ClaimNOVA: Evidence-grounded Claim-level Novelty Evaluation for Academic Papers

Anonymous EMNLP submission

Abstract

Novelty assessment is a central yet highly challenging task in peer review, requiring reviewers to determine whether a paper's claimed contributions meaningfully differ from prior work. Existing LLM-based systems have explored evidence-grounded novelty assessment and can generate novelty evaluation reports. However, novelty is often not a single undifferentiated dimension: different novelty claims require different comparison criteria, such as comparing method claims with prior methods. Ignoring these type-specific distinctions may lead to poorly matched evidence and misleading novelty judgments. To address this issue, we propose ClaimNOVA, a type-aware evidence-grounded framework that assists reviewers in making claim-level novelty judgments. Following prior work, we use novelty claims in introduction as the main input, as they summarize the authors' self-claimed novelty. For each claim, we construct claim-targeted background evidence, selects relevant prior work, performs evidence-grounded novelty evaluation, and organizes the results into positive, neutral, and negative. In addition, we use novelty type as a structured prior for evidence selection and type-aware comparison. Experiments show that, compared with zero-shot, few-shot, and RAG baselines, our framework better aligns with source novelty claims and covers human reviewer novelty evaluations. Further analyses reveal the roles of evidence grounding, novelty type, and claim-level evaluation.

1 Introduction

Novelty is one of the key criteria for assessing the value and contribution of scholarly publications (Zhao and Zhang, 2025). It is also a central concern in peer review (Alberts et al., 2008), as reviewer guidelines across major venues and journals explicitly ask reviewers to assess a paper's novelty and contribution relative to prior work.[1] However, the rapid growth of academic submissions has placed increasing pressure on the peer review system (Huang et al., 2025b). Novelty assessment is challenging because it requires reviewers to situate a paper's claimed contributions within prior work and distinguish substantial innovations from incremental extensions. Under growing review workloads and increasing disciplinary specialization, such judgments are difficult to make consistently and often result in vague feedback such as "not novel enough (Afzal et al., 2026)."

Recent developments show that the role of LLMs in peer review is being actively debated, tested, and regulated. For example, ICLR 2025[2] piloted an AI feedback agent for reviewers, AAAI-2026 (Biswas et al., 2026) introduced an AI-assisted peer-review pilot with AI-generated review content and discussion summaries, while ICML[3] and PLOS[4] have issued policies emphasizing reviewer responsibility and disclosure when LLMs are used. These efforts indicate that LLMs may provide useful support in review-related workflows, but they should not replace expert judgment.

Among different peer-review tasks, novelty judgment is a particularly important setting for LLM-assisted reviewer support. Rather than making final decisions, such a system should help reviewers situate each novelty claim within relevant prior work and provide evidence that supports more informed novelty judgments. Recent studies have begun to explore this direction from both benchmark and system perspectives. NovBench (Wu

[1] https://aclrollingreview.org/reviewerguidelines/, https://neurips.cc/Conferences/2025/ReviewerGuidelines/, https://www.nature.com/nature-portfolio/editorial-policies/peer-review/

[2] https://blog.iclr.cc/2025/04/15/leveraging-llm-feedback-to-enhance-review-quality/

[3] https://icml.cc/Conferences/2026/LLM-Policy/

[4] https://journals.plos.org/plosone/s/ethical-publishing-practice/

et al., 2026) provides a benchmark for evaluating LLM-generated academic novelty assessments using novelty descriptions in introduction and human reviewer novelty evaluations. OpenNovelty (Zhang et al., 2026) and NoveltyAgent (Hou et al., 2026) further develop evidence-based systems that retrieve real papers and generate structured novelty reports. These systems demonstrate the value of grounding novelty judgments in retrieved literature. However, they either do not explicitly support point-wise novelty evaluation, or they do not consider how novelty types can guide the evaluation process. This limits their ability to model the structured form of reviewer-written novelty evaluations, where reviewers often assess different contribution points separately and express positive, neutral, or negative judgments for different claims. This limitation is especially important because novelty is not a single-dimensional concept. Prior work in the science of science shows that different novelty types, such as new methods, theories, and results, are associated with different forms of scientific influence (Leahey et al., 2023). This suggests that novelty-judgment support systems should not evaluate novelty using a single criterion. A method claim should be compared with prior methods, a dataset claim with existing resources, and a task claim with prior problem formulations. Ignoring such distinctions may lead to inappropriate evidence selection and misleading novelty judgments.

