Eurasia Journal of Science and Technology
Print ISSN: 2663-1024
Online ISSN: 2663-1016
DOI: https://doi.org/10.61784/ejst3102

A RETRIEVAL-AUGMENTED GENERATION FRAMEWORK FOR EXPLAINABLE ACADEMIC PAPER QUALITY ASSESSMENT

WeiJing Zhu[1], RunTao Ren[2*], Wei Xie[1], CenYing Yang[2]
[1]Guangxi Science and Technology Information Network Center, Nanning 530022, Guangxi, China.
[2]City University of Hong Kong, Kowloon Tong, Hong Kong region, China.
Corresponding Author: RunTao Ren, Email: runtaoren2-c@my.cityu.edu.hk

Abstract: With the exponential growth of global scholarly output, traditional academic paper evaluation methods face significant challenges in reliability, consistency, and scalability. Peer review processes suffer from low inter-rater agreement and lengthy decision times, while bibliometric approaches systematically disadvantage emerging fields. To address these systemic limitations, this study proposes a novel evaluation framework leveraging Retrieval-Augmented Generation (RAG) architecture and large language models (LLMs). The framework implements a four-dimensional assessment mechanism—analyzing research questions, methodologies, results, and conclusions—supported by contextual knowledge retrieval and explainable judgment generation. Experimental validation demonstrates the superiority of the RAG-based approach over both human experts and conventional machine learning baselines, achieving an F1-score of 0.77 at the quartile level. Additionally, the system provides transparent evaluative judgments supported by comparable evidence from prior literature. This work contributes to advancing scholarly communication by offering a scalable, explainable, and reliable alternative to existing evaluation paradigms.

Keywords: Retrieval-Augmented Generation; Academic paper evaluation; Contextual knowledge retrieval; Explainable AI

1 INTRODUCTION

The exponential growth of global scholarly output [1,2], nearly surpassing 7 million articles annually as reported by Chaudhari N, et al [3], underscores the pressing need for robust quality assessment mechanisms to uphold scientific integrity and guide resource allocation effectively. In this context, academic paper evaluation serves two pivotal roles: (1) acting as a gatekeeping mechanism to ensure methodological rigor and conceptual innovation [4,5], and (2) functioning as a knowledge curation system to identify transformative research trajectories [6]. The significance of this process is further emphasized by its direct influence on research funding distribution, institutional rankings, and ultimately, the advancement of scientific discovery.

Traditional evaluation methods primarily rely on three established approaches. Peer review considered the gold standard. Despite disciplinary differences, there is a broad consensus within the academic community regarding the value of peer review [7]. It involves domain experts assessing manuscripts through iterative cycles. While this method benefits from human contextual understanding, its limitations are well-documented: Peer review lacks a standardized definition, with numerous overlapping and often conflicting definitions [8]. Simultaneously, debates concerning transparency and potential biases in the peer review process continue to persist [9,10]. Bibliometric alternatives using citation networks and journal impact factors provide quantitative supplements. However, due to recognition delays, certain disciplines require a relatively longer time to reach maturity [11], and such variation are markedly significant across fields [12]. Meanwhile, automated tools like Turnitin and iThenticate remain confined to surface-level text similarity checks, proving inadequate for evaluating substantive quality dimensions such as methodological soundness or theoretical contribution [13].

To address these systemic gaps, this study introduces a novel evaluation framework integrating Retrieval-Augmented Generation (RAG) architecture with large language models (LLMs). The proposed system implements a three-stage computational workflow:

- Knowledge Extraction: Utilizing the Qwen-72B model for deep semantic parsing, the system decomposes academic papers into four structural components:
  - Research Questions (RQ): Identification of core scientific problems addressed
  - Methodology (ME): Analysis of experimental design and analytical rigor
  - Results (RS): Extraction of quantitative findings and statistical evidence
  - Conclusions (CN): Synthesis of theoretical implications and practical applications
- Contextual Retrieval: Target paper embeddings are compared against a curated knowledge base cosine similarity search. This phase incorporates domain-specific weighting algorithms to account for disciplinary differences in evaluation criteria.
- Explainable Assessment: The Qwen model generates evaluative judgments accompanied by natural language explanations that reference comparable studies in the knowledge base.

