GraphMind: Interactive Novelty Assessment System for Accelerating Scientific Discovery

Italo Luis da Silva, Hanqi Yan, Lin Gui, Yulan He
Department of Informatics
King's College London, UK
{italo.da_silva,hanqi.yan,lin.1.gui,yulan.he}@kcl.ac.uk

Abstract

Large Language Models (LLMs) show strong reasoning and text generation capabilities, prompting their use in scientific literature analysis, including novelty assessment. While evaluating novelty of scientific papers is crucial for peer review, it requires extensive knowledge of related work, something not all reviewers have. While recent work on LLM-assisted scientific literature analysis supports literature comparison, existing approaches offer limited transparency and lack mechanisms for result traceability via an information retrieval module. To address this gap, we introduce **GraphMind**, an easy-to-use interactive web tool designed to assist users in evaluating the novelty of scientific papers or drafted ideas. Specially, **GraphMind** enables users to capture the main structure of a scientific paper, explore related ideas through various perspectives, and assess novelty via providing verifiable contextual insights. **GraphMind** enables users to annotate related papers through various relationships, and assess novelty with contextual insight. This tool integrates Semantic Scholar with LLMs to support annotation, classification of papers. This combination provides users with a rich, structured view of a scientific idea's core contributions and its connections to existing work. **GraphMind** is available at https://oyarsa.github.io/graphmind and a demonstration video at https://youtu.be/wKbjQpSvwJg. The source code is available at https://github.com/oyarsa/graphmind.

1 Introduction

Peer reviewing of scientific papers is a challenging task essential for scientific collaboration. It requires a thorough understanding of the paper being evaluated, as well as knowledge of the scientific literature around the topic. However, with the increasing number of publications, ranging from peer-reviewed conference and journal articles to preprints, it is increasingly difficult for reviewers to stay up to date within their research domains. Concurrently, advances in LLMs' reasoning and text generation capabilities have spurred interest in their use for scientific literature review (Yuan et al., 2021; Lin et al., 2023; Chitale et al., 2025).

There are two key dimensions to consider when evaluating a research paper. 1) at the **macro level**, how does the paper relate to existing work? Has similar research already been published? Compared to the cited or relevant prior work, does this paper present a groundbreaking idea or merely an incremental contribution? 2) at the **micro level**, is the paper well-organized and clearly motivated? How is the structure laid out, and do the sections logically support one another? Is there coherence across chapters, and does each part contribute meaningfully to the overall argument? 3) Building on these two dimensions, a third consideration is the **interaction between macro and micro levels**: can each micro-level component (e.g., specific methods, results, or arguments) be supported or contextualized by the macro-level literature? In other words, does the paper consistently draw connections between its internal structure and the broader research landscape? 4) Finally, based on all of the above, the ultimate question is: how can we synthesize these observations into an evaluation metric that effectively measures the paper's overall contribution?

However, existing review tools often fall short in capturing all the key features necessary for comprehensive paper evaluation and remain limited in scope. Some focus primarily on retrieving potentially related papers via citation networks, such as Connected Papers[1] and Inciteful[2], while others rely on surface-level semantic search, such as keyword or title matching, as seen in tools like Litmaps[3]. Additional systems attempt to assess paper quality

[1] https://connectedpapers.com/
[2] https://inciteful.xyz/
[3] https://litmaps.com/

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 286–294
November 4-9, 2025

based on content alone (Lin et al., 2024; Kang et al., 2018). In general, these tools typically lack integration between the internal content of a paper and its broader research context, which is an essential factor for evaluating scientific novelty.

To address these limitations, we propose **GraphMind**, an interactive tool designed to support novelty assessment by analyzing both macro-level and micro-level information, as well as the interaction between them. **GraphMind** enables a systematic and comprehensive analysis of a paper's contribution. In practice, it allows users to identify key research components and explore related work through a combination of citation-based and semantic relationships. Unlike traditional methods, our system decomposes abstracts into distinct background and target components, which enhances its ability to retrieve semantically related papers. This deeper contextual understanding facilitates a more robust and informed evaluation of novelty.