To address these gaps, we propose ClaimNOVA, an evidence-grounded framework for assisting reviewers in claim-level novelty judgment. Rather than replacing reviewers' decisions, our goal is to provide structured support that helps reviewers inspect each novelty claim with relevant background literature. ClaimNOVA preserves novelty claims in introduction, uses novelty type information to select relevant background evidence, performs claim-level novelty judgment, and organizes the resulting evaluations into positive, neutral, and negative. Our contributions are summarized as follows:

- We propose ClaimNOVA, a type-aware and evidence-grounded framework for claim-level novelty judgment. It decomposes novelty assessment into evidence retrieval, claim-level judgment, and structured evaluation.
- We use novelty type as a structured prior for evidence selection and type-aware comparison. This helps compare different claims with more appropriate prior work.
- Experiments on NovBench show that ClaimNOVA improves alignment with source novelty claims and coverage of human reviewer evaluations over zero-shot, few-shot, and RAG baselines. Further analyses reveal the roles of evidence grounding, novelty type, and claim-level evaluation, as well as trade-offs introduced by retrieval-augmented evidence.[5]

2 Related Works

2.1 Automated Scholarly Paper Review

Automated Scholarly Paper Review (ASPR) refers to the use of computational systems to evaluate scholarly papers and generate review reports automatically (Lin et al., 2023). Early studies mainly focused on predicting review scores or recommendation decisions (Kang et al., 2018; Ghosal et al., 2019; Wang et al., 2020), while later work explored automatic review generation by fine-tuning pre-trained models on paper-review pairs (Yuan et al., 2022; Yuan and Liu, 2022).

With the strong generation ability of LLMs (Staudinger et al., 2024; Bao et al., 2025), LLM-assisted peer review has become an active research direction. Recent studies evaluate whether LLMs can generate useful paper reviews (Liang et al., 2024; Zhou et al., 2024; Du et al., 2024), showing that LLMs can provide meaningful feedback but often lack critical depth, specificity, and expert-level insight. Subsequent work has explored prompting strategies, model adaptation, retrieval augmentation, and multi-agent frameworks to improve review generation (Jin et al., 2024; D'Arcy et al., 2024; Zhu et al., 2025; Chang et al., 2025; Idahl and Ahmadi, 2025).

Despite these advances, most existing methods still focus on improving the overall quality of review generation. They often rely on parametric knowledge or abstracts of related papers. This may miss crucial details and limit analytical depth. Moreover, novelty is a central aspect of paper evaluation and requires dedicated assessment rather than being treated as part of general review generation.

2.2 LLM-based Novelty Assessment of Academic Paper

Novelty is a key criterion for assessing the quality and contribution of academic papers. Recent work has explored LLMs for scientific novelty-related

[5] All resources publicly available at https://anonymous.4open.science/r/ClaimNOVA-5B87/.

