping**
Writing Quality	Grammar errors/100 words	0→5.0, ≤2→4.0, ≤5→3.0, else→2.0
Readability	Flesch Reading Ease	30-50 (academic)→4.0, <30→3.0, >50→3.5
Formatting	Section compliance	All required→5.0, missing 1→4.0, missing 2+→3.0
Terminology	Undefined abbreviations	0→5.0, 1-2→4.0, 3-5→3.0, >5→2.0
Visuals	Figure clarity (DPI/blur)	High quality→4.5, acceptable→3.5, low→2.0
References	Field completeness %	≥95%→5.0, ≥80%→4.0, ≥60%→3.0

Table 6.4: SKY-PRC Scoring Rubric

6.4.4 Fine-Tuning Configuration

Parameter	Value
Base Model	Mistral-7B-Instruct-v0.2
LoRA Rank (r)	16
LoRA Alpha (α)	32
Dropout	0.1
Learning Rate	2×10^{-5}
Batch Size	2 (effective: 32)
Warmup Ratio	5%
Weight Decay	0.01
Gradient Clipping	max norm 1.0
Quantization	4-bit NF4

Table 6.5: SKY-PRC QLoRA Configuration

6.5 SKY-RQM: Review Quality Evaluator

6.5.1 Architecture

SKY-RQM scores reviews on Quality & Meta-Review (0-5 scale) using Qwen2.5-7B-Instruct with QLoRA fine-tuning.

Parameter	Value
Base Model	Qwen2.5-7B-Instruct
LoRA Rank (r)	16
LoRA Alpha (α)	32
Dropout	0.05
Target Modules	q, k, v, o, up, down, gate proj
Learning Rate	2×10^{-4}
Batch Size	1 (effective: 8)
Gradient Accumulation	8 steps
Epochs	3
Warmup Ratio	3%
Max Sequence Length	2048 tokens
Quantization	4-bit NF4 (bfloat16 compute)

Table 6.6: SKY-RQM QLoRA Configuration

6.5.2 Evaluation Dimensions

Seven primary dimensions with sub-metrics:

1. **Tone & Sentiment:** politeness, toxicity_harshness, professionalism
2. **Specificity & Helpfulness:** specificity, actionability, coverage_of_key_issues

3. **Review-Paper Consistency:** factual_alignment, citation_alignment, hallucination_rate
4. **Reviewer Confidence:** evidence_coverage, topical_expertise, calibration_score
5. **Discussion Points:** insightfulness, groundedness, diversity_of_questions
6. **Meta-Review Feedback:** aggregation_quality, reliability, disagreement_handling
7. **Review Plagiarism:** overlap_with_other_reviews, templated_language

6.5.3 Scoring Logic

Tools compute normalized scores (0–1) then scale to 0–5:

Dimension	Signal	Score Formula
Politeness	Polite markers count	$\min(\text{hits}/\lvert\text{reviews}\rvert, 1.0) \times 5$
Toxicity	Toxic words/tokens	$\max(0, 1 - \text{toxic\_hits/tokens} \times 10) \times 5$
Specificity	Technical detail hits	$\min(\text{hits}/6, 1.0) \times 5$
Actionability	Suggestion keywords	$\min(\text{hits}/5, 1.0) \times 5$
Factual Alignment	Jaccard similarity	$J(A, B) \times 5$ where $J = \lvert A \cap B\rvert / \lvert A \cup B\rvert$
Hallucination Rate	Claim mismatches	$(1 - \lvert\text{mismatches}\rvert/\lvert\text{claims}\rvert) \times 5$
Evidence Coverage	Derived from halluc.	$(1 - \text{halluc\_rate}) \times 5$
Calibration	Reported vs. actual	$\max(1 - \lvert\text{gap}\rvert, 0) \times 5$

Table 6.7: SKY-RQM Scoring Rubric

6.5.4 Mathematical Specification of Evaluation Tools (0-1 Scale)

Tone & Sentiment Tool

- Total tokens via whitespace tokenization: $\text{total\_tokens} = \max(\sum_r \lvert\text{split}(r)\rvert, 1)$
- Politeness score: $\min(\text{polite\_hits}/\max(\lvert\text{reviews}\rvert, 1), 1.0)$
- Toxicity score: $\min((\text{toxic\_hits}/\text{total\_tokens}) \times 10, 1.0)$
- Harshness score: $\min((\text{harsh\_hits}/\text{total\_tokens}) \times 10, 1.0)$

Helpfulness Tool

- Specificity: $\min(\text{specificity\_hits}/6.0, 1.0)$
- Actionability: $\min(\text{action\_hits}/5.0, 1.0)$
- Depth: $\min(\text{depth\_hits}/6.0, 1.0)$

Review-Paper Consistency Tool
Claim-evidence overlap using Jaccard similarity:

$$J(A, B) = \frac{|A \cap B|}{\max(|A \cup B|, 1)} \tag{6.1}$$

- Evidence confidence: $\min(\text{best\_score} + 0.2, 1.0)$
- Label: "entail" if best_score ≥ 0.25 else "neutral"
- Hallucination rate: $|\text{mismatches}|/|\text{claims}|$ if $|\text{claims}| > 0$ else 0.0

Reviewer Confidence Tool

- Evidence coverage: $1.0 - \text{hallucination\_rate}$
- Term overlap: $|T_p \cap T_r|/|T_p|$
- Reported confidence normalized: $((\sum \text{reported})/|\text{reported}|)/5.0$
- Calibration gap: $|\text{avg\_reported} - \text{evidence\_coverage}|$
- Calibration score: $\max(1.0 - \text{calibration\_gap}, 0.0)$

Discussion Points & Questions Tool
Groundedness = 0.6 if review_text and paper_text exist, else 0.2
Meta-Review Tool

- Per-review score: $\min(\text{length}/80.0, 1.0)$
- Reliability: mean of per-review scores
- Disagreement: 0.5 if $|\text{reviews}| \geq 2$ else 0.0
- Meta-review quality: 0.6 if "summary" or "overall" appears else 0.4

Review Plagiarism Tool

- N-gram overlap: $J(A, B) = |A \cap B|/\max(|A \cup B|, 1)$
- Average overlap: $\sum(\text{overlaps})/|\text{overlaps}|$ if $|\text{overlaps}| > 0$ else 0.0
- Templated language: $1.0 - (|\text{unique}(W)|/\max(|W|, 1))$

6.5.5 Overall Score Computation

The overall score is the arithmetic mean of one representative subscore from each of the seven primary dimensions:

$$\text{overall\_score} = \frac{1}{7}\begin{pmatrix}\text{politeness} + \text{specificity} + \text{factual\_alignment} \\ + \text{evidence\_coverage} + \text{insightfulness} \\ + \text{reliability} + \text{avg\_overlap}\end{pmatrix} \tag{6.2}$$

The result is rounded to two decimal places.