To support this functionality, we leverage **a combination of existing tools to collect macro-level information**. For instance, using the arXiv API, we can search for papers and retrieve their full LaTeX content. Additionally, the Semantic Scholar API[4] allows us to access metadata, citation information, and recommended related papers. At the **micro level, we utilize LLMs**, such as GPT-4o[5] and Gemini 2.0 Flash[6], to extract key elements from each paper, including claims, methodologies, experiments, background, and research objectives. Using this information, we construct a hierarchical graph of related papers, integrating both top-cited works and semantically similar papers, to support novelty assessments. This graph forms the basis for generating reports that highlight both supporting and contrasting evidence from the literature.

To ensure flexibility and usability, GraphMind features a web-based frontend that allows users to explore and select papers for evaluation, and a backend server that handles API queries and vector-based similarity search. Users can operate in three modes: (1) browsing a curated set of papers with pre-computed analyses, (2) dynamically evaluating new papers from arXiv using live data retrieval and analysis, or (3) evaluating draft papers by directly entering the title and abstract:

- We introduce **GraphMind**, a tool designed to assist paper reviews by generating structured reports that combine a paper's key elements with insights from related works.
- We leverage APIs from arXiv and Semantic Scholar, with LLM-based extraction, to identify the key aspects of scientific papers.
- Our tool allows the creation novelty assessment reports that integrate detailed analysis of a paper's contribution and its position within the surrounding research landscape.

2 Related Work

There has been substantial progress in both macro and micro-level for LLM-assisted relevant paper recommendations and automated peer review.

Macro level: Automated paper recommendation. Existing retrieval approaches focus on citation-based connections rather than content-level analysis (Uzzi et al., 2013; Yang et al., 2025; Kreutz and Schenkel, 2022). Tools like Connected Papers, Inciteful, Litmaps, and ResearchRabbit[7] build paper graphs using co-citation, bibliographic coupling, or title/keyword similarity, but often miss deeper semantic relationships and lack transparency in their algorithms. SPECTER (Singh et al., 2022) improves relatedness scoring via citation-informed embeddings, but does not extract or leverage specific content from papers. Scholar Inbox (Flicke et al., 2025) considered both citation and semantic information, but only document level without fine-grained analysis. Guo et al. (2020) use title-abstract attention relations for retrieval, but do not identify detailed semantic relationships between components.

Micro level: LLM-assisted scientific paper review. Other works focus on evaluating and reviewing papers exclusively from their own content. PeerRead (Kang et al., 2018) includes a small subset with expert-annotated aspects such as clarity, impact, and originality. SciND (Gupta et al., 2024) constructs a knowledge graph from extracted novel entity triplets in publications to support novelty assessment, but it does not provide direct novelty annotations. SchNovel (Lin et al., 2024) extracted abstracts and metadata (e.g., institution, publication year) from 150,000 papers in the arXiv dataset[8]. However, it infers novelty through publication date, assuming newer work is more novel, a simplistic proxy that overlooks semantic content. Ai et al.

[4] https://www.semanticscholar.org/product/api
[5] https://openai.com/index/gpt-4o-system-card/
[6] https://deepmind.google/technologies/gemini/
[7] https://researchrabbitapp.com
[8] https://www.kaggle.com/datasets/Cornell-University/arxiv

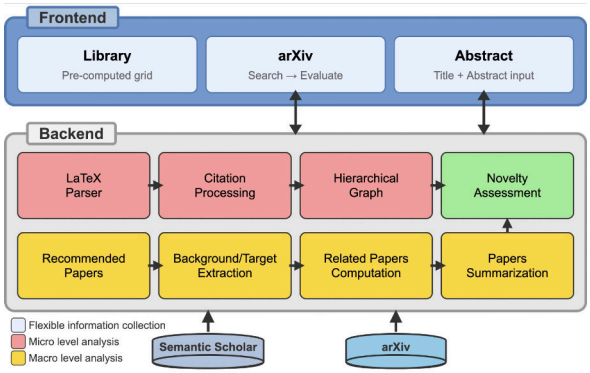

Figure 1: The overall architecture of **GraphMind**.