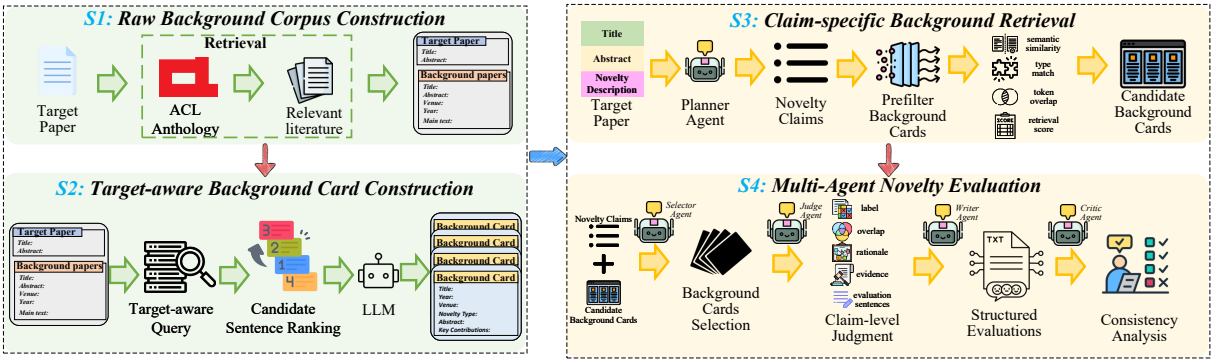

Figure 1: Overview of ClaimNOVA, a four-stage framework that constructs background evidence, retrieves claim-specific candidates, and generates structured novelty evaluations through multi-agent judgment.

tasks, including paper novelty assessment (Huang et al., 2025a; Liu et al., 2025; Lin et al., 2025; Wu et al., 2025a,b; Tan et al., 2026) and research idea generation (Shahid et al., 2025; Su et al., 2025; Kumar et al., 2025; da Silva et al., 2025). However, evaluating LLM-generated novelty assessments is highly challenging because they are usually free-form textual judgments. This makes it difficult to reliably assess semantic adequacy and aspect-specific DistAcc, both of which are crucial for evaluating free-form novelty assessments (Kuznetsov et al., 2024). Recently, NovBench (Wu et al., 2026) addresses this issue by providing a benchmark with novelty descriptions in introduction and expert-written novelty evaluations. Using this setting, we further study how type-aware background evidence can support claim-level novelty judgment.

Since novelty judgment is inherently comparative, recent systems ground novelty assessment in retrieved literature. OpenNovelty generates verifiable novelty reports through paper retrieval and contribution-level comparison (Zhang et al., 2026), while NoveltyAgent uses a multi-agent framework for evidence-based novelty report generation (Hou et al., 2026). Afzal et al. (2026) similarly frame novelty critique as content extraction, related-work retrieval, and evidence-based comparison.

Although these studies improve the transparency of novelty analysis, they do not provide positive, neutral, or negative assessments for different novelty claims, which limits their ability to offer intuitive and structured support. Moreover, they either do not explicitly support point-wise novelty evaluation or do not consider how novelty types can guide the evaluation. Prior work shows that different novelty types, such as new methods, theories, and results, are associated with different forms of scientific influence (Leahey et al., 2023). This suggests that different type of novelty claims should be compared with different kinds of prior work. Ignoring such distinctions may lead to broadly related but poorly comparable evidence. ClaimNOVA addresses this issue by using novelty type as a structured prior for evidence selection and type-aware comparison, then organizing claim-level judgments into structured novelty evaluations.

3 Method

We propose ClaimNOVA, a type-aware and evidence-grounded framework for assisting reviewers in claim-level novelty judgment. Rather than generating a free-form novelty review in one step, ClaimNOVA decomposes novelty evaluation into evidence construction, claim-specific retrieval, evidence-grounded judgment, and structured evaluation organization. It uses novelty type as a structured prior to compare different claims with appropriate prior work, while keeping the final output aligned with the original novelty claims.

Figure 1 shows the four-stage workflow of ClaimNOVA. The framework first constructs a raw background corpus from ACL Anthology papers and then converts relevant prior work into target-aware background cards. Next, it preserves the target paper's novelty claims and retrieves claim-specific candidate evidence based on semantic relevance and novelty-type compatibility. Finally, a multi-agent evaluation module selects evidence, judges each claim, organizes the generated evaluations into positive, neutral, and negative categories, and checks output consistency.