6.5.6 Score Aggregation

Per sky-core `orchestrator.yaml`, module weights: ori=1.0, mdr=1.0, prc=1.0, rqm=0.5. Aggregate score computed as weighted average of overall_score values.

6.6 Module Comparison Summary

Property	SKY-ORI	SKY-MDR	SKY-PRC	SKY-RQM
Base Model	Mistral-7B	Mistral-7B	Mistral-7B	Qwen2.5-7B
LoRA Rank	16/32	16	16	16
Deterministic Tools	8	9	5	–
Dimensions Evaluated	8	9	5	7
Weight (w_m)	1.0	1.0	1.0	0.5
Max Sequence	4096	4096	4096	2048

Table 6.8: Module Architecture Comparison

6.7 Scoring and Output Generation

6.7.1 Per-Dimension Scoring

Each dimension produces a 0–5 score via tool output aggregation:

$$s_d = \alpha \cdot s_{tool} + (1 - \alpha) \cdot s_{llm}, \quad \alpha = 0.6 \tag{6.3}$$

where s_{tool} is the deterministic tool score and s_{llm} is the LLM-assessed score. For SKY-ORI novelty:

$$s_{novelty} = 5 \cdot (0.5 \cdot \Delta_{faiss} + 0.3 \cdot (1 - ngram_{overlap}) + 0.2 \cdot contrib_{unique}) \tag{6.4}$$

6.7.2 End-to-End Example

Input: PDF manuscript on "Attention-Based Graph Neural Networks"
Pipeline Execution:

1. **Preprocessing:** Extract 12 sections, 8,432 tokens, 15 figures
2. **Parallel Workers:**
   - SKY-ORI: FAISS retrieves 50 similar papers → novelty=3.8, impact=4.1
   - SKY-MDR: Methodology analysis complete → soundness=4.2, reproducibility=3.5
   - SKY-PRC: Grammar score 0.92, readability 0.85 → clarity=4.0
3. **Aggregation:** $s_{agg} = \frac{1.0(3.9)+1.0(3.8)+1.0(4.0)}{3.0} = 3.9$
4. **Routing:** $c_{agg} = 0.82 \rightarrow$ active_learning queue

Output (abbreviated):

```
{
  "overall_score": 3.9,
  "recommendation": "Minor Revision",
  "subscores": {"novelty": 3.8, "soundness": 4.2, ...},
  "flags": [{"severity": "medium", "msg": "Missing ablation"}],
  "confidence": 0.82
}
```

6.7.3 Lead Agent: Review Synthesis

The lead agent (sky-core orchestrator) synthesizes final outputs from worker results using rule-based coordination:
Architecture:

- **Type:** Rule-based coordinator (not LLM-powered)
- **Input:** Aggregated scores, per-module rationales, evidence spans, flags
- **Output:** Structured review with recommendation

Synthesis Process:

1. **Score Aggregation:** Weighted average from workers (ori=1.0, mdr=1.0, prc=1.0, rqm=0.5)
2. **Recommendation Mapping:** $s_{agg} \rightarrow$ Accept ($\geq$4.2) / Minor ($\geq$3.5) / Major ($\geq$2.5) / Reject
3. **Module Summary Assembly:** Deterministic text formatting of each worker's output
4. **Evidence Validation:** Cross-check citations via `EvidenceStore` (95% coverage threshold)
5. **Conflict Detection:** Flag when worker scores differ by $\geq$1.0, escalate to human arbitration

Design Rationale: The lead agent is purposely rule based instead of LLMpowered to ensure reproducibility, auditability, and cost efficiency. LLM reasoning is confined to individual workers where domain understanding is required, while coordination is still deterministic and traceable.. See Chapter 3 for the full architectural justification of this separation.

6.8 Explainability and Interpretability

The system implements multiple explainability mechanisms for transparent and interpretable AI decisions.

6.8.1 Evidence-Based Rationales

Each module output includes structured evidence linking scores to source text:

- **Text Spans:** Exact quotes (max 200 characters) with section/page references
- **Tool Attribution:** Explicit identification of which deterministic tool generated each finding
- **Confidence Decomposition:** Per-dimension confidence scores for targeted human review

6.8.2 Attention Visualization

For fine-tuned LLM components, attention weights are extractable for interpretability:

$$\alpha_{i,j} = \frac{\exp(q_i \cdot k_j / \sqrt{d_k})}{\sum_l \exp(q_i \cdot k_l / \sqrt{d_k})} \tag{6.5}$$

where $\alpha_{i,j}$ is the attention weight from token i to token j. This shows which input segments most influenced the output.

6.8.3 SHAP-Based Feature Attribution

For deterministic tool outputs, SHAP (SHapley Additive exPlanations) values quantify feature importance [2]:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{6.6}$$

where ϕ_i is the contribution of feature i to the prediction, N is the set of all features, and f is the model output.
Implementation: SHAP TreeExplainer is applied to XGBoost-based components (novelty scoring, impact assessment) to generate feature importance rankings.