(2024) determines the novelty of a document by comparing its atomic content units (ACUs). It retrieves similar ACUs by cosine similarity and calculates the novelty depending on how salient the ACUs are in the corpus. While algorithmically interesting, this approach fails to capture scientific innovation and creativity.

In summary, existing tools either lack macro or micro-level understanding, barely discuss the information interaction, such as supporting evidence between two levels. Therefore, GraphMind aims to fill in the gap by combining structured paper understanding with relationship-aware analysis, to further support the novelty assessment.

3 Architecture of GraphMind

GraphMind is composed of a web-based frontend and a backend server. Together, they enable users to search, analyze, and evaluate the novelty of scientific papers. An overview of the system is presented in Figure 1. The **Frontend** is designed to support a more flexible information collection paradigm, while the **Backend** focuses on enabling a comprehensive analysis of the given paper by considering both macro- and micro-level information.

Frontend The web interface is built using vanilla TypeScript as a Multi-Page Application, with separate files for each page. It communicates with the server via a REST API. There are two main pages: *Search* and *Detail*. In the *Search* page, the user can select papers from either a pre-computed *Library* or perform a dynamic *arXiv* search. The *Library* contains a selected subset of papers from the ICLR 2022-2025 and NeurIPS 2022-2024 conferences, that have been fully pre-processed. This enables quick access without relying on external APIs. The *arXiv* search allows users to query any arXiv paper. If data is available, the system runs the full evaluation pipeline. A progress bar is shown during processing, and users can enable notifications upon completion. Results are cached locally to avoid repeated processing. Regardless of how the user chose the paper, they eventually arrive at the *Detail* page, where they can see key metadata for the paper, extracted information, related works, and the final structured novelty report.

Backend The server is written in Python with FastAPI. It presents a REST API with three endpoints: `/search` (*Search*), `/evaluate` (*Evaluate*) and `/abstract` (*Abstract*). The *Search* endpoint uses the arXiv API to query the papers by title, and returns the basic meta data as JSON. The *Evaluate* endpoint gets a paper title, arXiv ID, and some settings such as the number of related papers to retrieve and the LLM choice. It then uses the arXiv API to retrieve the paper contents, the Semantic Scholar API to retrieve full metadata from the paper and its citations, and recommended papers to use as the base for the semantic similarity method. It uses Server-Side Events (SSE) to to stream live updates to the frontend during evaluation. The full evaluation result, including the main paper, related work, extracted data and final report, is returned as JSON. It supports GPT-4o, GPT-4o mini and Gemini 2.0 Flash as the LLMs for data extraction and evaluation, and uses SentenceTransformers (Reimers and Gurevych, 2019) to generate vector embeddings from the text[9]. The *Abstract* endpoint operates similarly but accepts only the title and abstract as input. From these, it retrieves semantically related papers and generates the novelty assessment. However, without access to the full paper content, it cannot extract citations or construct the structured graph.

4 GraphMind

GraphMind is a multi-source information display platform designed to help users analyze scientific ideas by comparing them with relevant literature. It presents a structured view of processed papers and supports novelty assessment through interactive statistical insights. By integrating evidence extracted from both the target paper and related works, the platform enables flexible retrieval and visualization of relevant multi-source literature. In what follows, we first describe the search interface (Section 4.1), followed by the assessment results page (Section 4.2). More details of the implementation can be found in Appendix C.

[9]The encoder model used is `all-MiniLM-L6-v2`.

4.1 Search

When the tool is launched, the user is directed to the *Search* page. On their first visit, a help message is displayed, offering instructions on how to use the tool. After dismissing this message, the user has three options to explore: the *Library*, which contains a collection of pre-computed papers, the *arXiv* search, and the *Abstract* evaluation. Figure 2 shows the *Search page* interface.

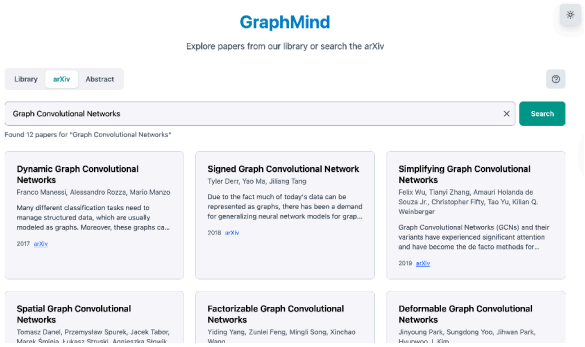

Figure 2: Search page showing the available tabs with the arXiv tab selected.

