GraphMind: Unveiling Scientific Reasoning through Contextual Graphs for Novelty Assessment

Anonymous ACL submission

Abstract

Assessing scientific novelty is inherently complex, requiring evaluation of both a paper's internal structure and its contribution within the broader research landscape. Existing large language model (LLM) approaches often rely on surface-level similarity or citation retrieval, overlooking the integration between a paper's content and its contextual grounding in related literature. To address this gap, we introduce SciNova, a benchmark containing 3,063 papers from both ICLR and NeurIPS, with peer review ratings for novelty prediction. SciNova represents each paper as a hierarchical graph that captures its claims, methods, and experiments (*micro-level*), while linking it to related paper graphs that encode both cited and semantically similar works (*macro-level*). Building on this benchmark, we propose GraphMind, a model that jointly processes micro- and macro-level structures for novelty prediction and rationale generation. Experiments show that by incorporating the bi-level graph, GraphMind significantly outperforms baseline LLMs in both novelty rating prediction and rationale generation.

1 Introduction

Large Language Models (LLMs) have demonstrated significant potential in understanding and analyzing scientific literatures (Messeri and Crockett, 2024). They have been employed in various research-related tasks, including extracting key information (Dunn et al., 2022), generating novel research ideas (Si et al., 2025; Gu and Krenn, 2025), producing literature reviews (Yuan et al., 2022; Du et al., 2024), and even supporting entire research pipelines, i.e., performing research independently and communicating their findings (Buehler, 2024; Lu et al., 2024). Among these applications, scientific novelty assessment is particularly critical, as it serves as the foundation for key research tasks such as literature review, hypothesis generation, and research evaluation (Zhao and Zhang, 2025).

Scientific novelty assessment is inherently challenging, extending far beyond conventional document-level information extraction. It should involve evaluating not only the internal quality and organization of a paper, including structure, motivation, and logic between sections, but also its position within the broader research landscape. Therefore, this includes determining how the paper relates to prior work: Has similar research already been conducted? Does the paper introduce a genuinely groundbreaking idea, or is it merely an incremental contribution compared to existing literature?

Existing LLM-based approaches to novelty assessment can be broadly categorized into two groups. The first focuses on retrieving potentially related papers through citation networks, as implemented in tools such as Connected Papers[1] and Inciteful[2]. The second relies on surface-level semantic similarity, employing keyword or title matching techniques, as seen in tools like Litmaps[3]. Additional systems attempt to assess paper quality based solely on textual content (Kang et al., 2018; Lin et al., 2024). However, these approaches typically lack integration between a paper's internal content and its broader research context, a critical component for effectively evaluating scientific novelty.

To address these limitations, we introduce SciNova, a benchmark specifically designed to systematically evaluate the capability of LLMs to assess novelty in research papers. Unlike existing datasets that typically provide only the plain text from either abstract or review, SciNova extracts key information from the full paper and represents it as a hierarchical graph (**micro-level**) as shown in Figure 1. Within each paper (left), we extract its claims, methodologies, and experiments, providing comprehensive evidence of idea formulation and

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

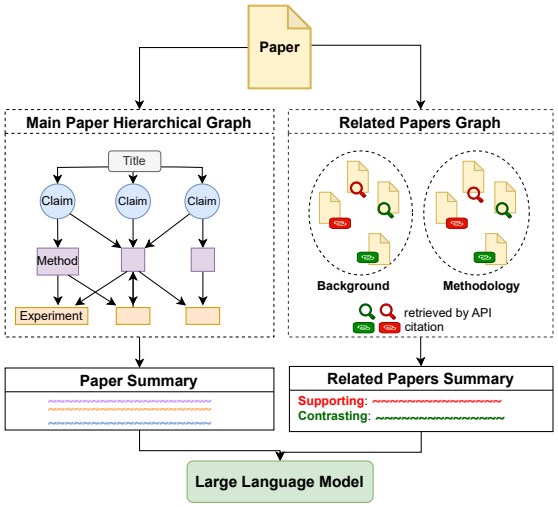

Figure 1: Overview of proposed benchmark, SciNova, and the proposed method, GraphMind, for novelty assessment. We extract the hierarchical graph from the target paper, as well as the related paper graph as part of the benchmark. GraphMind summarises the structured information from the benchmark for novelty assessment (novelty score prediction and rationale generation).

validation. This design aligns with the multifaceted definition of novelty, which may arise from introducing a new research problem or proposing new methods for existing problems. On the right, we build a related paper graph (**macro-level**) that integrates both top-cited works and semantically similar to support novelty assessments, providing richer contextual grounding from the literature. The combination of the micro-level and macro-level graph forms the basis for generating reports that highlight both supporting and contrasting evidence from the literature.

Building upon this dataset, we propose GraphMind, a method to leverage a hierarchical graph architecture to effectively process and integrate contextual information. Extensive experiments demonstrate that by incorporating both the hierarchical papers graph and the related paper graph, GraphMind consistently outperforms existing baseline models on the novelty assessment task. In summary, our contributions are:

- We introduce SciNova, a **new benchmark** for scientific novelty assessment, built from ICLR and NeurIPS papers with structured paper graphs and related works retrieved via the Semantic Scholar API.
- We propose a **novel method for novelty assessment**, GraphMind, which leverages both hierarchical paper graphs and related papers graphs, explicitly modelling the interactions between these components, going beyond simple contextual similarity.
- Empirically, we show that with the incorporation of the new benchmark SciNova, GraphMind demonstrates a stronger understanding of both the literature context and paper-level details, achieving higher accuracy in novelty classification. Furthermore, our qualitative analysis of LLM-generated rationales indicates improved faithfulness and factuality.

2 Related Work

Paper assessment benchmarks. We summarise existing benchmarks for paper assessment in Table 1. Most benchmarks target general approval decisions (Yuan et al., 2021; Bharti et al., 2021; Fernandes and Vaz-de Melo, 2022; Chitnis et al., 2024) rather than novelty specifically. PeerRead (Kang et al., 2018) provides expert-annotated aspects including novelty, while SchNovel (Lin et al., 2024) provides novelty assessment through temporal comparison of papers. Benchmarks also differ in input data granularity, ranging from full parsed paper text (Kang et al., 2018; Yuan et al., 2021; Bharti et al., 2021) to reviews only (Fernandes and Vaz-de Melo, 2022) or abstract and metadata alone (Lin et al., 2024). Only SchNovel provides related papers, randomly sampled for few-shot prompting. SciNova combines novelty-focused assessment with full structured paper content and related papers obtained through citations and similarity matching.