3.1 Task Formulation

We build our study on the EMNLP 2023 subset of NovBench (Wu et al., 2026). Each paper contains novelty descriptions from the introduction and

human reviewer novelty evaluations. We use the novelty descriptions in introduction because they explicitly summarize the contributions that authors claim to be novel. This follows the NovBench setting and does not imply that novelty can only be assessed from the introduction.

Given these novelty descriptions, our task is to generate reviewer-style novelty evaluations. Different from free-form review generation, the output should remain grounded in the original novelty claims while expressing explicit evaluative judgments. This design keeps the evaluation aligned with the paper's novelty claims and reduces semantic drift introduced by unconstrained generation.

Formally, given a target paper p, we denote its novelty claims as $\mathcal{C}_p = \{c_1, c_2, \ldots, c_n\}$, where n is the number of extracted novelty claims. For each claim c_i, our framework retrieves or selects relevant background evidence $\mathcal{B}_i$. Here, relevant background evidence refers to prior-work information that is comparable to the target claim and useful for judging whether the claim is substantially new, moderately new, or mainly incremental. The framework then predicts a novelty polarity label $y_i \in \{\text{POSITIVE}, \text{NEUTRAL}, \text{NEGATIVE}\}$ and generates a claim-grounded evaluation sentence e_i.

The three polarity labels correspond to sentiment toward the novelty of a claim. POSITIVE indicates that the claim is judged as substantially novel. NEUTRAL indicates moderate, partial, or incremental novelty. NEGATIVE indicates limited novelty or strong overlap with prior work. The evaluation sentence e_i is a concise reviewer-style comment that preserves the key content of the original claim while explicitly stating its novelty judgment.

The output is a structured novelty evaluation:

$$\mathcal{E}_p = \{\mathcal{E}_p^+, \mathcal{E}_p^0, \mathcal{E}_p^-\} \tag{1}$$

Where $\mathcal{E}_p^+$, $\mathcal{E}_p^0$, and $\mathcal{E}_p^-$ contain positive, neutral, and negative novelty evaluations, respectively.

3.2 Raw Background Corpus Construction

The first stage of ClaimNOVA constructs a raw background corpus from ACL Anthology[6] papers. Specifically, we collect papers published in three leading NLP conferences, ACL, EMNLP, and NAACL, from 2019 to 2022 as candidate background literature. These papers provide temporally relevant prior work for assessing the novelty of target papers while avoiding overlap with the target evaluation period.

For each background paper, we download the PDF from ACL Anthology when available and parse the full paper text using PyMuPDF[7]. Since most papers follow the ACL-style double-column format, the extracted text may contain noise such as headers, footers, page numbers, references, and appendix content. We therefore apply text cleaning and remove tail sections that are unlikely to provide useful evidence for novelty comparison.

This stage produces a raw background corpus with structured metadata and cleaned paper text. Each background paper includes its title, abstract, venue, year, and main text. The corpus is then used to construct target-aware background cards.

3.3 Target-aware Background Card Construction

Directly using full background papers as evidence is inefficient and may introduce irrelevant information. Therefore, the second stage converts raw background papers into compact target-aware background cards. For each target paper, we use its title, abstract, and novelty descriptions to build a target-aware query. This query is used to rank candidate sentences from each background paper. The top-ranked sentences are then provided to the LLM to construct a compact background card.

Each background card contains structured fields, including title, year, venue, novelty type, summary, and key contribution sentences. The summary captures the main contribution of the background paper. The key contribution sentences provide more fine-grained evidence for later comparison. The novelty type field follows the same NLP-oriented type schema defined in Appendix A.

This card-level representation reduces the context length for later agent reasoning. It also makes background evidence easier to compare through structured contribution information. For each target paper, the cards form the evidence pool for claim-specific retrieval.

3.4 Claim-specific Background Retrieval

The third stage retrieves candidate background evidence for each novelty claim. For each target paper, we use the novelty descriptions in introduction as the source claims. A planner agent preserves and organizes these descriptions into claim sentences. It is instructed to keep the original wording whenever possible, rather than freely rewriting the

[6] https://aclanthology.org/

[7] https://github.com/pymupdf/pymupdf

claims. This reduces semantic drift and keeps later judgments aligned with the novelty claims.