Library This page shows a curated list of pre-computed papers available for immediate exploration. Each paper is presented as a card containing the title, abstract, and publication venue. Users can click on a card to go to the *Detail* page to see the novelty assessment report.

arXiv This page allows users to search for papers on arXiv by title, using live data retrieval and dynamic analysis. Results are presented similarly to those in the *Library* tab. Upon selecting a paper, users are shown a configuration panel (see Figure 3) where they can customize evaluation settings, such as the number of citations to include, recommended papers to retrieve, related papers with which to build the graph, and the LLM to use to extract the required information. They can also choose whether to include all related papers or only those published before the main paper. Once confirmed, the assessment process starts (See Appendix C). During this process, a progress bar indicates the current steps being executed. Users can cancel at any time. When completed, they are redirected to the *Detail* page, which displays both micro- and macro-level novelty assessment results.

Abstract This tab allows users to directly enter a paper's title and abstract. This enables the evaluation of papers not in the arXiv. The system extracts relevant information from the abstract and performs a macro-level novelty evaluation. Note that micro-level analysis is not possible due to lack of full arXiv metadata.

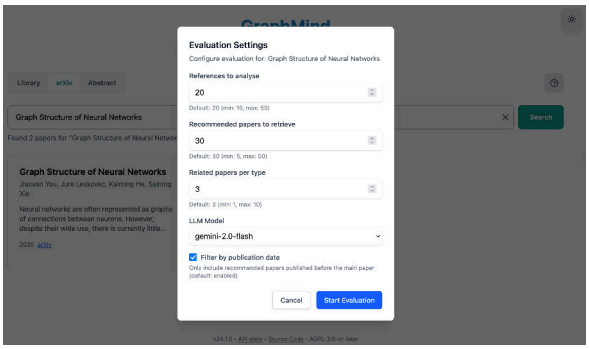

Figure 3: Customizable evaluation configuration panel.

4.2 Assessment Results

The **Detail** page provides comprehensive paper analysis and novelty assessment results, including **Metadata**, with basic paper information and predicted novelty label, **Novelty Assessment**, with synthesized evaluation results and its supporting evidence, **Paper Structured Graph** with micro-level elements extracted from the paper, and **Related Papers**, with macro-level relevant papers retrieved using both the citation network and semantic matches.

Metadata Shows essential paper information: title, authors, publication year, conference name, acceptance status (if available), arXiv link, extracted keywords, and abstract. A novelty score, expressed as a percentage, is also provided. This score is obtained by prompting the evaluation model for novelty assessment multiple times, and averaging the predictions. Figure 4 shows the metadata section.

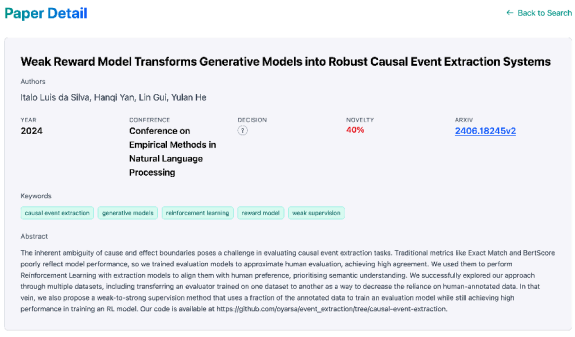

Figure 4: Metadata section on the Paper Detail page.

Paper Structured Graph Presents the micro-level elements of the paper extracted from the full

paper content. It included the core claims, the methods used to validate those claims, and the experiments conducted as evidence for the methods. These elements are interlinked, illustrating how claims are substantiated by methods, and how those methods are, in turn, validated through experiments. This structure helps users understand the key aspects of the paper, how they support each other and how they contribute to the overall argument. Users can interact with each node in the graph to view the corresponding segments extracted directly from the paper. Figure 5 shows the structured graph.