Novelty assessment methods. To conduct the novelty assessment, many existing works rely on lexicon similarity, from sentence-level to document-level (Ghosal et al., 2021; Tsai and Zhang, 2011; Ai et al., 2024; Ruan et al., 2023). For example, Ai et al. (2024) proposes NovAScore, a method of determining the novelty of a document in a given corpus by comparing its atomic content units (ACUs). The novelty assessment method then retrieves similar ACUs by cosine similarity, and calculates the final score depending on how novel and salient the ACUs are concerning the corpus. However, we argue that such semantic similarity fails to capture the scientific innovation and creativity (Zhao and Zhang, 2025).

[4]PeerRead also includes papers from ACL and NeurIPS, but these don't have the originality score.

Benchmarks	Paper Content	Related papers	Target Metric	Size	Paper Source
PeerRead (Kang et al., 2018)	parsed PDF	N.A.	originality score	183	ICLR 2017[4]
ASAP-Review (Yuan et al., 2021)	parsed PDF	N.A.	approval	8,877	NeurIPS 2016-2019, ICLR 2017-2020
PEERAssist (Bharti et al., 2021)	parsed PDF	N.A.	approval	4,500	ICLR 2017-2020
Fernandes and Vaz-de Melo (2022)	peer reviews	N.A.	approval	4,035	ICLR 2019-2021
SchNovel (Lin et al., 2024)	abstract and metadata	random sample	comparative novelty	15,000	ArXiv dataset
AutoRef (Chitnis et al., 2024)	original PDF	N.A.	approval	300	ICLR 2023
SciNova (ours)	**structured full paper**	**citation, semantic API**	novelty score	3063	ICLR 2022-2025, NeurIPS 2022-2024

Table 1: Comparisons of existing scholarly paper novelty assessment. Beyond plain text in abstract and review, we (i) extract hierarchical information throughout the paper (ii) incorporate related paper for novelty assessment.

3 SciNova Benchmark

We collect data at both the micro and macro levels. That is, within the paper itself (§3.2) and in relation to the broader scientific landscape in related paper graph (§3.3).

3.1 Metadata for the Benchmark

Our benchmark overview is shown in Figure 1. For each target paper, we extract a ***Hierarchical Graph*** from the target paper, including the title, claims, methods, and experiments (e.g. models tested, datasets, findings and conclusions). To compare against the related papers, we build the ***Related Paper Graph***, with both cited papers (references included by the target paper), and related papers retrieved by the Semantic Scholar API. To support novelty evaluation beyond simple semantic similarity, we consider at least two types of novel contributions in academic papers. The first involves introducing a new research question that has previously been overlooked; the second entails proposing a novel methodology for an existing problem setup. Accordingly, we categorise related papers based on their background relatedness and methodological relatedness (as illustrated by the two dashed circles in the related paper graph). Furthermore, we classify these related papers into two groups: positive (supporting) and negative (contrastive).

Data resource. We collect target papers from ICLR from 2022 to 2025 and NeurIPS from 2022 to 2024 [5], extracting metadata such as title, abstract, authors, and publication date. We also collect all reviews, including the main peer reviews with their review scores, and meta reviews containing the final approval decisions. Where available, we use the technical and empirical novelty ratings as our target. Otherwise, we adopt the contribution rating as a proxy. We use the median scores from peer reviews as our ground truth. In case of ties, we use the review with the highest confidence. These scores range from 1 to 5 (see Appendix J for the complete scoring rubric), where we treat scores of 1-3 as "not novel" and 4-5 as "novel".

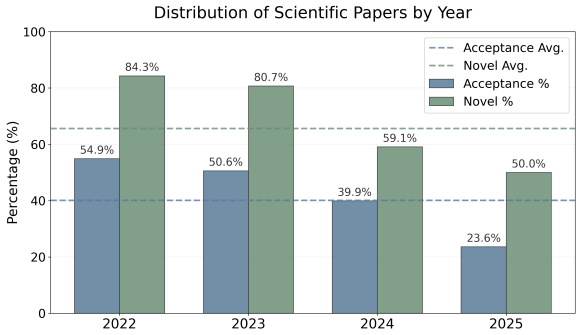

Figure 2: Distribution of scientific papers by year with acceptance rate and novelty rates.

As withdrawn/rejected papers are not always available on arXiv, the initial set we collected is skewed towards accepted papers, which does not reflect the actual ICLR acceptance rate[6]. To correct this, we resample the dataset to match the real-world acceptance ratio. In the ICLR statistics, around 40% of papers are approved, excluding the withdrawals.

Data statistics. We have 3,063 papers as the resulting benchmark for all experiments. Figure 2 shows the dataset's distribution of acceptance rate and novelty labels across publication years. The overall acceptance rate is 40%, though this varies by year. An important observation is that acceptance does not necessarily imply novelty: papers may be accepted for other strengths, such as strong experimental results or thorough empirical validation."

3.2 Micro-level: Hierarchical Paper Graph

Existing novelty assessment methods typically rely on either the abstract or the full paper as input (Kang et al., 2018; Wang et al., 2024). While

[5] We use the OpenReview API to extract the metadata, and the arXiv API to download the LaTeX source for accurate paper content parsing.

[6] ICLR 2024 Fact Sheet.

the abstract offers a high-level summary, it often omits critical details necessary for evaluating novelty, such as related work, methodological specifics, and experimental results. On the other hand, using the full paper presents challenges for LLMs, particularly in effectively extracting key information from long and complex contexts (Li et al., 2025).