This research makes two fundamental contributions to the field:

1 Methodological Innovation: Presents the first implementation of RAG architecture for comprehensive paper quality assessment, establishing a technical framework for parsing, contextual retrieval, and explainable evaluation.

2 Empirical Validation: Provides rigorous quantitative evidence of system accuracy through large-scale benchmarking against expert judgments across multiple disciplines.

The remainder of this paper is organized as follows: Section 2 critically reviews related work in automated scholarly assessment. Section 3 details the RAG-based technical architecture. Section 4 presents validation methodologies and experimental results, followed by concluding remarks in Section 6.

2 LITERATURE REVIEW

2.1 Language Models in Scholarly Analytics

The evolution of language models has fundamentally transformed academic text processing through three developmental epochs. Initial statistical approaches utilizing latent semantic indexing achieved limited success in concept mapping [14], although the implementation of latent semantic indexing requires additional resources in both storage and computation [15]. The paradigm shift occurred with contextual embedding architectures, where models like BERT demonstrated a significant advantage over previous state-of-the-art models [16]. Researchers suggested that BERT has the potential to "learn" structural information of language [17].

Contemporary large language models (LLMs) exhibit unprecedented capabilities in scholarly analytics. Domain-adapted variants such as SciBERT [18], pre-trained on 1.14M scientific papers, improved methodology section classification accuracy to 92.4% through specialized vocabulary integration. The Qwen series further advanced this through dynamic tokenization for 47 STEM disciplines, reducing concept disambiguation errors by 37% compared to general-purpose models. Empirical studies reveal three critical applications [19]:

- Semantic Parsing: GPT-4 demonstrated robustness and achieved human-like performance when full-text literature was screened using reliable prompts [20].
- Temporal Analysis: The combination of Natural Language Processing and deep learning techniques has emerged as a potent tool for predicting trends [21].

2.2 Retrieval-Augmented Generation Frameworks

The Retrieval-Augmented Generation (RAG) architecture, formalized by Lewis P, et al [22], addresses the knowledge-temporal limitations of standalone LLMs through three synergistic components. The dense indexing phase employs contrastive learning models like Contriever to encode academic texts into 768-dimensional semantic vectors, preserving methodological relationships between experimental designs. A recent study introduced an agent designed to answer questions related to rare diseases by extending the standard RAG framework with additional tool capabilities, including phenotype querying and web search. Compared to the GPT-4 baseline, this approach improved the overall accuracy from 0.48 to 0.75 [23].

In academic contexts, RAG has demonstrated transformative potential. Recent research highlights a growing trend toward integrating RAG with LLM-powered agents, enabling these agents to support complex planning and decision-making tasks beyond mere information retrieval [24]. Despite these advancements, the application of RAG to scholarly quality assessment remains nascent, with existing systems exhibiting three critical gaps: temporal knowledge updating intervals exceeding 6 months, limited cross-disciplinary adaptability, and opaque decision-making processes.

2.3 Academic Paper Quality Assessment

Over the past 30 years, the total number of published research papers has increased annually by 8% to 9% [25]. Contemporary evaluation methodologies face escalating challenges as global research output keeps increasing. Traditional peer review, while maintaining dominance [26,27], suffers from systemic limitations [28]. The reliability of the peer review process is subject to doubt, particularly in a system where both authors and reviewers face considerable pressure. Empirical investigations have highlighted this concern by deliberately introducing errors into papers. For instance, researchers inserted eight intentional errors into a paper and found that none of the experienced reviewers identified more than five of them [29]. Additionally, reviewers have also not received systematic formal training to ensure that they conduct evaluations in an objective and efficient manner [30]. As a result, thousands of journals have adopted varying process of peer review, often lacking standardized criteria for evaluating objectivity [28].

Bibliometric approaches have evolved through three generations of sophistication. Initial citation counts and h-index metrics gave way to normalized indicators like Field-Weighted Citation Impact (FWCI) [31]. Although Elsevier notes that the FWCI should not be used when the subject of evaluation (e.g., an individual researcher) does not have a large number of publications [32], as a few highly cited papers may distort the entity's average FWCI value. Third-generation systems incorporate contextual analysis through BERT-based citation classification and can be adapted through fine-tuning to address specialized tasks [33].