We further introduce novelty type as a structured prior for evidence retrieval and comparison. Prior work suggests that scientific novelty is multidimensional, with different types of novelty associated with different forms of scientific influence (Leahey et al., 2023). Motivated by this, we assign paper-level novelty types from novelty descriptions in introduction and use them to guide type-aware background retrieval. These types are used as auxiliary priors rather than final prediction targets. Detailed type definitions and assignment criteria are provided in Appendix A.

For each novelty claim c_i, ClaimNOVA retrieves a candidate set of background cards from the background pool of the target paper. Candidate cards are ranked using semantic similarity, token overlap, retrieval score, and novelty type matching. Details of these prefiltering signals are provided in Appendix B. The output of this stage is a claim-specific candidate evidence set, which is passed to the multi-agent novelty evaluation stage.

3.5 Multi-Agent Novelty Evaluation

The fourth stage performs a multi-agent novelty evaluation based on the claim-specific candidate evidence retrieved in S3. It consists of selector agent, judge agent, writer agent, and critic agent.

Given a novelty claim c_i and its candidate background cards, the selector agent chooses evidence for novelty judgment. It prefers cards that are directly comparable to the target claim. The selected cards should help judge whether the claim is substantially novel, moderately novel, or mainly incremental. This step produces the background evidence selected $\mathcal{B}_i$ for the judge agent.

Given c_i and $\mathcal{B}_i$, the judge agent performs evidence-grounded novelty assessment. It predicts a polarity label y_i from three categories. A POSITIVE label indicates that the claim is substantially novel relative to the background evidence. A NEUTRAL label indicates moderate or partial novelty. A NEGATIVE label indicates limited novelty or mainly incremental contribution.

In addition to the polarity label, the judge produces auxiliary judgment information, including overlap level with prior work, whether the claim appears incremental, a short rationale, and supporting evidence from the selected background cards. Most importantly, the judge generates a concise evaluation sentence e_i for the claim. This sentence preserves key technical terms from the original claim while expressing an explicit evaluative judgment. More details of the judge output schema and decision criteria are provided in Appendix C.

After all claims have been judged, the writer agent organizes the generated evaluation sentences according to their predicted polarity labels. The final output contains three sections: *Positive Novelty Evaluations*, *Neutral Novelty Evaluations*, and *Negative Novelty Evaluations*. Unlike free-form review generation, the writer does not rewrite the claim-level judgments. Instead, it preserves the judge-generated evaluation sentences and renders them into the required structured format.

The critic agent checks whether important claims are missing, whether the grouped output contradicts point-level judgments, and whether the generated evaluation sentences drift too far from the original novelty descriptions. The critic is used only for consistency analysis and does not modify the final output used in evaluation.

4 Experiments

4.1 Implementation Details

Dataset. We perform experiments mainly on the NovBench (Wu et al., 2026) which contain 1684 EMNLP 2023 samples. Each instance contains a target paper, novelty descriptions in introduction, and human reviewer novelty evaluations. We use them as source claims and reference evaluations, respectively. Final outputs follow the NovBench format: Positive, Neutral, and Negative Novelty Evaluations.

Background Evidence. To support evidence-grounded novelty assessment, we construct a background corpus from ACL Anthology papers published in ACL, EMNLP, and NAACL from 2019 to 2022. For each target paper, we retrieve 50 relevant background papers and convert them into target-aware background cards. Each card includes structured fields such as title, year, venue, novelty type, abstract, and key contribution sentences. Candidate cards are first prefiltered using semantic similarity, token overlap, retrieval score, and novelty type matching. The selector then keeps up to five cards for each novelty claim.