Importantly, research papers inherently follow a hierarchical structure, with sections, subsections, and logical connections between different components. The paper introduction lays out the key aspects of the paper and makes claims about what the paper aims to accomplish. The methodology describes the methods used to execute the claims, such as the tasks, algorithms and models. The experiments validate the claims in the form of hard evidence from executing the methods. Each of these can be distilled to its main idea and summarised.

Therefore, we propose the hierarchical graph, an approach to extract the hierarchical information from the target papers, and the related papers graph to support multi-faceted novelty assessment. Specially, our extracted hierarchical graph is a Directed Acyclic Graph (DAG), where each node represents a key concept, and edges define their dependencies. Each node is composed of labels summarising each idea, and detail texts explaining each one in depth.

- For claim nodes, we summarise what the paper claims to contribute, especially claims made in the abstract, introduction, discussion and conclusion.
- For method nodes, we identify the methods used to validate the claims. These include the key components: algorithms, theoretical framework or novel techniques introduced.
- For experiment nodes, we include the models, baselines, datasets, etc. used in experiments to validate the methods and their conclusions.

The detailed description of the nodes and links is provided in Appendix B.2. The graph is constructed by prompting GPT-4o-mini with instructions about the different types of nodes, their relationships, the paper title, abstract, and the full textual content, including tables formatted as Markdown using the prompt in Figure E (Appendix).

3.3 Macro-level: Related Paper Graphs

In addition to the information in the target paper itself, the comparison with related papers can situate the paper within the broader research landscape for novelty assessment. Therefore, we retrieve the related papers through the reference section

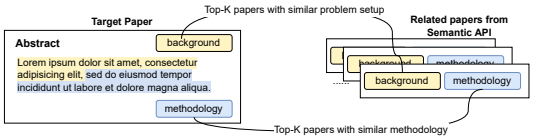

Figure 3: Diagram of semantic neighbour retrieval.

within the paper and query Semantic Scholar API for broader comparison, indicated as *citation* and *retrieved by API*, respectively, in Figure 1: related papers graph. The details of how to extract the papers from the two sources are as follows.

Citation graph. For references included in the target paper, we prioritise the ones with the highest similarity with the target paper. We divide the references into two polarities: positive, supporting the target paper's claim, or negative, contrasting or critiquing the cited work. This polarity is determined using an LLM that analyses the citation context (the sentence where the citation appears) and classifies it as supporting or contrasting. Finally, we retrieve the top-K supporting and contrasting citations, for a total of $2 \times K$ citations. Section B.3 describes this process in more detail.

Semantic neighbours graph. Relying solely on citations has limitations: authors may miss relevant work or omit certain references due to bias. To mitigate this, we use Semantic Scholar's recommendation API, which suggests related papers based on content similarity. For each target paper, we retrieved 30 recommended papers using the Semantic Scholar recommendation API, including only papers published before the target paper's publication date.

Connecting related papers with multifaceted novelty aspects. Due to the lack of citation context, we are unable to determine the citation polarity. Instead, we consider two types of novelty contributions in scientific papers: novel problem setup and methodology. To achieve it, we take the abstract of each retrieved paper and use an LLM to separate it into two parts: *background* and *methodology*.

- The *background* describes the problem setup, motivation, rationale, task and previous works.
- The *methodology* describes the methods, objectives, goals, findings, results, implications, and limitations.

We extract backgrounds and methodologies from

the target paper and each retrieved paper, computing similarity scores between corresponding components using SentenceTransformers (Reimers and Gurevych, 2019). We select the top-K papers by background similarity and top-K by methodology similarity, yielding $2 \times K$ papers total. Figure 3 illustrates this process.

As a result, each target paper is linked to $2 \times K$ semantically related papers from both background-related and methodology-related, along with $2 \times K$ related papers from citation graph. All together provide a richer, more structured understanding of its novelty.

4 Multi-faceted Novelty Evaluation

Based on the benchmark dataset, with the hierarchical graph representing the target paper and the related papers graph, we discuss how we organise those structured inputs for LLMs in paper novelty.

Figure 4 displays the multiple components in the prompt fed to LLMs, including the paper summary obtained from the hierarchical graph and the related paper summaries from supporting and contrasting papers from the related paper graph. Section D (Appendix) shows the full prompt text. Moreover, we provide the novel assessment criteria according to the ACL 2016 reviewer guidelines, via Kang et al. (2018). These criteria are based on whether the paper introduces a new topic, methodology or analysis to its field, and how innovative its research is. This means that even if the paper's results aren't convincing (e.g. performance results are poor), it could still be novel.

Prompt for Novelty Assessment

```
The following data contains information about a
scientific paper.

It includes the target paper's title, a summary
of its key points and some related papers. The
paper summary describes... The related papers
are...

Based on this, decide whether the paper is
novel. It is novel if... First, generate the
rationale for your novelty label, then give the
final novelty label.

-Data-
Title: {title}
Abstract: {abstract}
Paper summary: {text graph}
Supporting papers: {supporting}
Contrasting papers: {contrasting}
```

Figure 4: The template used to prompt LLM for novelty assessment, consisting of i) description of the target paper and ii) comparison with the related papers.

The components used to build the prompt are:

- **`title`**: Target paper title
- **`abstract`**: Target paper abstract
- **`text graph`**: Textual version of the hierarchical graph
- **`contrasting`**: Summaries of the contrasting related papers
- **`supporting`**: Summaries of the supporting related papers

Target paper hiearchical graph. To incorporate the hierarchical paper graph as part of the input for the evaluation model, we need to convert it to a textual format. Since our hierarchical graph is a Directed Acyclic Graph (DAG), we can use topological sorting to transform the graph structure into a linear sequence of nodes. We convert each node into sentences by incorporating their types, labels and detail texts. The collection of these sentences is used as the textual representation of the graph in our prompt as **`text graph`**.

Related papers. We fetch four types of papers from the related papers graph: positive citations, negative citations, background- and methodology-related papers, with K papers for each category, totalling $4 \times K$ related papers. We use GPT-4o-mini to summarise their abstracts relative to the target paper to obtain a *related paper summary* for each paper. We use different prompts for supporting and contrasting papers, as shown in Figure F (Appendix).

When adding related papers to the prompt, we combine the titles and *related paper summaries* as the **`supporting`** and **`contrasting`** components.