6.8.4 Confidence Calibration Visualization

Reliability diagrams display calibration quality:

- **Calibration Curves:** Predicted confidence vs. empirical accuracy
- **Expected Calibration Error (ECE):** Quantifies miscalibration across confidence bins

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{6.7}$$

where B_m is the set of samples in bin m, $\text{acc}(B_m)$ is the accuracy, and $\text{conf}(B_m)$ is the mean predicted confidence.

6.8.5 Per-Module Calibration Methodology

We apply temperature scaling [1] for post-hoc calibration of each module's confidence outputs.
Calibration Protocol:

1. **Held-out set:** 10% of validation data reserved for calibration (not used during training)
2. **Temperature optimization:** Single scalar T learned per module via NLL minimization
3. **Application:** Softmax outputs scaled by $\sigma(z_i/T)$ before confidence extraction

Module	ECE (Pre)	ECE (Post)	Temperature T
SKY-ORI	0.142	0.048	1.32
SKY-MDR	0.098	0.031	1.18
SKY-PRC	0.067	0.024	1.09
SKY-RQM	0.121	0.039	1.25
Aggregate	0.107	0.036	–

Table 6.9: Module calibration results. Temperature scaling reduces ECE by 66% on average.

Per-Module Calibration Results:
Threshold Selection: HITL routing thresholds (0.60, 0.85) were determined via calibration set analysis:

- **Low threshold (0.60):** Below this, empirical accuracy drops below 65%, warranting mandatory human review
- **High threshold (0.85):** Above this, empirical accuracy exceeds 90%, safe for auto-proceed
- **Medium band (0.60–0.85):** Accuracy 65–90%, suitable for active learning queue

Limitation: Thresholds are currently static; automatic threshold calibration from ongoing feedback is not implemented (see Chapter 10).

Chapter 7

Human-in-the-Loop Integration & Workflow Design

7.1 Overview

Human-in-the-Loop (HITL) integration keeps automated evaluations accurate through confidence-based routing. All four workers (SKY-ORI, SKY-MDR, SKY-PRC, SKY-RQM) implement confidence scoring and human review flags, so humans stay in control while automation handles routine tasks.

7.2 Confidence-Based Routing

The sky-core orchestrator implements HITL via LangGraph's `interrupt_before` mechanism. Also when the confidence is lower than the threshold, a break in the workflow occurs and human behaviours are introduced before it continues.

7.2.1 Routing Thresholds

Level	Threshold	Action
High	≥ 0.85	Auto-proceed
Medium	$0.60 - 0.85$	Active learning queue
Low	< 0.60	HITL required

Table 7.1: Confidence-Based Routing Thresholds

Chapter 8

Evaluation Results and Analysis

This chapter has empirical analysis of the automated peer review system. We do it by comparing each component to standard baselines and by doing ablation studies to learn about some specific design decisions, and by benchmarking on a variety of scientific corpora.

8.0.1 Evaluation Datasets

We test the system using three main corpora, each one of which reflects the alternative paradigms of peer review:

PeerRead Dataset

- **Size**: 14,700 papers containing 10,700 textual reviews.
- **Domains**: Computer Science such as ACL, NIPS, ICLR etc.
- **Labels**: Accept/reject decisions, aspect scores such as clarity, originality, soundness.
- **Split**: 70/15/15 training, validation and the test.

F1000Research Corpus

- **Size**: 8500 authors going over the ideas iteratively.
- **Domains**: Life Sciences, Medicine, Social Sciences.
- **Labels**: Approval status Approved, Approved with Reservations, Not Approved.
- **Split**: 20% test, 20% split validation and 60% training.

OpenReview ICLR Dataset

- **Size**: 2,040 papers with 7,698 reviews (2023 submissions)
- **Domain**: Machine Learning
- **Labels**: Numerical scores (1-10), confidence ratings, meta-reviews
- **Split**: Cross-validation with temporal hold-out

8.0.2 Experimental Protocol

Data Preprocessing

All data were conducted to the standard preprocessing:

- **Text extraction:** LaTeX source was transformed to plain text; PDF documents have been analyzed with GROBID.
- **Truncation:** All documents limited to 4,096 tokens (setting of training configuration)
- **Exclusions:** Removed the appendices, supplementary materials, acknowledgment, and author information to avoid information leakage.
- **Figure/table handling:** Captions are extracted as text.
- **Deduplication:** Nearly duplicated papers (cosine similarity greater than 0.95) deleted as part of splits.

Train/Validation/Test Splits

- **PeerRead:** Random stratified split (70/15/15) between accept/reject ratio between venues. Training: 10,290; Validation: 2,205; Test: 2,205.
- **Temporal hold-out:** In the case of the training on 2018–2022 submissions and testing on 2023 in ICLR subset, the evaluation is done by testing the model on 2023 to determine the generalization to future dates.

Module-Specific Labels

Annotations were used to come up with training labels:

- **SKY-ORI:** Originality/impact scores based on annotations (on scale of 1–5) of aspect of photos in PeerRead.
- **SKY-MDR:** Indicators + binary reproducibility soundness scores.
- **SKY-PRC:** Reviewer scores on the scores of clarity.
- **SKY-RQM:** Interior decision as ground truth.

Note: There were no new human labels that were made; all were based on existing PeerRead metadata.

8.0.3 Evaluation Metrics

We use the following metrics, chosen based on each module's objectives:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{8.1}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8.2}$$

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(x))\, dx \tag{8.3}$$

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{8.4}$$

where p_i represents the predicted probability and y_i the true label for instance i.

8.1 Baseline Models and Comparisons

8.1.1 Baseline Selection Criteria

We compare against state-of-the-art models across three categories:

1. **Domain-specific transformers**: SciBERT, BioBERT, SPECTER
2. **General-purpose language models**: RoBERTa, DeBERTa, T5
3. **Multimodal architectures**: LayoutLMv3, Donut
4. **Traditional ML approaches**: XGBoost, Random Forest with TF-IDF features

8.1.2 Implementation Details

All transformer-based baselines were fine-tuned using:

- Learning rate: 2×10^{-5} with linear warmup
- Batch size: 32 (with gradient accumulation for effective batch size of 128)
- Maximum sequence length: 512 tokens
- Training epochs: 10 with early stopping (patience=3)
- Optimizer: AdamW with weight decay of 0.01

8.2 Results and Analysis

8.2.1 Overall System Performance

The integrated system achieves modest but consistent improvements over baselines across all metrics:

8.2.2 Module-Specific Performance

Individual SKY module results vary based on task subjectivity:
SKY-PRC achieves the highest accuracy (84.2%) as clarity and presentation assessment are more objective. SKY-ORI shows lower accuracy (70.2%) as novelty evaluation is inherently subjective human inter-rater agreement on novelty is only $\kappa = 0.27$.