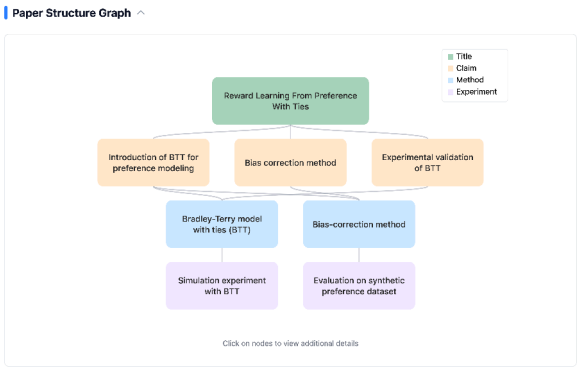

Figure 5: Paper structured graph.

Novelty Assessment Presents the results of novelty assessment using our proposed method. It begins with a *Result* summary that integrates the key findings from both micro- and macro-level analyses, explaining how they collectively inform the final novelty score. Following the summary, two sections provides deeper context: (1) **Supporting Evidence**: Highlights external papers that reinforce the novelty and soundness of the main paper's approach. These examples show how the paper introduces new ideas or methods that are distinct from prior work. (2) **Contradictory Evidence**: Identifies related papers that challenge the originality or effectiveness of the proposed approach, pointing out overlaps or limitations. Each evidence item links to a related paper (see below) and highlights how that paper contributes to novelty assessment. Figure 6 shows the Novelty Assessment section with *Supporting Evidence*, highlighting a comparison of background information between the main and a related paper.

Related papers The *Related Papers* section displays the extracted information for each retrieved related paper. This includes both the citations, categorized as either *supporting* or *contrasting* based on the context in which they are cited, and the

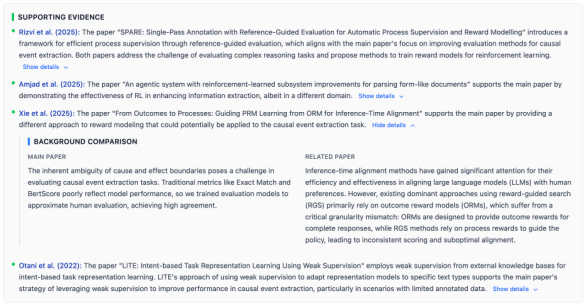

Figure 6: Novelty Assessment section showing supporting evidence comparison.

semantically related papers, which are classified as either *background* (works that inform or motivate the main paper) or *target* (works that address similar goals). For each related paper, we show a semantic similarity score (both as a percentage and a visual scale), the original abstract, and a relation-aware summary explaining the connection to the main paper. For cited papers, we also include the citation contexts with their corresponding polarities (supporting or contrasting). For semantically related papers, we show either the extracted background content (for background papers) or the target content (for target papers).

Abstract evaluation For papers evaluated through the *Abstract* input method, some features are unavailable due to the absence of full arXiv LaTeX content. Specifically, we cannot generate the **Paper Structured Graph** or include citation-based entries in the **Related Papers** section. However, we do provide semantically related papers found based solely on the abstract content, and derive the novelty assessment from those relationships. This allows authors to position their in-progress work within the relevant literature, even in the absence of a full paper submission.

5 Model Evaluation

To evaluate the performance of our methodology, we conduct experiments using published papers from ICLR (2022-2025) and NeurIPS (2022-2024) conferences. The papers were collected via the OpenReview API[10], and our full assessment pipeline was applied to them. We use the median originality scores from peer reviews as our ground truth. These scores range from 1 to 5 (see Appendix E for the complete scoring rubric), where we treat scores of 1-3 as "not novel" and 4-5 as "novel". The predicted novelty labels are compared against the

[10] https://docs.openreview.net/

ground truth using standard classification metrics: precision, recall, F1 score, and accuracy. Our evaluation includes several LLMs: GPT-4o, Gemini 2.0 Flash, Qwen 2.5 7B (Yang et al., 2024) and Llama 3.1 8B (Dubey et al., 2024).