Rating and rationale generation. After aggregating all the extracted information, we prompt the LLM to predict a novelty score (0 or 1) and generate a structured rationale justifying its prediction. This rationale summarises the paper, highlights both supporting and contrasting evidence, and gives the final rating. We chose a binary classification instead of a multiple-class scale (e.g. 1-5) because we found that the models disproportionately favoured the extremes (e.g. often 1 or 5, rarely 2 or 4). Using binary outputs circumvents that issue and allows us to focus on improving the model's underlying reasoning capabilities.

5 Experiments

We provide the `SciNova` and PeerRead datasets with extracted structured information to LLMs to

Model	Precision	Recall	F1	Accuracy
SciNova (ours)				
$\text{Basic}_{\text{Llama}}$	0.5480	0.5555	0.5475	0.6532
NovAScore	0.4706	0.4947	0.4418	0.7171
$\text{SciMON}_{\text{GPT-4o}}$	0.5509	0.5951	0.3792	0.6466
$\text{Basic}_{\text{GPT-4o}}$	0.6063	0.6163	0.6068	0.6280
$\text{GraphMind}_{\text{GPT-4o}}$	**0.6364**	**0.6499**	**0.6265**	**0.7360**
PeerRead				
$\text{Basic}_{\text{Llama}}$	0.4489	0.3747	0.3578	0.7758
NovAScore	0.4851	0.4901	0.4856	0.8304
$\text{SciMON}_{\text{GPT-4o}}$	0.4052	0.4278	0.4161	0.8364
$\text{Basic}_{\text{GPT-4o}}$	0.4753	0.4800	0.4726	0.8118
$\text{GraphMind}_{\text{GPT-4o}}$	**0.4889**	**0.4932**	**0.4882**	**0.8471**

Table 2: Evaluation results on the SciNova and PeerRead datasets.

verify if the newly added structured information can improve novelty assessment in § 5.1. We also ablate the effects of each component in §5.2, as well as provide evaluation on generated rationale in §5.3 and §5.4.

Models. We evaluate the novelty score prediction with the four backbone models: Llama-3.1-8B (Dubey et al., 2024), Qwen-2.5-7B (Yang et al., 2024), GPT-4o and Gemini-2.0-Flash. We train the Llama and Qwen models to predict the binary novelty score as a generation task using supervised fine-tuning on the SciNova training set, given the paper information and instruction (see Figure A6 for dataset statistics). GPT-4o and Gemini-2.0-Flash are directly prompted.

5.1 Comparison for Novelty Score Prediction

Comparison with baselines To verify the effectiveness of our benchmark, pipeline and proposed method, GraphMind, we test the novelty assessment on multiple LLM variants. We also compare against the following baselines. Appendix I describes our baseline implementations in detail.

Basic: A simple baseline that uses only the paper title and abstract as input.

SciMON (Wang et al., 2024): Retrieves "inspirations" from an existing corpus of papers via semantic and citation neighbours. These inspirations are used as additional input for the novelty assessment method.

NovAScore (Ai et al., 2024): Breaks down documents into Atomic Content Units and uses those to calculate a score aggregating novelty and salience. This score is converted to a binary novelty label.

The results in Table 2 show that GraphMind is the best across both datasets, achieving better accuracy, precision, recall and F1 score, with an improvement of 17% in accuracy between GraphMind and the $\text{Basic}_{\text{GPT-4o}}$ baseline on SciNova and 4% on PeerRead. Note that the results from SciNova and PeerRead differ significantly. This is because PeerRead is more imbalanced than SciNova (88% vs 65% novel papers).

Results with different backbones. We sampled a dataset of 100 items with balanced labels (SciNova-100), which we used to evaluate more combinations across different backbones. We also include LLMs facilitated with search tools, using their built-in tools from the respective APIs. The search LLMs are instructed to search the web for similar papers and we remove the in-context demonstrations to avoid any confusion.

Table 3 shows the results of this comparison subset. For GPT-4o, search often results in more noise than helpful results, contributing to the worsening performance when compared to the Basic version. Gemini, on the other hand, had more relevant search results, which contributed to an improvement in performance relative to the Basic version. However, this was still not enough to surpass the full graph version.

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 3: Results on the SciNova-100 dataset for various LLM backbones.

5.2 Ablation Studies

We select a few variations of our method to show how each component contributes to the performance. Table 4 shows the results on our benchmark dataset SciNova.

Abstract only: only the paper title and abstract as the input, without the hierarchical graph or the related papers graph. This is the same as the *Basic* baseline.

Abs. + Related G.: only title and abstract, with full related paper graph.

Variants	Precision	Recall	F1	Accuracy
Abstract only	0.6063	0.6136	0.6068	0.6280
Abs. + Related G.	**0.6613**	**0.6699**	0.6231	0.7244
H.G. only	0.5914	0.5901	0.5907	0.7313
H.G. + Citation	0.6152	0.6255	0.6131	0.7290
H.G. + Semantic	0.6242	0.6363	0.6114	0.7174
GraphMind	0.6364	0.6499	**0.6265**	**0.7360**

Table 4: Ablation results on SciNova. The best results for each metric are in **bold**, and the second best are underlined.

H.G. only: only the hierarchical graph, but no related papers.
H.G. + Semantic: with the hierarchical graph, but only semantic related papers.
H.G. + Citation: with the hierarchical graph, but only cited related papers.

The results demonstrate that both components are essential. The abstract-only baseline achieves 0.6280 accuracy, while the full method reaches 0.7360, a 17% increase. Individual components substantially outperform the baseline: the hierarchical graph alone achieves 0.7313, and the related papers graph 0.7244. Within related papers, citations alone (0.7290) outperform semantic neighbours alone (0.7174), with both combined yielding 0.7244. The full method's superior performance (0.7360) over any single component confirms that the components provide complementary information.

5.3 Automated Rationale Evaluation

In addition to evaluating the novelty score as a classification problem, we also use LLM-as-judge to evaluate our generated rationales and compare them with the original peer-review. We run a pairwise tournament amongst all evaluated models and the original human rationale. We use a Bradley-Terry model (Bradley and Terry, 1952) to generate model ratings for each metric. We use GPT-4o as our judge with zero-shot prompts describing each evaluation metric. The full tournament setup and prompt can be seen in Appendix C.