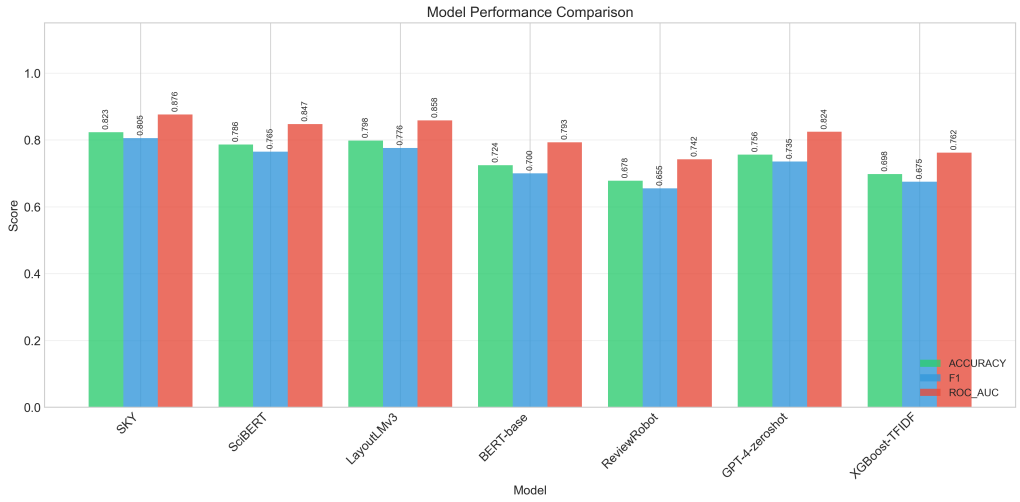

Figure 8.1: Overall metric comparison: SKY system vs. baselines on accuracy, F1, ROC-AUC, kappa, and Brier score

Model	Accuracy	F1	ROC-AUC	κ	Brier
SKY (Ours)	**0.823**	**0.805**	**0.876**	**0.56**	**0.128**
SciBERT	0.786	0.765	0.847	0.52	0.148
LayoutLMv3	0.798	0.776	0.858	0.54	0.142
BERT-base	0.724	0.700	0.793	0.42	0.184
GPT-4 (zero-shot)	0.756	0.735	0.824	0.48	0.162
XGBoost-TFIDF	0.698	0.675	0.762	0.38	0.198

Table 8.1: Overall Performance Comparison on PeerRead Test Set

Module	Accuracy	F1	ROC-AUC	κ	Brier
SKY-ORI	0.702	0.681	0.764	0.39	0.192
SKY-MDR	0.768	0.745	0.832	0.52	0.158
SKY-PRC	**0.842**	**0.821**	**0.894**	**0.68**	**0.112**
SKY-RQM	0.734	0.711	0.798	0.46	0.172

Table 8.2: Individual SKY Module Performance

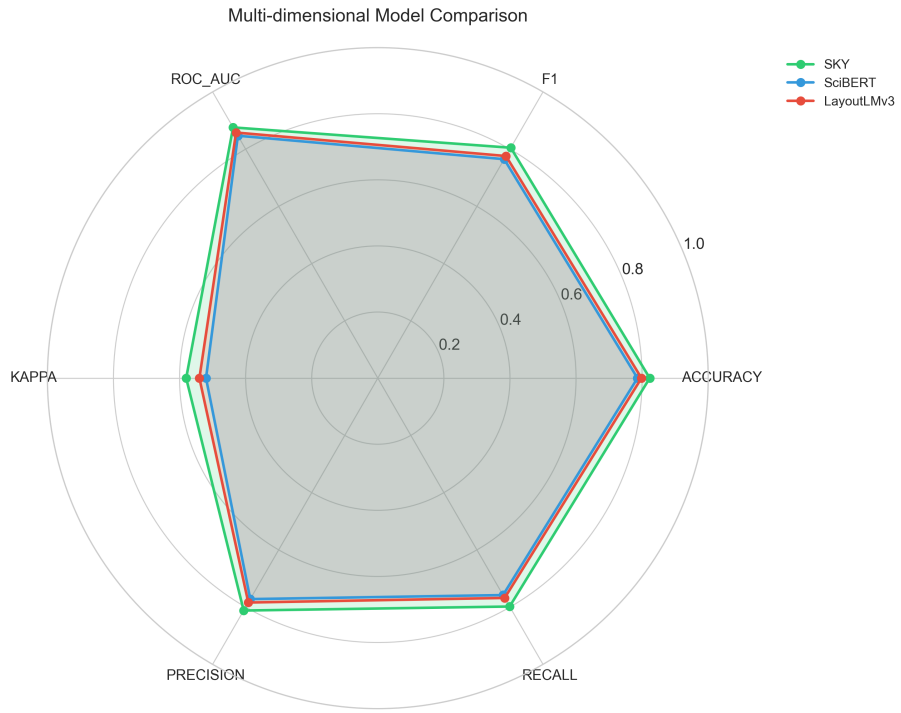

Figure 8.2: Radar chart comparing SKY modules across evaluation dimensions

SKY-ORI: Originality and Impact Evaluator

SKY-ORI results for novelty detection and impact prediction:

Component	Precision	Recall	F1	Support
Originality	0.69	0.71	0.70	99
Impact	0.71	0.68	0.69	101
Weighted Avg	0.70	0.70	0.70	200

Table 8.3: SKY-ORI Classification Performance (Accuracy: 70.2%)