Table 1 shows the performance of our method against baselines. The *Basic* baselines directly prompt a given LLM using only the paper title and abstract, without our hierarchical graph or related papers. The *Search* baselines are provided with our hierarchical graph. But they use the models' own web search capabilities to find relevant papers, not our curated retrieval pipeline. Note that both Qwen and Llama do not support search capabilities; therefore, we report results only for their Basic versions.

Model	Precision	Recall	F1	Accuracy
$\text{Basic}_{\text{GPT-4o}}$	0.6863	**0.7000**	0.6931	0.6900
$\text{Search}_{\text{GPT-4o}}$	0.6667	0.6400	0.6531	0.6600
$\text{GraphMind}_{\text{GPT-4o}}$	**0.7805**	0.6400	**0.7033**	**0.7300**
$\text{Basic}_{\text{Gemini}}$	0.5169	**0.9200**	0.6619	0.5300
$\text{Search}_{\text{Gemini}}$	0.6667	0.7200	0.6923	0.6800
$\text{GraphMind}_{\text{Gemini}}$	**0.7800**	0.7222	**0.7500**	**0.7400**
$\text{Basic}_{\text{Qwen}}$	**0.6680**	0.7371	0.7008	0.5500
$\text{GraphMind}_{\text{Qwen}}$	0.5946	**0.8800**	**0.7097**	**0.5800**
$\text{Basic}_{\text{Llama}}$	0.5062	**0.8200**	**0.6260**	0.5100
$\text{GraphMind}_{\text{Llama}}$	**0.5125**	0.8000	0.6247	**0.5200**

Table 1: Evaluation result with different LLMs.

We find that the *Basic* baseline struggles to accurately assess novelty, as it lacks sufficient context and relies solely on the LLM's internal knowledge. The *Search* baseline retrieves papers less relevant than our related paper retrieval method, which leads to lower performance with noisy data. Our method, *GraphMind*, consistently outperforms both baselines by leveraging a structured representation of the paper and high-quality retrieved related works.

In addition to the baselines, we also select a few variations on our method to show how each component contributes to the performance. Table 2 presents the results on our benchmark dataset `SciNova` and PeerRead.

Variants	Precision	Recall	F1	Accuracy
`GraphMind`	0.7800	**0.7222**	**0.7500**	**0.7400**
No citation	0.7506	0.6704	0.7083	0.7200
No semantic	**0.8167**	0.6772	0.7403	0.7200
No related	0.8151	0.6703	0.7353	0.6900
No graph	0.7217	0.6693	0.6942	0.6900

Table 2: Ablation results with updated runs.

We also evaluate the generated rationales using a Bradley-Terry tournament model (Bradley and Terry, 1952; Chiang et al., 2023). We employ an LLM (GPT-4o) as a judge to perform pairwise comparisons between the *Basic* baseline, our *GraphMind* method, and the original human reviews. The comparisons are scored across the following five dimensions:

Different aspects of the generated rationales:

Clarity: how easy is it to understand and to follow its ideas?

Faithfulness: does the rationale justify the novelty label? For example, if the text is mostly positive, so should the label.

Factuality: is the rationale is correct grounded in scientific facts from the target and related papers?

Specificity: does the rationale cover information specific to the paper, or doe sit make overly generic statements?

Contributions: does the rationale effectively compare the target paper with the related papers?

Table 3 displays the results, showing that GraphMind outperfoms the baseline in general, and closely matches or exceeds human reviews in several aspects, particularly in *Faithfulness*, *Factuality*, and *Specificity*.

Model	Clarity	Faithful	Factuality	Specificity	Contrib.
Human	**1547**	1476	1470	1443	**1584**
Basic	1520	1507	1386	1369	1430
GraphMind	1520	**1552**	**1609**	**1657**	1540

Table 3: Bradley-Terry ratings from automated pairwise tournament with GPT-4o as a judge.

The experimental results indicate that *GraphMind* is an effective system for assessing novelty of scientific papers. It produces more accurate novelty labels compared to directly prompting LLMs, and generates rationales that are on par with, or even superior to, human-written reviews in terms of faithfulness, factual grounding, and specificity.