We evaluate the following multiple 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 correctly grounded in scientific facts from the target and related papers?

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

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

Evaluation Results of Generated Rationales The results in Table 6 support the idea GraphMind has access to more information than the baselines (the hierarchical and related paper graphs). These enable it to ground the rationales in an effective understanding of the paper (coming from the hierarchical graph and supporting the clarity and faithfulness ranks), and the surrounding research context (from the related papers and supporting the specificity and contribution ranks). In comparison, the baselines lack the same type of information.

It should be noted that the rationales generated by the automated methods concern only the paper's novelty. The human rationales we compare against them in this evaluation also discuss other factors, such as soundness, typos, style, etc. This explains how some human ratings could be lower, as the rationales are being compared with more focused assessments from the models.

This explains why the Human variant ranks low on items such as specificity, as it covers topics other than novelty, and faithfulness, as the rationale's sentiment can differ among aspects of the paper. On the other hand, human-written rationales are still easier to understand than automated ones, and human reviewers show good understanding of the paper contributions.

5.4 Case Study and Error Analysis

To examine the improvement source, we display the generated rationale from our baseline (*Basic*) and GraphMind in Table 5. A key issue with *Basic* is that it doesn't have enough information for an informed decision. It offers almost no contextualization of the target paper in the literature, and its understanding of the paper content is superficial, as it only has information from the abstract. On the other hand, our full method highlights the relevant information in both the target paper content, via the hierarchical graph, and literature, via related papers, explicitly mentioning contrasting and supporting works.

It's also interesting to note that the rationale generated from our method follows a specific and desirable structure: it first describes the target paper, then the evidence for and against it, and finally summarises everything into a single evaluation explana-

Generated Rationale	
Basic	This paper addresses ... The authors propose ... Their findings reveal significant drops in accuracy when model selection strategies depend on target labels, which is a critical insight for the field. The introduction of the BenchmarkPDA framework for reproducibility is also a valuable contribution. **However, while the work is important for the community, it does not introduce fundamentally new algorithms or methodologies; rather, it critiques existing practices and provides a structured evaluation.** Therefore, while the findings are significant, **the novelty lies more in the evaluation approach than in the development of new techniques.**
`GraphMind`	The target paper presents a significant contribution to the field of Partial Domain Adaptation (PDA) by providing a realistic evaluation of existing methods and their model selection strategies. It highlights the critical impact of using target labels for model selection, revealing a substantial drop in accuracy when these labels are not available. This finding is crucial as it addresses a gap in the current understanding of PDA methodologies.The supporting papers reinforce the importance of systematic evaluation and model selection in similar contexts,while the contrasting papers introduce alternative approaches and critiques that do not undermine the novelty of the target paper but rather highlight its relevance in the ongoing discourse. Overall, the target paper's focus on realistic evaluation and its findings contribute new insights to the field, warranting a novelty rating of 1.

Table 5: Rationale comparison between versions with highlighted differences. In the Basic version, highlighted in **bold** are the key parts of the rationale. The `GraphMind` version shows improvements, with entities from the hierarchical graph, such as **claims**, **methods** and **experiments** nodes. It also shows supporting and contrasting related papers.

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

tion, and gives the predicted label. This allows the reader to understand how each aspect influenced the final prediction.

Error analysis. We also noticed two main sources of disagreement between `GraphMind` and the human evaluation:

Logical incoherence: the human annotation gives a novelty label incompatible with the rationale. For example, if the review is mostly positive but the paper is annotated as not novel.

Insufficient evidence: the model concludes that the paper provides enough evidence (such as literature review and experiments) to deem the paper novel, but the human annotations require more evidence to reach the same conclusion.

The logical incoherence issue is reasonable, as it can happen when writing reviews. The reviewer might have forgotten to add details that support their argument, or mistakenly believed their point was sufficiently explained. It's an inherent problem with human annotations, and it's something an automated approach can mitigate, as it can be more consistent.

The insufficient comparison aspect highlights that the model still has a lacklustre understanding of the paper, even if it's better than other approaches. It's usually able to understand the goals and proposed contributions, but understanding the experiments and whether they provide enough evidence can be challenging.

6 Conclusion

We created `SciNova`, a novelty assessment benchmark based on ICLR papers from 2022 to 2025 and NeurIPS papers from 2022 to 2024, with submission information from OpenReview and paper content from arXiv LaTeX files. We also included top-cited and recommended papers from the Semantic Scholar API to build a network of related papers. Through experiments evaluating the classification results and generated rationales across our original method, `GraphMind`, existing baselines and several ablations, we've shown that our method performs well and addresses the limitations of existing approaches.

Limitations

Our dataset was limited to only ICLR and NeurIPS conferences, as they were the only ones where retrieving a large number of papers was feasible through the OpenReview API that also had the required information. Not all papers were used, as we could only use those with LaTeX code on arXiv. An alternative would be to parse the PDFs available in OpenReview directly, but we found that unreliable.

References

Lin Ai, Ziwei Gong, Harshsaiprasad Deshpande, Alexander Johnson, Emmy Phung, Ahmad Emami, and Julia Hirschberg. 2024. Novascore: A new auto-

mated metric for evaluating document level novelty. *Preprint*, arXiv:2409.09249.

Prabhat Kumar Bharti, Shashi Ranjan, Tirthankar Ghosal, Mayank Agrawal, and Asif Ekbal. 2021. Peerassist: leveraging on paper-review interactions to predict peer review decisions. In *International Conference on Asian Digital Libraries*, pages 421–435. Springer.

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.

Markus J. Buehler. 2024. Accelerating scientific discovery with generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning. *Machine Learning: Science and Technology*.

Soham Chitnis, Manasi Patwardhan, Ashwin Srinivasan, Tanmay Tulsidas Verlekar, Lovekesh Vig, and Gautam Shroff. 2024. Autoref: Generating refinements of reviews given guidelines. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 175–190.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S Weld. 2019. Pretrained language models for sequential sentence classification. *arXiv preprint arXiv:1909.04054*.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, and 21 others. 2024. LLMs assist NLP researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099, Miami, Florida, USA. Association for Computational Linguistics.