Model and Baselines. We use Qwen3-14B as the backbone model for ClaimNOVA. For a controlled comparison, we compare it with zero-shot, few-shot, and RAG prompting baselines using the same

Method	Relevance	Coverage	Clarity	DistAcc
Zero-shot	3.3892	0.1974	0.6634	0.6627
Few-shot	3.3656	0.2238	0.6487	**0.6734**
RAG	3.2015	0.1673	**0.6681**	0.6245
Ours	**3.5775**	**0.2634**	0.6674	0.5716

Table 1: Main comparison using Qwen3-14B as the backbone model.

Variant	Relevance	Coverage	Clarity	DistAcc
Full	3.5775	**0.2634**	**0.6674**	0.5716
w/o Bg.	3.5996	0.2466	0.6665	0.5722
w/o Sel.	3.5798	0.2588	**0.6674**	**0.5797**
w/o Type	3.5662	0.2599	**0.6674**	0.5402
w/o Eval.	**4.7841**	0.2205	0.6612	0.5739

Table 2: The results of ClaimNOVA ablation studies. "Bg." denotes background evidence, "Sel." denotes background selector, and "Eval." denotes judge-generated evaluation sentences.

backbone. Zero-shot and few-shot baselines generate structured novelty evaluations directly from introduction novelty claims. The RAG baseline additionally uses retrieved background evidence. Full results with additional general-purpose and review-specialized LLMs are reported in Appendix D.

Generation Settings. All agent modules are implemented with vLLM (Kwon et al., 2023) using deterministic decoding. We set the temperature to 0 and require structured JSON outputs for intermediate agent decisions, including background selection, claim-level novelty judgment, and evaluation sentence generation. The final output is rendered into Positive, Neutral, and Negative Novelty Evaluations. The prompts for the selector, judge, writer, and critic agents are provided in Appendix E.

Evaluation Metrics. Following NovBench (Wu et al., 2026), we evaluate all methods with four automatic metrics: Relevance, Coverage, Clarity, and DistAcc. Relevance measures semantic alignment with the novelty claims in the introduction. Coverage measures how well the output covers human reviewer novelty evaluations. Clarity measures readability and surface quality, including conciseness, fluency, and informativeness. DistAcc measures alignment with the polarity distribution of human reviewer evaluations.

4.2 Main result

Table 1 presents a controlled comparison between ClaimNOVA and three prompting baselines. Compared with standard prompting baselines, our method achieves the best Relevance and Coverage. This suggests that claim-level decomposition improves both alignment and coverage. By preserving claims, selecting background evidence, judging each claim, and organizing the outputs, the model better aligns with novelty claims in introduction and covers more novelty-related information.

Our method also achieves Clarity comparable to RAG and higher than few-shot prompting. However, few-shot prompting obtains the highest DistAcc, indicating that polarity distribution calibration remains challenging for our framework. This motivates our subsequent ablation study, background evidence analysis, and type-wise analysis, which further examine how different components affect alignment, coverage, and polarity calibration.

Full prompting baseline results across additional general-purpose and review-specialized LLMs are reported in Appendix D. These results show that large general-purpose models such as GPT-4o and review-specialized models such as SEA-S remain competitive, especially in DistAcc. Since these models differ substantially in scale and specialization, our main analysis focuses on the controlled Qwen3-14B comparison.

4.3 Ablation study

Table 2 reports the ablation results of ClaimNOVA. The full model achieves the highest Coverage, suggesting that the complete pipeline is most effective at preserving and evaluating key information from the original novelty claims. Removing background evidence leads to a clear drop in Coverage. This indicates that external background papers help produce more complete novelty evaluations beyond the target paper's self-description.

The role of novelty type is most evident in DistAcc. Removing novelty type information substantially decreases DistAcc from 0.5716 to 0.5402, showing that type-aware comparison helps calibrate the polarity distribution of novelty judgments. This supports our use of novelty type as an auxiliary structured prior for background matching and claim-level judgment.