6 Conclusion and Future Work

GraphMind is an easy-to-use interactive web tool where users can generate evaluation reports from scientific papers. It aims to assist peer reviewers and other academic users in the task of assessing the novelty of a paper using both its key elements and relationship with the scientific literature. We achieve this by fetching the paper's full content from the arXiv, using it to extract the key elements,

and pairing the paper with relevant papers from the literature extracted from a related papers graph.

In the future, we will consider further expanding our related paper retrieval to larger datasets and developing more refined ways of finding relevant papers. This includes building our own database containing millions of scientific papers and using more advanced LLM-driven methods to enhance the search process. We'll also consider incorporating interactive user feedback and evaluation on domains beyond Machine Learning.

Acknowledgments

This work was supported in part by the UK Engineering and Physical Sciences Research Council through a Turing AI Fellowship (grant no. EP/V020579/1, EP/V020579/2).

References

Lin Ai, Ziwei Gong, Harshsaiprasad Deshpande, Alexander Johnson, Emmy Phung, Ahmad Emami, and Julia Hirschberg. 2024. Novascore: A new automated metric for evaluating document level novelty. *Preprint*, arXiv:2409.09249.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Wei-Lin Chiang, Tianle Li, Joseph E. Gonzalez, and Ion Stoica. 2023. Chatbot arena - new models & elo system update. https://blog.lmarena.ai/blog/2023/leaderboard-elo-update/. Accessed: 2025-07-05.

Maitreya Prafulla Chitale, Ketaki Mangesh Shetye, Harshit Gupta, Manav Chaudhary, and Vasudeva Varma. 2025. Autorev: Automatic peer review system for academic research papers. *arXiv preprint arXiv:2505.14376*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Markus Flicke, Glenn Angrabeit, Madhav Iyengar, Vitalii Protsenko, Illia Shakun, Jovan Cicvaric, Bora Kargi, Haoyu He, Lukas Schuler, Lewin Scholz, Kavyanjali Agnihotri, Yong Cao, and Andreas Geiger. 2025. Scholar inbox: Personalized paper recommendations for scientists. *Preprint*, arXiv:2504.08385.

Guibing Guo, Bowei Chen, Xiaoyan Zhang, Zhirong Liu, Zhenhua Dong, and Xiuqiang He. 2020. Leveraging title-abstract attentive semantics for paper recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):67–74.

Komal Gupta, Ammaar Ahmad, Tirthankar Ghosal, and Asif Ekbal. 2024. Scind: a new triplet-based dataset for scientific novelty detection via knowledge graphs. *International Journal on Digital Libraries*, 25:639–659.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. *Preprint*, arXiv:1804.09635.

Christin Katharina Kreutz and Ralf Schenkel. 2022. Scientific paper recommendation systems: a literature review of recent publications. *Preprint*, arXiv:2201.00682.

Ethan Lin, Zhiyuan Peng, and Yi Fang. 2024. Evaluating and enhancing large language models for novelty assessment in scholarly publications. *Preprint*, arXiv:2409.16605.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023. Automated scholarly paper review: Concepts, technologies, and challenges. *Information Fusion*, 98:101830.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. Scirepeval: A multi-format benchmark for scientific document representations. In *Conference on Empirical Methods in Natural Language Processing*.

Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. Atypical combinations and scientific impact. *Science*, 342(6157):468–472.

Alex J Yang, Fanming Wang, Yujie Shi, Yiqin Zhang, Hao Wang, and Sanhong Deng. 2025. Beyond surface correlations: Reference behavior mediates the disruptiveness-citation relationship. *Journal of Data and Information Science*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021. Can we automate scientific reviewing? *Preprint*, arXiv:2102.00176.

Appendix

A Evaluation dataset

Table A1 shows the distribution of novelty labels in our dataset.

Year	Count	Count %	Novel	Novel %
2022	534	17.4%	450	84.3%
2023	688	22.5%	555	80.7%
2024	929	30.3%	549	59.1%
2025	912	29.8%	456	50.0%
Total	3063	100.0%	2010	65.6%

Table A1: Distribution of scientific papers by year with novelty rates.

B More screenshots

Figures A1 and A2 show the *Paper Analysis* and *Related Papers* sections, respectively.