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.

Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S. Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. 2022. Structured information extraction from complex scientific text with fine-tuned large language models. *Preprint*, arXiv:2212.05238.

Arpad E Elo. 1967. The proposed uscf rating system, its development, theory, and applications. *Chess life*, 22(8):242–247.

Gustavo Lúcius Fernandes and Pedro O. S. Vaz-de Melo. 2022. Between acceptance and rejection: challenges for an automatic peer review process. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, JCDL '22, New York, NY, USA. Association for Computing Machinery.

Tirthankar Ghosal, Vignesh Edithal, Asif Ekbal, Pushpak Bhattacharyya, Srinivasa Satya Sameer Kumar Chivukula, and George Tsatsaronis. 2021. Is your document novel? let attention guide you. an attention-based model for document-level novelty detection. *Nat. Lang. Eng.*, 27(4):427–454.

Xuemei Gu and Mario Krenn. 2025. Interesting scientific idea generation using knowledge graphs and llms: Evaluations with 100 research group leaders. *Preprint*, arXiv:2405.17044.

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.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2025. Long-context LLMs struggle with long in-context learning. *Transactions on Machine Learning Research*.

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

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

Lisa Messeri and MJ Crockett. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58.

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.

Xuanmin Ruan, Weiyi Ao, Dongqing Lyu, Ying Cheng, and Jiang Li. 2023. Effect of the topic-combination novelty on the disruption and impact of scientific articles: Evidence from pubmed. *Journal of Information Science*, 0(0):01655515231161133.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2025. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *The Thirteenth International Conference on Learning Representations*.

Flora S. Tsai and Yi Zhang. 2011. D2s: Document-to-sentence framework for novelty detection. *Knowl. Inf. Syst.*, 29(2):419–433.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. Scimon: Scientific inspiration machines optimized for novelty. *Preprint*, arXiv:2305.14259.

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.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *J. Artif. Int. Res.*, 75.

Yi Zhao and Chengzhi Zhang. 2025. A review on the novelty measurements of academic papers. *Scientometrics*, 130:727–753.

Appendix

A Baseline LLMs evaluation

To highlight how poorly baseline LLMs perform in the novelty assessment task, we built a small baseline experiment. We sampled 100 entries from our benchmark dataset SciNova and prompted GPT-4o[7] and O3-mini[8] from OpenAI, Gemini 2.0 Flash from Google[9] and DeepSeek V3-0324[10].

Figure A shows the prompt used. The demonstrations came from randomly sampled entries of the training split: 5 from not novel papers and 5 from novel ones. They contain the paper title, the novelty label and the rationale.

The models with search used the prompt shown in Figure A. The GPT model used for search was GPT-4o-search-preview with low search context size, and the Gemini one was Gemini-2.0-Flash.

Prompt for baseline LLMs

```
The following data contains
information about a scientific
paper. It includes the paper's
title and abstract.

Based on this content, decide
whether the paper is novel enough
or not. If it is, give it a label
of 1. If it isn't, give it a
label of 0. This should reflect
how much the paper brings and
develops new ideas previously
unseen in the literature. First,
generate the rationale for your
novelty rating, then give the
final novelty rating.

The output should have the
following format:

```
Rationale: <text>

Label: <0 or 1>
```

#####
{demonstrations}

-Data-
Title: {title}
Abstract: {abstract}
```

Prompt for search LLMs

```
The following data contains
information about a scientific
paper. It includes the paper's
title and abstract.

First, search the web for
publications related to the
paper. Your goal is to find
relevant papers to compare the
target paper with. This would be
important to determine if the
paper's contributions are novel.

Based on this content, decide
whether the paper is novel enough
or not. If it is, give it a label
of 1. If it isn't, give it a
label of 0. This should reflect
how much the paper brings and
develops new ideas previously
unseen in the literature. First,
generate the rationale for your
novelty rating, then give the
final novelty rating.

The rationale must include the
documents retrieved by web
search. It must be pure plain
text without any formatting.
Instead of writing the titles and
links to the documents inside the
rationale, assign each a number
and list them (number, title and
link) at the bottom of the text.

The output should have the
following format:

```
Label: <0 or 1>

Rationale: <text>
```

#####
-Data-
Title: {title}
Abstract: {abstract}
```

B Details of Benchmark Creation

B.1 Full paper content

To build our paper hierarchical graph, we need the full content. The OpenReview API gives us the PDF for each paper, but parsing PDFs is not reliable enough for our case. Instead, we use the arXiv API to locate papers submitted to the conferences that were also uploaded as preprints. We take the LaTeX code from arXiv, parse the references (including their surrounding contexts in the text) and transform the content to Markdown. However, only about 40% of the papers from the OpenReview

[7] https://openai.com/index/gpt-4o-system-card/
[8] https://openai.com/index/openai-o3-mini/
[9] https://deepmind.google/technologies/gemini/
[10] https://github.com/deepseek-ai/DeepSeek-V3

API have a corresponding arXiv version, and those are skewed towards accepted papers. We focus on this subset to ensure reliable content extraction and resample the dataset to match the acceptance observed in the overall dataset.

B.2 Hierarchical Graph Components

Our hierarchical graph has the following node types:

- Title: the title of the paper.
- Keywords: keywords summarising the main topics of the paper.
- Primary area: what scientific primary area the paper is from. The possible areas come from ICLR[11].
- TLDR: a sentence that summarises the paper from the abstract.
- Claim: summarises what the paper claims to contribute, especially claims made in the abstract, introduction, discussion and conclusion.
- Method: for each claim, identifies the methods used to validate the claims from the method sections. These include the key components: algorithms, theoretical framework or novel techniques introduced.
- Experiment: what models, baselines, datasets, etc. were used in experiments to validate the methods and their conclusions.