The background selector shows a more mixed effect. Removing it slightly improves DistAcc but reduces Coverage compared with the full model. This suggests that the selector mainly contributes to evaluation completeness by selecting more targeted evidence, while its effect on polarity calibration is less consistent. This also indicates that more fine-

Setting	Relevance	Coverage	Clarity	DistAcc
Claim-only	**3.6006**	0.2501	0.6663	0.5713
Retrieved	3.5810	**0.2658**	**0.6676**	0.5791
Selected	3.5743	0.2638	**0.6676**	0.5728
Rand-Ret	3.5581	0.2625	0.6674	**0.6085**

Table 3: Effect of background evidence. "Rand-Ret" denotes randomly sampled background cards from the retrieved candidate pool.

grained evidence selection remains a promising direction for future improvement.

Finally, the variant without judge-generated evaluation sentences obtains the highest Relevance but the lowest Coverage and Clarity. This is expected because this variant is closer to extractive grouping of the original novelty claims, which are also used as the reference for Relevance. However, higher semantic closeness to source claims does not necessarily indicate better novelty evaluation, since the output becomes less evaluative and less complete. Overall, the full model provides the most balanced setting, achieving the best Coverage while maintaining competitive Clarity and DistAcc.

4.4 Background Evidence Analysis

Table 3 shows the effect of different background evidence strategies. Compared with the claim-only setting, all retrieval-augmented settings improve Coverage and Clarity, indicating that external prior work helps the model produce more complete and better-formed novelty evaluations. The retrieved-background setting provides the largest Coverage gain, improving Coverage from 0.2501 to 0.2658, which suggests that directly retrieved evidence is useful for expanding the amount of novelty-related information covered by the output. Interestingly, Rand-Ret achieves the highest DistAcc, increasing it from 0.5713 to 0.6085. This indicates that the initial retrieval pool already contains broadly useful prior work, and even randomly sampled candidates from this pool can help calibrate the polarity distribution of generated evaluations.

At the same time, claim-only achieves the highest Relevance. This is expected because outputs without external evidence tend to stay closer to the original novelty claims in the introduction. By contrast, retrieval-augmented settings encourage the model to generate more evidence-informed evaluations, which can improve completeness but may slightly reduce source-claim similarity. Overall, this result provides an interesting finding: RAG-style background augmentation is beneficial for novelty evaluation, especially in Coverage and DistAcc, but simply adding or selecting evidence does not uniformly improve all metrics. Effective reviewer support therefore requires not only retrieval, but also better claim-level and type-aware control over how retrieved evidence is used.

Type	#	Relevance	Coverage	Clarity	DistAcc
Method	1368	3.5643	0.2587	0.6674	0.5680
Result	1070	3.5542	0.2597	0.6676	0.5748
Dataset	520	3.5669	**0.3176**	0.6676	0.5794
Task	513	3.6068	0.2903	0.6673	0.5774
Application	52	3.6162	0.2045	**0.6684**	**0.5940**
Theory	77	**3.6283**	0.2320	0.6660	0.5738

Table 4: Type-wise performance of the full model.

4.5 Novelty Type-wise Analysis

Table 4 reports the performance of ClaimNOVA across different novelty types. Since novelty type is multi-label, a paper may appear in multiple type-specific subsets. Clarity remains relatively stable across types, suggesting that the generated evaluation sentences are consistently readable.

The main variation appears in Coverage. Dataset- and task-type claims achieve higher Coverage, indicating that novelty claims about newly constructed resources or newly defined tasks are often more explicit and easier to align with the generated evaluations. By contrast, application- and theory-type claims obtain lower Coverage. This may be due to their smaller sample sizes and their more abstract or context-dependent nature.

Application-type claims achieve the highest DistAcc, while method-type claims score lower. Since application and theory contain only 52 and 77 papers, their results should be interpreted cautiously. The lower score for method claims suggests that method novelty is harder to calibrate, as it often requires distinguishing new mechanisms from incremental extensions. Overall, ClaimNOVA performs better on dataset- and task-oriented novelty claims in terms of coverage. In contrast, method-oriented novelty remains harder to calibrate.