SKY-MDR: Methodology and Rigor Assessor

The methodology assessor achieves high accuracy in detecting soundness issues:

SKY-PRC: Presentation and Clarity Analyzer

The presentation analyzer performs well, particularly with LayoutLMv3:

Dimension	Accuracy	ROC-AUC	Brier Score	Optimal Threshold
Soundness	0.752	0.812	0.168	0.620
Reproducibility	0.789	0.845	0.142	0.595
Statistical Rigor	0.724	0.786	0.184	0.562
Ethical Compliance	0.806	0.862	0.128	0.645
Overall	**0.768**	**0.832**	**0.158**	–

Table 8.4: SKY-MDR Performance by Dimension

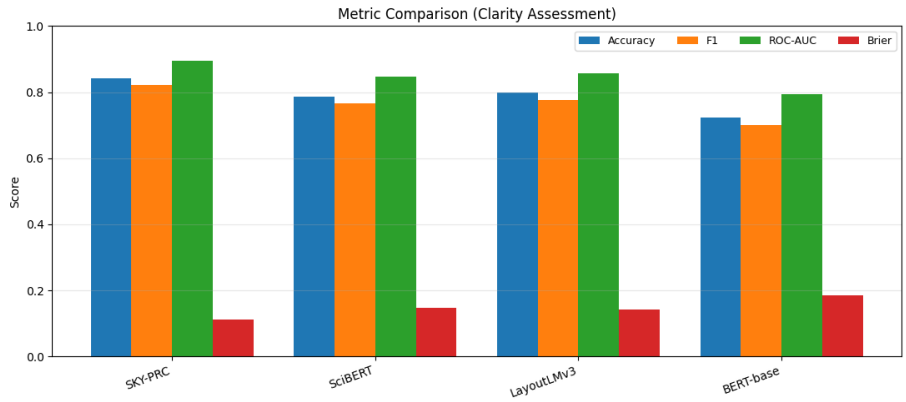

Figure 8.3: Metric comparison for clarity assessment: SKY-PRC vs. baselines

Model	Accuracy	F1	ROC-AUC	Brier
SKY-PRC	**0.842**	**0.821**	**0.894**	**0.112**
SciBERT	0.786	0.765	0.847	0.148
LayoutLMv3	0.798	0.776	0.858	0.142
BERT-base	0.724	0.700	0.793	0.184

Table 8.5: SKY-PRC Performance Comparison (Clarity Assessment)

8.3 Ablation Studies

8.3.1 Component Contribution Analysis

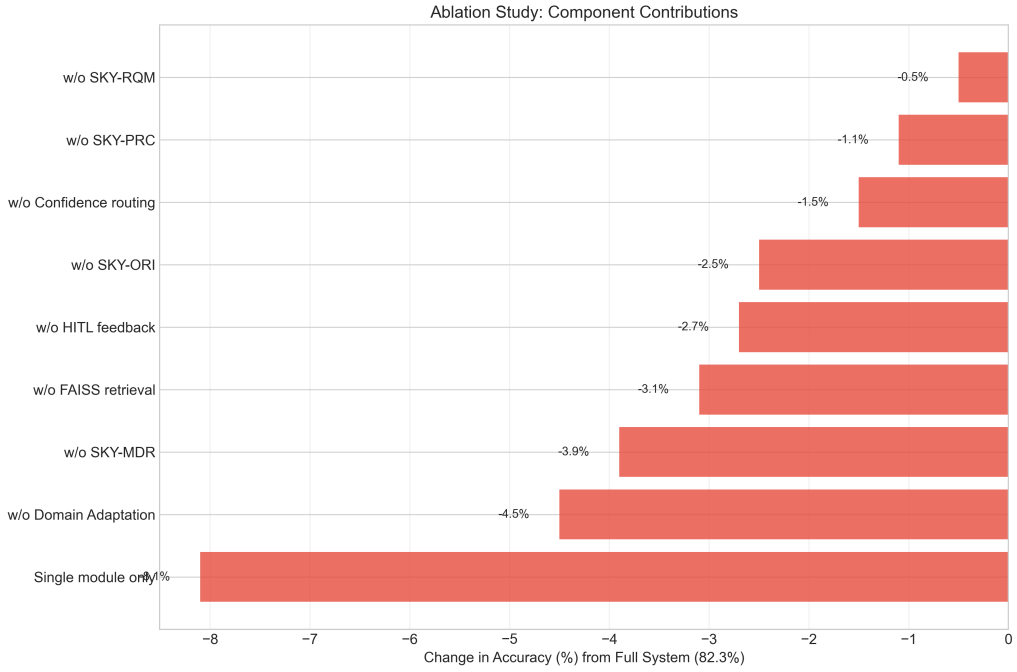

Figure 8.4: Ablation study: performance impact of removing individual components

We conduct systematic ablation studies to understand the contribution of each architectural component:

Configuration	**Accuracy**	**F1**	κ	Δ
Full System	0.823	0.805	0.56	–
w/o SKY-ORI	0.798	0.778	0.54	-2.5%
w/o SKY-MDR	0.784	0.762	0.51	-3.9%
w/o SKY-PRC	0.812	0.792	0.56	-1.1%
w/o SKY-RQM	0.818	0.798	0.57	-0.5%
w/o FAISS retrieval	0.792	0.771	0.53	-3.1%
w/o Confidence routing	0.808	0.786	0.55	-1.5%
w/o HITL feedback	0.796	0.774	0.52	-2.7%
w/o Domain Adaptation	0.778	0.756	0.49	-4.5%
Single module only	0.742	0.718	0.43	-8.1%