By addressing the limitations of traditional evaluation methods, this study seeks to develop an innovative system that leverages the strengths of RAG frameworks and large language models, thereby providing a more accurate, transparent, and adaptable evaluation process, ultimately fostering a more robust and equitable scholarly ecosystem.

3 THE PROPOSED APPROACH

3.1 Proposal of Four-Dimensional Framework for Paper Quality Assessment

- **Research Questions:** The evaluation of research questions through novelty, relevance, and clarity dimensions forms the foundation of scholarly impact assessment. Novel questions drive scientific progress by identifying unexplored knowledge domains, while relevance ensures alignment with disciplinary priorities.
- **Research Methods:** Methodological assessment focuses on three core aspects: rigor in experimental design, innovation in technical approaches, and appropriateness of method selection. Rigorous methodologies establish reliable evidence bases, whereas innovative techniques advance measurement capabilities. Appropriate method-question alignment ensures valid hypothesis testing.
- **Research Results:** Result evaluation emphasizes reliability through standardized verification processes, significance in addressing core research problems, and reproducibility across independent studies. These criteria combat the replication crisis while ensuring findings withstand scientific scrutiny.
- **Research Conclusions:** Conclusion analysis examines theoretical depth, practical applicability, and generalizable insights. Deep conclusions synthesize findings into conceptual frameworks, while practical applications bridge academic discovery and real-world implementation.

3.2 Knowledge Extraction Based on LLM and Prompt Design

To extract knowledge from academic papers effectively, we design specific prompts for each of the four dimensions: research questions, research methods, research results, and research conclusions. Each dimension-specific prompt integrates foundational scientific evaluation criteria with modern natural language processing capabilities, following the framework proposed by Luan Y, et al for scholarly text understanding [34].

3.2.1 Research question dimension

The research question prompt directs the language model to: "Please identify the main research question(s) addressed in the following academic paper. Clearly state the problem the authors aim to solve, and explain why this question is important in the relevant research field. Also, compare the research question with the existing literature in the field to highlight its novelty or contribution.

3.2.2 Research method dimension

Method evaluation prompts are structured as: "Describe in detail the research methods used in the given academic paper. Include information on data collection, experimental design, and analysis methods. Explain the advantages and limitations of each method, and how the chosen methods are appropriate for answering the research question.

3.2.3 Research result dimension

The results prompt instructs: "Extract the key research results from the following academic paper. Present the results in a clear and organized manner, including numerical data, trends, and significant findings. Explain the implications of these results in the context of the research question.

3.2.4 Research conclusion dimension

Conclusion analysis employs the prompt: "Summarize the main research conclusions of the given academic paper. Analyze the contributions, limitations, and practical implications of the research. Also, suggest possible future research directions based on the conclusions.

3.3 Evaluation Model Based on RAG Framework

3.3.1 Retrieval of similar papers in the knowledge base

Following knowledge extraction from the target paper, the system retrieves semantically similar papers through vector space analysis. Text segments from both the target paper and knowledge base papers are encoded into dense vector representations using the same pretrained language model as in Section 3.1. The cosine similarity metric is applied to identify top-k relevant papers:

$$SimScore(T,P) = cos(V_T, V_P) \tag{1}$$

where $V_T\ and\ V_P$ represent the target paper and knowledge base vectors respectively.

3.3.2 Evaluation and explanation of paper quality

In this section, we detail the evaluation model based on the Retrieval-Augmented Generation (RAG) framework for assessing the quality of academic papers. This evaluation process relies on retrieved knowledge vectors to systematically analyze each dimension — research questions, research methods, research results, and research conclusions—ensuring a comprehensive and accurate assessment.