The nodes have specific relationships between them that a valid graph must maintain:

- There is a single title node, and it connects to the primary area, TLDR and keyword nodes.
- TLDR connects to all claim nodes.
- Each claim node connects to one or more method nodes.
- Each method node connects to one or more experiment nodes.
- Every method or experiment node must be connected.

B.3 Citation Graph

Citation context polarity. To build our related paper graph, we need to determine whether a citation supports or contradicts the target paper making it. This is because authors can cite papers so that they add supporting evidence for their claims, or use them as contrasting points to, for example, argue that their methodology is superior to the ones that came before.

[11] https://iclr.cc/Conferences/2024/CallForPapers

We use GPT-4o-mini with zero-shot prompting as our method for determining these polarities. For each reference in the target paper, we query the LLM with the citation context (the sentence where the citation appears) and ask it to provide a positive or negative label.

Suppose there are multiple citation contexts for a given reference (i.e. the same paper was cited in different contexts). In that case, we obtain the polarity for each context separately and use the majority label to obtain the final polarity for the reference. If there is a tie, we default to a positive label.

We validated the quality of these classification results by manually annotating 100 entries from the PeerRead dataset. Table A1 shows the classification metrics.

Metric	Value
Precision	0.9000
Recall	0.8710
F1	0.8852
Accuracy	0.8372

Table A1: Evaluation metrics for citation context polarity classification.

Abstract splitting. We use GPT-4o-mini to split a paper abstract into sections concerning the context (problem setup, motivation, task, etc.) and target (methods, objectives, results, etc.). Figure B.3 shows the prompt we used. The `demonstrations` are taken from the CSAbstruct dataset (Cohan et al., 2019).

Prompt for abstract splitting

```
Given the paper abstract, your
goal is to extract the paper's
background context and target.

The background context describes
the problem setup, motivation,
rationale, task and previous
knowledge. The target describes
the methods, objectives, goals,
findings, results or implications.

The output will contain two
fields: `background` and
`target`. For each of them,
collect the sentences of the
appropriate type. The output
should be the relevant sentences
combined for each type. All
sentences in the abstract must be
either a background context or a
```

```
target sentence. No sentence in
the abstract should be missing
from the output.

#####
{demonstrations}
-Data-

Abstract: {abstract}
```

C Rationale evaluation

We use GPT-4o as an LLM-as-judge model to evaluate the rationales generated by our automated methods. We use the prompt in Figure C.

The evaluations are performed as a tournament between the methods. For each paper instance in the evaluation dataset, we run pairwise matches between all models, twice for each pair in swapped orders. For example, a tournament with 3 methods A, B and C would have the matches A vs B, A vs C, B vs C, B vs A, C vs A, C vs B. We do this to account for potential position bias in the model. This means we have $n(n-1)$ comparisons for each instance in the evaluation dataset. This is done separately for each metric. In total, we have $n(n-1)M$ comparisons, where M is the number of metrics ($M = 5$, in this case).

Finally, we use the Bradley-Terry algorithm (Bradley and Terry, 1952) to compute the final rankings per metric. Bradley-Terry was chosen instead of the more commonly used Elo (Elo, 1967) because the latter is sensitive to the match order, while the former isn't. This is relevant because all matches should have equal weight, regardless of when they happened. This is because we use fixed model codes, so performance doesn't change over time.

Prompt for rationale evaluation

```
Abstract: {abstract}

# Rationales to Compare
## Rationale A
{rationale_a}

## Rationale B
{rationale_b}

# Evaluation Instructions
Compare these two rationales and
determine which one is better
specifically in  terms of
"{metric}".

{metric}: {definition}

# Output Format
Your output must be structured as
follows:
- Winner: A or B
- Explanation: A brief
explanation of your decision
```

D Novelty assessment

Figure D shows the full prompt used for novelty assessment using the hierarchical graph and related papers. Note that the prompt uses the graphs in their textual form, so there's no explicit mention of the graphs, only what they were transformed to.

Prompt for novelty assessment

```
The following data contains
information about a scientific
paper. It includes the target
paper's title, a summary of its
key points and some related
papers.

The paper summary describes the
most important information about
the paper and its  contents. It
summarises key aspects, which you
can use to build a more
comprehensive  understanding of
the paper.

The related papers are split into
"supporting" papers (those that
corroborate the paper's  ideas,
methods, approach, etc.) and
"contrasting" papers (those that
go against the  paper's ideas).
Use these related papers to
understand the context around the
target paper,  so you know what
other works exist in comparison
with the target paper.

Based on this, decide whether the
paper is novel. It is novel if
brings new ideas or  develops new
ideas previously unseen. Make
sure that the ideas are truly
unique. The  paper is not novel
if anything similar to it has
been done before. Be very
thorough.  When in doubt, tend
towards the not novel label.

First, generate the rationale for
your novelty label, then give the
final novelty label.  It should
be 1 for a novel paper, or 0
otherwise. If you're uncertain,
assign the 0  (not novel) label.

{demonstrations}

-Data-
```

```
Title: {title}
Abstract: {abstract}

Paper summary:
{text graph}

Supporting papers:
{supporting}

Contrasting papers:
{contrasting}
```

E Hierarchical graph extraction

Figure D shows the full prompt used to extract the hiearchical graph entities from the paper content.

Prompt for hierarchical graph extraction

```
The following data contains
information about a scientific
paper. It includes the
paper's title, abstract, and the
main text. The goal is to
represent all the relevant
information from the paper as a
graph.

Your task is to extract entities
of the following types and the
relationships between
them. All entities must have
different text descriptions.
- title: the title of the paper.
- primary_area: what scientific
primary area the paper is from.
It must be one from
the following list:
{primary_areas}.
- tldr: a sentence that
summarises the paper from the
abstract.
- claim: summarise what the paper
claims to contribute, especially
claims made in the
abstract, introduction,
discussion and conclusion. Pay
attention to the key phrases
that highlight new findings or
interpretations.
- method: for each claim,
identify the methods used to
validate the claims from the
method sections. These include
the key components: algorithms,
theoretical framework
or novel techniques introduced.
- experiment: what models,
baselines, datasets, etc. were
used in experiments to
validate the `methods` and their
conclusions.

Extract these entities and the
relationships between them. The
paper title is the root
entity. You must follow these
rules when generating the
entities and relationships:

- The title connects to the
primary_area, the keywords and
the tldr sentence.
- The tldr sentence connects to
all the claims.
- Each claim must connect to one
or more methods.
- Each method must connect to one
or more experiments.
- Every method and experiment
must connected to at least one
entity.

All entity types should be
present in the output. None of
the lists in the output can be
empty.

#####
-Data-
Title: {title}
Abstract: {abstract}

Main text:
{main_text}
```

F Related paper summarisation

Figure F shows the prompts used for related paper summarisation for supporting and contrasting papers, respectively.