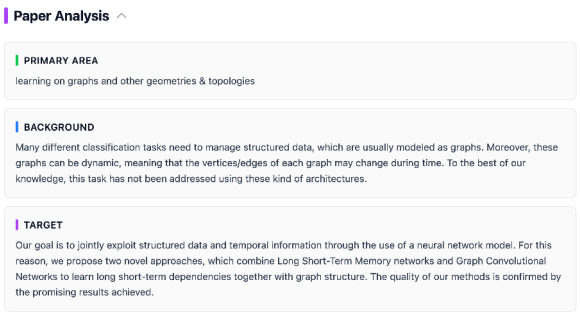

Figure A1: Paper Analysis section

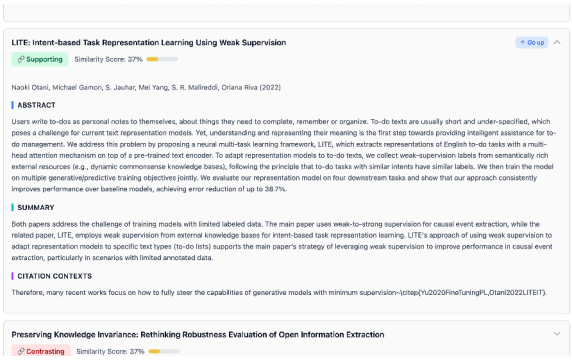

Figure A2: Related Papers section.

C Assessment Pipeline

Sections 4.1 and 4.2 show how our tool looks like from the user's point of view. This section describes the pipeline that generates that information. This includes three components: **Graph extraction**, **Related papers retrieval** and the final **Novelty assessment generation**. This section describes the steps required for each one.

Graph extraction We extract the graph from the full paper content from the arXiv. First, we take the LaTeX code, extract citations from the bibliography and convert the content to Markdown using Pandoc[11]. For each citation, we identify all sentences in the main text where it appears as citation contexts. An LLM then extracts key components from the Markdown: claims, methods, experiments, and their interconnections. For each component, we also extract supporting excerpts from the paper.

Related Papers We retrieve two types of papers: citations and semantically related. Citations are extracted from the parsed bibliography and filtered by semantic similarity to retain only the most relevant ones. We then use an LLM to classify each citation context as positive or negative. Citations are classified as supporting when contexts are mostly positive, and as contrasting when contexts are mostly negative.

To find semantic neighbours, we first retrieve recommended papers via the Semantic Scholar API. We then use an LLM to extract targets and backgrounds from their abstracts, calculate semantic similarity with the main paper's targets and backgrounds, and select the most relevant matches.

Novelty Assessment We combine the paper graph and related papers to generate the final novelty assessment. First, we convert the paper graph to text using topological sorting to create a linear chain of nodes, then transform each node into a paragraph using its label and supporting text. Second, we compile the related papers into an evidence list, converting each paper into a paragraph containing its title, relation type, and summary. Finally, we provide both components as input to an evaluation LLM, which generates a novelty label and structured rationale with result summary, supporting evidence, and contradictory evidence

D Cost and Time

Table A2 shows the average time and cost of performing the full evaluation of an arXiv paper for each model available in the tool. Note that the cheapest and fastest model (Gemini 2.0 Flash) is also the best performing (see 1.

[11] https://pandoc.org/

Model	Time (s)	Cost (USD)
Gemini 2.0 Flash	61.91	0.023213
GPT-4o	75.06	0.477835
GPT-4o mini	86.07	0.030429

Table A2: Time and cost of full novelty evaluation per model.

E Full novelty definition

Our definition of novelty comes from the PeerRead paper (Kang et al., 2018). We reproduce it here:

Novelty Definition

How original is the approach? Does this paper break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?
Note that a paper could score high for originality even if the results do not show a convincing benefit.

5 = Surprising: Significant new problem, technique, methodology, or insight – no prior research has attempted something similar.

4 = Creative: An intriguing problem, technique, or approach that is substantially different from previous research.

3 = Respectable: A nice research contribution that represents a notable extension of prior approaches or methodologies.

2 = Pedestrian: Obvious, or a minor improvement on familiar techniques.

1 = Significant portions have actually been done before or done better.