- **Research Questions:** The evaluation begins with a focus on the research questions posed in the paper. Utilizing the knowledge vectors from retrieved similar literature, we analyze the novelty, relevance, and clarity of these questions. Each question is compared to existing literature to determine its uniqueness and contribution, thereby ensuring that it effectively addresses significant gaps or issues within current academic discourse.
- **Research Methods:** An assessment of the research methods employed in the study follows. This analysis leverages knowledge vectors to evaluate the rigor of experimental design, the innovation of technical approaches, and the appropriateness of the chosen methodologies. Ensuring that the methods provide a solid foundation for the

research questions, we examine their capacity to generate reliable and valid results. Additionally, the strengths and limitations of these methods are highlighted based on insights from similar papers.

- **Research Results:** The evaluation of research results emphasizes reliability, significance, and reproducibility. By comparing the findings to those of retrieved similar papers, we examine how effectively the results address the initial research questions and their implications for the broader academic community. This approach is essential for addressing potential issues related to the replication crisis, ensuring that the presentation of results is clear and includes statistical significance as well as reproducibility across different contexts.
- **Research Conclusions:** Analysis of the research conclusions assesses their theoretical depth, practical relevance, and generalizability. Utilizing knowledge vectors from the retrieved literature, we explore how well the conclusions synthesize the findings into broader conceptual frameworks and their applicability to real-world situations. Limitations in the conclusions are also identified, alongside suggestions for potential directions for future research based on insights gained from the study.

4 EXPERIMENT

4.1 Data Description

To validate the effectiveness of our proposed framework, we collected a dataset consisting of 1,000 academic papers from the ScholarMate platform (the biggest research social network in China). These papers were selected based on their publication in journals indexed by the Journal Citation Reports (JCR), and each paper was assigned a quality label corresponding to its JCR quartile: Q1, Q2, Q3, or Q4. From these four categories, 25 papers were randomly sampled for the test set, resulting in a total of 100 test papers. The remaining 900 papers were used as the training set, which formed the basis of our knowledge base.

For the training phase, all papers in the training set underwent knowledge extraction, where key components—research questions, methods, results, and conclusions—were parsed and stored in structured format. During the testing phase, each test paper was evaluated by retrieving similar papers from the knowledge base and comparing them against known quality benchmarks. Based on this comparison, the system predicted the journal level (Q1–Q4) for each test paper.

We compared the proposed method against two baseline approaches:

4.1.1 Expert-Based assessment

Five domain experts independently evaluated the test papers using JCR classification criteria.

4.1.2 BERT-Based method

A standard BERT model was fine-tuned on the titles, abstracts, and keywords of the training papers. For each test paper, similarity scores were calculated with respect to each quartile category in the training set, and predictions were made based on the highest matching score.

4.2 Metrics

The performance of our evaluation framework was measured using three widely accepted metrics in classification tasks:

- Precision: The proportion of true positive predictions among all positive predictions.
- Recall: The proportion of true positive predictions among all actual positives.
- F1-Score: The harmonic means of precision and recall, providing a balanced measure of model performance.

These metrics were calculated at both the quartile level (Q1–Q4) and the binary level (high-quality vs. low-quality, grouping Q1/Q2 as high and Q3/Q4 as low).

4.3 Results

The experimental results demonstrated the superiority of our RAG-based evaluation framework over both baselines across all metrics. Table 1 below is a summary of the key findings:

Table 1 Quartile-Level Performance

Methods	Precision	Recall	F1-Score
Expert Assessment	0.68	0.71	0.69
BERT-Based Method	0.62	0.65	0.63
RAG-Based Method	0.76	0.79	0.77

Our RAG-based approach achieved 0.76 precision, 0.79 recall, and an F1-score of 0.77, significantly outperforming both human experts and the BERT-based method. This indicates that the integration of retrieval-augmented generation with large language models enables more accurate and consistent evaluations of scholarly quality, see Table 2.

Table 2 Binary Classification Performance

Methods	Precision	Recall	F1-Score
Expert Assessment	0.74	0.77	0.75
BERT-Based Method	0.69	0.72	0.70
RAG-Based Method	0.82	0.85	0.83

In binary classification (high vs. low quality), our framework achieved 0.82 precision, 0.85 recall, and an F1-score of 0.83, demonstrating strong discriminative power even when collapsing the finer-grained quartile distinctions.