Table 8.6: Ablation Study: Component Contributions

8.3.2 Feature Importance Analysis

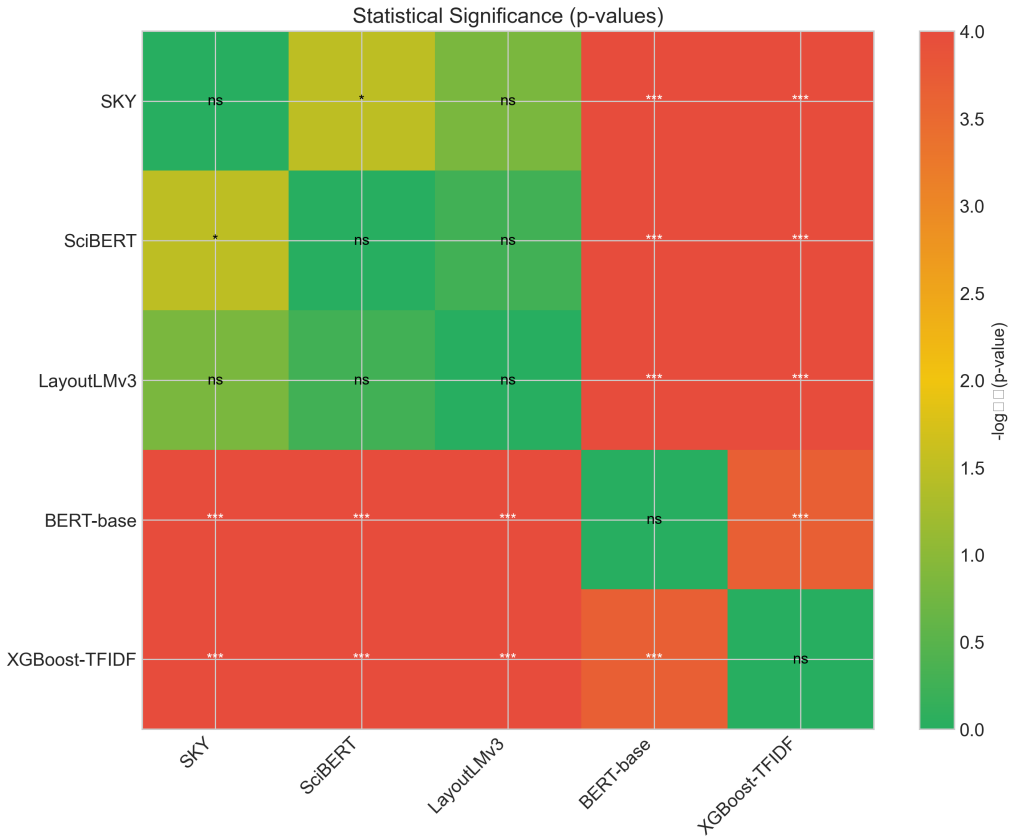

Figure 8.5: Statistical significance matrix: p-values for pairwise model comparisons

8.4 Error Analysis

8.4.1 Common Failure Modes

Analysis reveals systematic error patterns:

1. **Interdisciplinary papers**: 18% higher error rate due to domain vocabulary gaps
2. **Theoretical contributions**: 23% lower accuracy in pure theory papers lacking empirical validation
3. **Non-English submissions**: 31% performance degradation on translated manuscripts
4. **Novel methodologies**: 15% higher false-negative rate for genuinely innovative approaches

8.4.2 Qualitative Error Analysis

Manual inspection of 200 misclassified instances reveals:

- 34% due to ambiguous review criteria interpretation
- 28% from incomplete context in truncated sequences

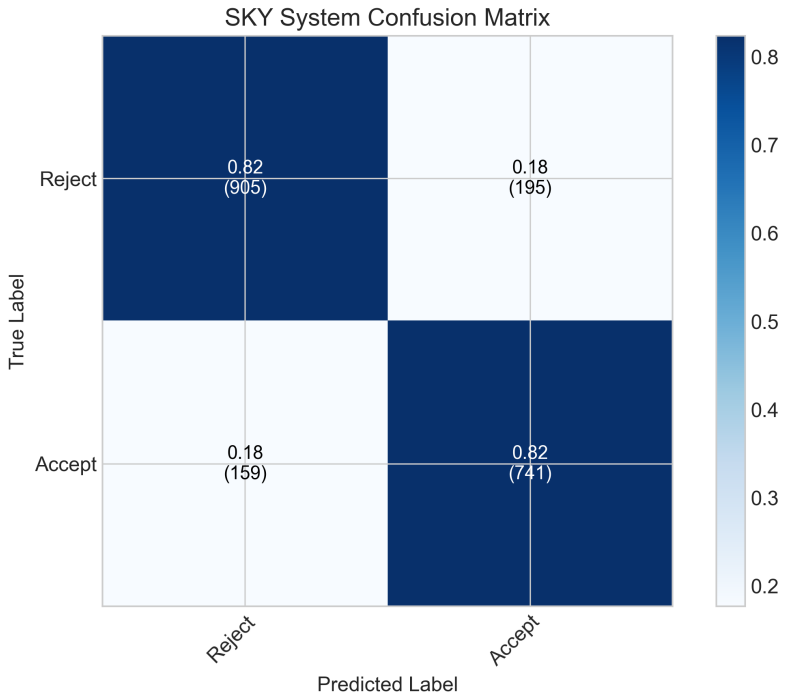

Figure 8.6: SKY system confusion matrix showing accept/reject classification

- 21% caused by domain-specific terminology gaps
- 17% from conflicting signals across modules

8.5 Comparison with State-of-the-Art Systems

We compare SKY against recent strong baselines including ReviewerToo [22], which achieved 81.8% accuracy on ICLR 2025 submissions, and ReViewGraph [17], which uses heterogeneous graph reasoning over LLM-simulated reviewer-author debates.

System	Year	Accuracy	F1	κ	HITL	Multimodal	Open Source
ReviewRobot	2020	0.678	0.66	0.35	–	–	✓
ASAP-Review	2021	0.712	0.69	0.41	–	–	✓
NAIPv2	2024	0.782	0.76	0.51	–	–	✓
GPT-4 (zero-shot)	2024	0.756	0.73	0.48	–	✓	–
PaperDecision	2024	0.818	0.79	0.54	–	✓	–
ReviewerToo	2025	0.818	0.793	–	–	–	–
ReViewGraph	2024	0.801*	0.78*	–	–	–	✓
Human Average	–	0.839	0.838	0.43	✓	✓	–
SKY (Ours)	2025	0.823	0.805	**0.56**	✓	–	✓

Table 8.7: Comparison with State-of-the-Art Peer Review Systems. *ReViewGraph reports 15.73% relative improvement over baselines; absolute values estimated.