Prompt for related supporting paper summarisation

```
The following data contains
information from a Target Paper
and a Related Paper. The Related
Paper has a positive relation to
the Target Paper. It contains
supporting information that
strengthens the target paper
claims.

Your task is to generate a
summary that highlights how the
Related Paper supports the Target
Paper. Your summary should be
short and concise, comprising a
few sentences only.

#####
-Data-
# Target paper

Title: {title_target}
Abstract: {abstract_target}

# Related paper

Title: {title_related}
Abstract: {abstract_related}
```

Prompt for related contrasting paper summarisation

```
The following data contains
information from a Target Paper
and a Related Paper. The Related
Paper has a negative relation to
the Target Paper. It is used to
contrast the claims made by the
Target Paper.

Your task is to generate a
summary that highlights how the
Related Paper contrasts the
Target Paper. Your summary should
be short and concise, comprising
of a few sentences only.

#####
-Data-
# Target paper

Title: {title_target}
Abstract: {abstract_target}

# Related paper

Title: {title_related}
Abstract: {abstract_related}
```

G Benchmark Statistics

Table A2 shows the distribution of paper approval decisions in the original dataset, and Table A3 shows this distribution in the final dataset, after downsampling. Table A4 shows the distribution of novelty labels. Table A5 shows the distribution of approval and novelty labels by year of publication.

Value	Count	Actual %
False	4475	70.78%
True	1847	29.22%
Total	6322	100.00%

Table A2: Distribution of approval decisions before resampling.

Value	Count	Actual %
Reject	1840	60.00%
Accept	1223	40.00%
Total	3063	100.00%

Table A3: Distribution of approval decisions.

Value	Count	Actual %
Not novel	1038	66.14%
Novel	2028	33.86%

Table A4: Distribution of novelty labels.

H Dataset splits

The `SciNova` dataset after down-sampling to a realistic approval rating distribution has 3,063 items. One of our baselines uses Llama, which requires distinct training and testing splits. We use the sizes shown in Table A6.

I Baseline implementation details

We reimplement SciMON (Wang et al., 2024) and NovAScore (Ai et al., 2024) as baselines to compare to our method.

SciMON constructs three graphs: semantic, citation and knowledge neighbours. For semantic neighbours, we convert the abstract to a vector; for citations, we use the title, and for knowledge neighbours, we break the abstract down to key ideas and connect these across papers. We use cosine similarity to find the most similar papers for each type of neighbour. From these, we use the key ideas as part of the novelty evaluation prompt.

In our NovAScore implementation, we use GPT-4o-mini to extract Atomic Content Units (ACUs) from the abstracts, using the prompt from the original paper. Each ACU is converted to a vector and added to a vector database, which we use for queries. We also extract ACUs for each evaluated paper, then query them in the database to find similar excerpts. We use NovAScore's formula to calculate the final score, which we convert to a binary class using a threshold of 0.3. Other hyperparameters are $k = 5$ for the query, similarity threshold of 0.8 for which results to consider, and the salient weights are $\alpha = 0.5$, $\beta = 0.5$, and $\gamma = 1$. We chose these hyperparameters via empirical tuning.

In both SciMON and NovAScore, we use `sentence-transformers` with the `all-MiniLM-L6-v2` to convert text to vectors with 768 elements. We use the `Annoy` library for vector database indices and similarity queries.

J Novelty scoring rubic

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

Year	Count	Count %	Approved	Approval %	Novel	Novel %
2022	534	17.4%	293	54.9%	450	84.3%
2023	688	22.5%	348	50.6%	555	80.7%
2024	929	30.3%	371	39.9%	549	59.1%
2025	912	29.8%	215	23.6%	456	50.0%
Total	3063	100.0%	1227	40.1%	2010	65.6%

Table A5: Distribution of scientific papers by year with approval and novelty rates.

Name	Count	Percentage
Train	1500	49%
Dev	500	16%
Test	1063	35%
Total	3063	100%

Table A6: Dataset split distribution.

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.

K Responsible NLP

Artifact intended use. The artifacts we used, the PeerRead dataset, ICLR and NeurIPS papers from OpenReview, Semantic Scholar API and arXiv papers, are all freely available for use in research, matching our use.

Package	Version	License
`annoy`	1.17.3	Apache-2.0
`arxiv`	2.2.0	MIT
`datasets`	3.6.0	Apache-2.0
`google-genai`	1.24.0	Apache-2.0
`matplotlib`	3.10.3	PSF
`networkx`	3.5.0	BSD
`nltk`	3.9.1	Apache-2.0
`numpy`	2.3.1	BSD
`openai`	1.93.0	Apache-2.0
`openreview`	1.50.1	MIT
`peft`	0.16.0	Apache-2.0
`sentence-transformers`	5.0.0	Apache-2.0
`thefuzz`	0.22.1	MIT
`tiktoken`	0.9.0	MIT
`torch`	2.7.1	BSD
`transformers`	4.53.0	Apache-2.0

Table A7: Python packages used during development.

Model size and budget. Our usage of LLM APIs (Gemini, OpenAI, DeepSeek, etc.) cost around 1,000 USD between development and final experiments. Supervised fine-tuning used Llama 3.1 8B and Qwen 2.5 7B on our university's GPU cluster.

AI usage. Models such as Gemini 2.5 Pro, OpenAI GPT-5 and Anthropic Claude Opus 4 were used to write code for this project and edit sections of this paper.

Packages used. Table A7 shows the packages used during development with their versions and licenses.

A complete listing of packages used with full version specification is available with the source code.