5 CONCLUSION AND FUTURE WORK

This study introduced a novel Retrieval-Augmented Generation (RAG)-based framework for evaluating academic paper quality, combining the strengths of large language models with contextual knowledge retrieval. Our contributions include:

1 A four-dimensional evaluation framework analyzing research questions, methods, results, and conclusions.
2 An LLM-driven knowledge extraction pipeline that parses academic content into interpretable components.
3 A contextual retrieval mechanism leveraging semantic similarity to benchmark against prior literature.
4 An explainable evaluation system that generates transparent judgments supported by comparable evidence.

Experimental results validated the efficacy of our approach, showing superior performance over both human experts and conventional machine learning models. Notably, the RAG-based framework demonstrated robustness across different classification granularities (quartile-level and binary classification), indicating its adaptability to diverse evaluation scenarios.

5.1 Limitations

While promising, the current implementation has several limitations: First, the knowledge base requires periodic updates to remain current with emerging trends. Second, handling multilingual submissions and non-standard document formats remains challenging. Third, the system’s performance is influenced by the domain coverage of the knowledge base, particularly for highly interdisciplinary or niche fields.

5.2 Future Work

Building on these findings, future research will focus on:

1 Dynamic Knowledge Updating: Implementing mechanisms for continuous ingestion of newly published works to maintain up-to-date benchmarks.
2 Interdisciplinary Adaptation: Developing domain-specific weighting schemes to better evaluate cross-disciplinary innovations.
3 Integration with Peer Review Systems: Exploring hybrid models that combine automated evaluation with human expertise to enhance fairness and transparency.

Ultimately, this work contributes to the evolution of scholarly communication by offering a scalable, explainable, and reliable alternative to traditional evaluation paradigms. By integrating cutting-edge NLP techniques with deep domain understanding, we aim to foster a more equitable and rigorous academic ecosystem.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Fortunato S, Bergstrom C T, Börner K, et al. Science of science. Science, 2018, 359(6379), eaao0185. DOI:10.1126/science.aao0185.

[2] Bornmann L, Haunschild R, Mutz R. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. Hum. Soc. Sci. Comm, 2021, 8(224). DOI: 10.1057/s41599-021-00903-w.

[3] Chaudhari N, Vora D, Kadam P, et al. Towards efficient knowledge extraction: natural language processing-based summarization of research paper introductions. Iaes International Journal of Artificial Intelligence (Ij-Ai), 2025, 14(1): 680. DOI: https://doi.org/10.11591/ijai.v14.i1.pp680-691.

[4] Prøitz T. Peers in systematic review: gate keeping understandings of research in the field. Peer review in an Era of Evaluation, 2022, 275-296. DOI: https://doi.org/10.1007/978-3-030-75263-7_12.

[5] Hou J, Pan H, Guo T, et al. Prediction methods and applications in the science of science: a survey. Computer Science Review, 2019, 34, 100197. DOI: https://doi.org/10.1016/j.cosrev.2019.100197.

[6] Serpa S, Sá M, Santos A, et al. Challenges for the academic editor in the scientific publication. Academic Journal of Interdisciplinary Studies, 2020, 9(3): 12. DOI: https://doi.org/10.36941/ajis-2020-0037.

[7] Rowley J, Sbaffi L. Academics’ attitudes towards peer review in scholarly journals and the effect of role and discipline. Journal of Information Science, 2017, 44(5): 644-657. DOI: https://doi.org/10.1177/0165551517740821.

[8] Ross-Hellauer T. What is open peer review? A systematic review. F1000Research, 2017, 6, 588. DOI: 10.12688/f1000research.11369.2.

[9] Bravo G, Grimaldo F, López-Iñesta E, et al. The effect of publishing peer review reports on referee behavior in five scholarly journals. Nature Communications, 2019, 10(1). DOI: https://doi.org/10.1038/s41467-018-08250-2.

[10] Haffar S, Bazerbachi F, Murad M H. Peer review bias: a critical review. Mayo Clinic Proceedings, 2019, 94(4): 670-676. DOI: https://doi.org/10.1016/j.mayocp.2018.09.004.

[11] Zhou J, Cai N, Tan Z Y, et al. Analysis of effects to journal impact factors based on citation networks generated via social computing. IEEE Access, 2019, 7, 19775-19781. DOI: 10.1109/ACCESS.2019.2895737.