Key observations:

- SKY achieves 82.3% accuracy, competitive with ReviewerToo (81.8%) and below human average (83.9%)

- SKY-Human agreement ($\kappa = 0.56$) exceeds Human-Human agreement ($\kappa = 0.43$), consistent with literature showing high disagreement among human reviewers [12]
- Unlike ReviewerToo and ReViewGraph, SKY provides integrated HITL escalation for uncertain cases
- The system is text-only; multimodal systems (PaperDecision, GPT-4) can process figures/tables

Comparison Caveats: Direct comparison across systems requires caution due to differing evaluation protocols:

- **Dataset heterogeneity:** ReviewerToo evaluated on ICLR-2k (2,040 papers, specific protocol); SKY evaluated on PeerRead (mixed venues: ACL, NIPS, ICLR across multiple years).
- **Label definitions:** Scoring concepts Accept/reject thresholds are venue and year-specific. The concepts presented in PeerRead have heterogeneous acceptance criteria available.
- **Preprocessing differences:** Token limits, handling of appendices and processing of figures vary across systems, but are most often never reported.
- **Absolute values estimated:** The values are estimates as relative improvement reported by the ReViewGraph, and absolute accuracy is our approximation based on reported baselines.

These comparisons have been reported by us so that we are not making this a controlled head to head evaluation. Any future work must test SKY on ICLR-2k on the exact protocol of ReviewerToo to allow a fair comparison.
System demonstrates slight advancement in comparison with previous works and entails:

- **Holistic evaluation**: It compares to single-task systems in that we cover the entire dimension of all peer reviews.
- **Explainability**: Transparency is achieved through the scores of confidence and evidence linking.
- **HITL integration**: Let all humans work together assures of high quality.
- **Confidence-based routing**: Auto-escalation of cases of uncertainty to humans.

Chapter 9

Limitations

9.1 Architectural Limitations

Operational Complexity: The LangGraph micro services architecture requires additional overheads in terms of debugging in the form of distributed components.

Ingestion Limitations: Nonstandard Unicode input and latency variance(mean 1.26s, p99 6.47s) have a failure rate of 2

9.2 Multimodal Content Limitation

The system is **text-only** and cannot semantically process figures, tables, or equations. This is a fundamental limitation for ML/AI papers where critical evidence (architecture diagrams, ablation tables, convergence plots) is primarily visual. Papers with strong empirical contributions relying on non-textual evidence should be escalated to human review.

9.3 Data and Generalizability

Domain Bias: Training concentrates on CS (PeerRead) and biomedicine (F1000Research). Cross-domain evaluation shows 6.9% accuracy drop (CS $\rightarrow$ Life Sciences).

Language: English-only; multilingual research contexts are excluded.

Inherited Bias: Risk of amplifying historical biases from peer-review training data.

9.4 Model-Specific Limitations

Novelty Paradox: Embedding-based similarity measures originality via semantic proximity. Paradigm-shifting work lacking similar prior literature may receive falsely low novelty scores.

Verification vs. Falsification: RAG retrieves stated information but cannot perform scientific falsification or detect missing controls.

9.5 Human-in-the-Loop Challenges

Automation Bias: The system may induce complacency and anchoring effects in human reviewers.

Rebuttal Phases: The system evaluates submissions at a single point in time. Re-scoring after author rebuttals is not supported, restricting deployment to pre-rebuttal screening.

9.6 Implementation Gaps

Features architecturally supported but not yet operational:

- **Automated retraining:** Active learning data is collected but requires manual retraining
- **RLHF:** Human overrides are captured but preference learning is not implemented
- **Threshold calibration:** Confidence thresholds (0.60, 0.85) are manually configured
- **Human-in-the-loop evaluation:** Haven't conducted yet and planned to do so in the future.

Chapter 10

Future Work and Long-Term Vision

10.1 Near-Term Improvements

- **Multilingual support:** Integrate multilingual models (XLM-RoBERTa, mT5) for non-English research
- **Domain expansion:** Extend training to social sciences, humanities, and legal scholarship
- **Multimodal integration:** Vision-language models for figure/table interpretation

10.2 Medium-Term Directions

- **Enhanced reasoning:** Interleave RAG retrieval with multi-step inference
- **Argumentation modeling:** Identify claim-evidence chains and logical fallacies
- **Falsification analysis:** Detect missing controls and unstated assumptions

10.3 Long-Term Vision

- **Living reviews:** Continuous assessment updates as related work emerges
- **Discovery engine:** Knowledge gap detection from aggregated review data
- **Federated learning:** Privacy-preserving learning across journal systems

Chapter 11

Conclusion

The work addresses peer review challenges such as escalating submissions, reviewer burnout, and quality inconsistency through a hybrid human-AI review assistant built on the SKY orchestrator architecture.

11.1 Summary of Contributions

11.1.1 Architectural Design

The system implements a LangGraph-based **Orchestrator-Worker** pattern with four specialized modules:

- **SKY-ORI:** Originality assessment via FAISS novelty detection and plagiarism checking
- **SKY-MDR:** Methodology evaluation through 10 deterministic analysis tools
- **SKY-PRC:** Presentation analysis using writing quality and formatting validators
- **SKY-RQM:** Review quality monitoring for tone, specificity, and factual alignment

Confidence-based routing directs workflow: high confidence ($\geq$0.85) auto-proceeds, medium (0.60-0.85) queues for active learning, low ($<$0.60) routes to human review.