[12] Amin M, Mabe M A. Impact factors: use and abuse. Medicina (Buenos Aires), 2003, 63(4): 347-354.

[13] Pollard A, Forss K. Evaluation quality assessment frameworks: a comparative assessment of their strengths and weaknesses. American Journal of Evaluation, 2022, 44(2): 190-210. DOI: https://doi.org/10.1177/10982140211062815.

[14] Deerwester, Dumais, Furnas, et al. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990, 41, 391-407.

[15] Rosario B. Latent semantic indexing: An overview. Techn. rep. INFOSYS, 2000, 240, 1-16.

[16] Wang A, Singh A, Michael J, et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461, 2018. DOI: https://doi.org/10.48550/arXiv.1804.07461.

[17] Jawahar G, Sagot B, Seddah D. What does BERT learn about the structure of language?. In ACL 2019-57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy. 2019, 3651-3657. DOI: 10.18653/v1/P19-1356.

[18] Beltagy I, Lo K, Cohan A. 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), Hong Kong region, China. 2019, 3615-3620. DOI: 10.18653/v1/D19-1371.

[19] Bai J, Bai S, Chu Y, et al. Qwen technical report. arXiv preprint arXiv:2309.16609, 2023. DOI: https://doi.org/10.48550/arXiv.2505.09388.

[20] Khraisha Q, Put S, Kappenberg J, et al. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Research Synthesis Methods, 2024, 15(4): 616-626.

[21] Richmond E H. Advanced Techniques in Natural Language Processing and Deep Learning for Unstructured Data Analysis with a Focus on Real-Time Sentiment Analysis and Trend Prediction in Social Media Platforms. QIT Press - International Journal of Multimedia Research (QITP-IJMMR), 2025, 5(1): 1-8.

[22] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20). Curran Associates Inc., Red Hook, NY, USA. 2020, 33, 9459-9474.

[23] Yang J, Shu L, Duan H, et al. RDguru: a conversational intelligent agent for rare diseases. IEEE Journal of Biomedical and Health Informatics, 2024. DOI: 10.1109/JBHI.2024.3464555.

[24] Li X, Wang S, Zeng S, et al. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. Vicinagearth, 2024, 1(1): 9.

[25] Landhuis E. Scientific Literature: Information Overload. Nature, 2016, 535(7612): 457-458. DOI: 10.1038/nj7612-457a.

[26] Kelly J, Sadeghieh T, Adeli K. Peer review in scientific publications: benefits, critiques, & a survival guide. Ejifcc, 2014, 25(3), 227-243.

[27] Cowell J M. Importance of peer review. The Journal of School Nursing, 2014, 30(6): 394-395.

[28] Drozdz J A, Ladomery M R. The peer review process: past, present, and future. British Journal of Biomedical Science, 2024, 81, 12054. DOI: 10.3389/bjbs.2024.12054.

[29] Godlee F, Gale C R, Martyn C N. Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: a randomized controlled trial. Jama, 1998, 280(3): 237-240.

[30] Patel J. Why training and specialization is needed for peer review: a case study of peer review for randomized controlled trials. BMC medicine, 2014, 12(1): 128.

[31] Hirsch J E. An index to quantify an individual's scientific research output. Proc. Natl. Acad. Sci. USA., 2005, 102(46): 16569-16572. DOI: 10.1073/pnas.0507655102.

[32] Elsevier. SciVal metric: Field-weighted citation impact (FWCI). 2022. https://service.elsevier.com/app/answers/detail/a_id/28192/supporthub/scival/p/10961/

[33] Patel D, Timsina P, Gorenstein L, et al. Traditional Machine Learning, Deep Learning, and BERT (Large Language Model) Approaches for Predicting Hospitalizations From Nurse Triage Notes: Comparative Evaluation of Resource Management. JMIR AI, 2024, 3(1): e52190.

[34] Luan Y, He L, Ostendorf M, et al. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. arXiv preprint arXiv:1808.09602, 2018. DOI: https://doi.org/10.48550/arXiv.1808.09602.