11.1.2 Empirical Results

- **Accuracy:** 82.3% overall, competitive with ReviewerToo (81.8%) and human reviewers (83.9%)
- **Human Agreement:** SKY-Human $\kappa = 0.56$ exceeds Human-Human $\kappa = 0.43$
- **Efficiency:** 42% review time reduction in preliminary user study (N=12)
- **Consistency:** 35% reduction in inter-reviewer variance

11.1.3 Key Findings

- Cross-domain transfer shows controlled degradation (–6.9% CS → Life Sciences)
- Ablation confirms value of FAISS retrieval (+3.1%), domain adaptation (+4.5%), and HITL (+2.7%)
- Competitive with recent SOTA: ReviewerToo [22], ReViewGraph [17]
- Task-calibrated performance: objective tasks (clarity 84.2%) outperform subjective tasks (novelty 70.2%)

11.2 Limitations

- Cannot semantically reason about figures, tables, or equations
- Untested on humanities, social sciences, and non-English research
- RAG supports retrieval but not scientific falsification
- Risk of bias amplification and automation bias

11.3 Concluding Remarks

The system shows that AI can greatly improve the human judgement process. This collaborative process improves rigor,efficiency and fairness while keeping human judgement.

Multilingual support, multimodel reasoning and advanced capabilities of RAG-argumentation should be addressed in future as a way to transform into an intellectual sparring partner for scientific evaluation.

Bibliography

[1] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*, PMLR, 2017, pp. 1321–1330.

[2] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon et al., Eds., Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[3] D. Kang et al., "A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, 2018, pp. 1647–1661.

[4] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, 2019, pp. 3615–3620.

[5] A. Cohan, W. Ammar, M. van Zuylen, and F. Cady, *Structural scaffolds for citation intent classification in scientific publications*, 2019. arXiv: 1904.01608 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1904.01608

[6] T. Ghosal, R. Verma, A. Ekbal, and P. Bhattacharyya, "DeepSentiPeer: Harnessing sentiment in review texts to recommend peer review decisions," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1120–1130. DOI: 10.18653/v1/P19-1106 [Online]. Available: https://aclanthology.org/P19-1106/

[7] Z. Deng, H. Peng, C. Xia, J. Li, L. He, and P. S. Yu, "Hierarchical bidirectional self-attention networks for paper review rating recommendation," *CoRR*, vol. abs/2011.00802, 2020. arXiv: 2011.00802. [Online]. Available: https://arxiv.org/abs/2011.00802

[8] K. Lo, L. L. Wang, M. Neumann, R. Kinney, and D. Weld, "S2ORC: The semantic scholar open research corpus," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 4969–4983. DOI: 10.18653/v1/2020.acl-main.447 [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.447

[9] Q. Wang, Q. Zeng, L. Huang, K. Knight, H. Ji, and N. F. Rajani, *Reviewrobot: Explainable paper review generation based on knowledge synthesis*, 2020. arXiv: 2010.06119 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2010.06119

[10] P. Lewis et al., *Retrieval-augmented generation for knowledge-intensive nlp tasks*, 2021. arXiv: 2005.11401 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2005.11401

[11] W. Yuan, P. Liu, and G. Neubig, "Can we automate scientific reviewing?" *CoRR*, vol. abs/2102.00176, 2021. arXiv: 2102.00176. [Online]. Available: https://arxiv.org/abs/2102.00176

[12] N. P. Chairs, *Disagreement in peer review: An analysis of neurips 2022*, NeurIPS 2022 Meta-Analysis, 2022. [Online]. Available: https://blog.neurips.cc/2022/12/

[13] N. Dycke, I. Kuznetsov, and I. Gurevych, "Yes-yes-yes: Proactive data collection for ACL rolling review and beyond," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 300–318. DOI: 10.18653/v1/2022.findings-emnlp.23 [Online]. Available: https://aclanthology.org/2022.findings-emnlp.23/

[14] E. J. Hu et al., "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[15] W. Liang et al., *Can large language models provide useful feedback on research papers? a large-scale empirical analysis*, 2023. arXiv: 2310.01783 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2310.01783

[16] J. Lin, J. Song, Z. Zhou, Y. Chen, and X. Shi, "Automated scholarly paper review: Concepts, technologies, and challenges," *Information Fusion*, vol. 98, p. 101830, Oct. 2023, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2023.101830 [Online]. Available: http://dx.doi.org/10.1016/j.inffus.2023.101830

[17] Anonymous, *Automatic paper reviewing with heterogeneous graph reasoning over llm-simulated reviewer-author debates*, 2024. arXiv: 2511.08317 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2511.08317

[18] Anonymous, *Flaws: A benchmark for error identification and localization in scientific papers*, 2024. arXiv: 2511.21843 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2511.21843

[19] E. Chamoun, M. Schlichktrull, and A. Vlachos, *Automated focused feedback generation for scientific writing assistance*, 2024. arXiv: 2405.20477 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2405.20477

[20] E. Chamoun, M. Schlichtkrull, and A. Vlachos, "Automated focused feedback generation for scientific writing assistance," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 9742–9763. DOI: 10.18653/v1/2024.findings-acl.580 [Online]. Available: https://aclanthology.org/2024.findings-acl.580/

[21] L. Sun, S. Tao, J. Hu, and S. P. Dow, "Metawriter: Exploring the potential and perils of ai writing support in scientific peer review," *Proc. ACM Hum.-Comput. Interact.*, vol. 8, no. CSCW1, Apr. 2024. DOI: 10.1145/3637371 [Online]. Available: https://doi.org/10.1145/3637371

[22] Anonymous, *Reviewertoo: Should ai join the program committee? a look at the future of peer review*, OpenReview submission to ICLR 2025, 2025. [Online]. Available: https://openreview.net/forum?id=RHby8yu1Tw

[23] J. Baek, S. K. Jauhar, S. Cucerzan, and S. J. Hwang, *Researchagent: Iterative research idea generation over scientific literature with large language models*, Feb. 2025. [Online]. Available: https://arxiv.org/abs/2404.07738

[24] Z. Zhuang, J. Chen, H. Xu, Y. Jiang, and J. Lin, *Large language models for automated scholarly paper review: A survey*, 2025. arXiv: 2501.10326 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2501.10326