4.6 Case Study on Claim-level Novelty Evaluation

Figure 2 presents a case study on academic paper novelty evaluation. The example compares the original novelty claims, the extractive variant without judge-generated evaluation sentences, ClaimNOVA, and the human reviewer novelty evaluation.

The variant without evaluation sentences mostly

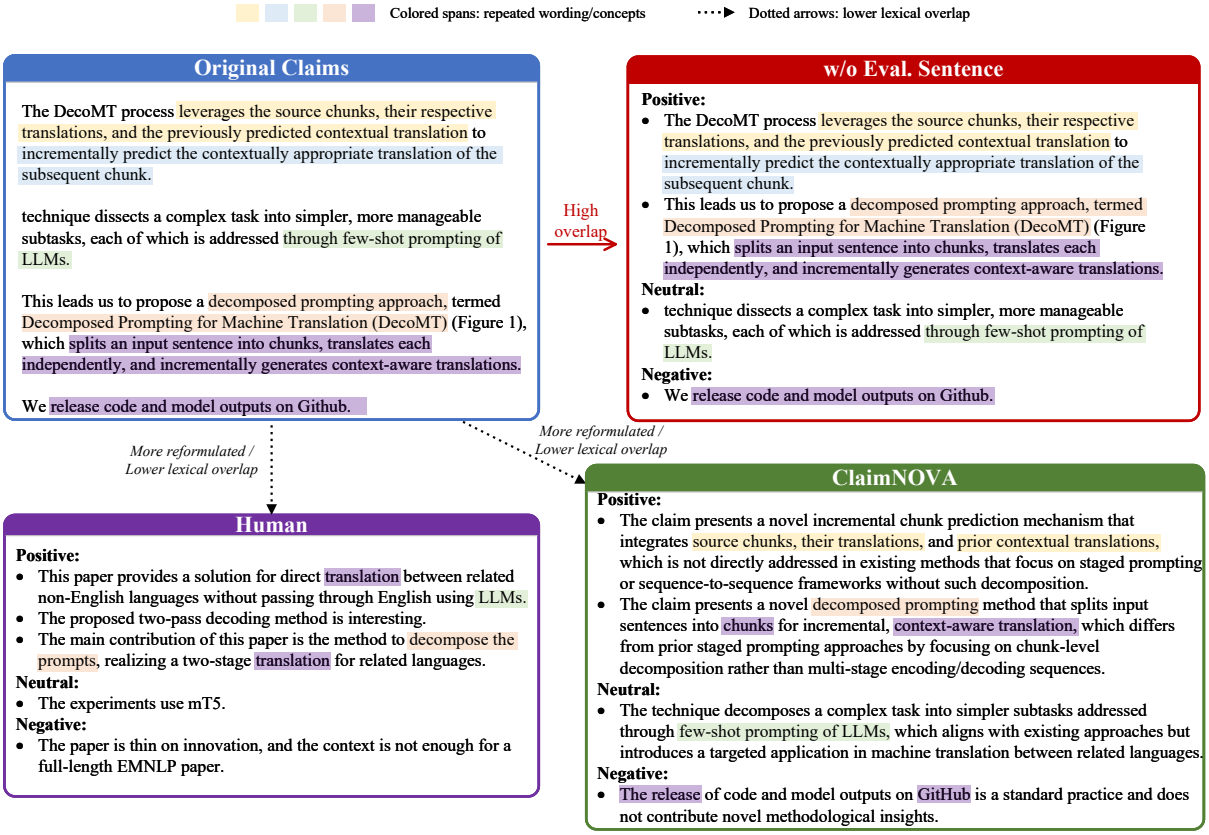

Figure 2: Case study of original claims, w/o Eval. Sentence, ClaimNOVA, and Human novelty evaluation.

copies the original introduction claims and groups them by polarity. This explains its high Relevance, but the output is closer to extractive claim grouping than to explicit novelty evaluation.

In contrast, CLAIMNOVA rewrites the original claims into claim-grounded evaluative comments. It identifies the incremental chunk prediction mechanism and decomposed prompting for machine translation as positive novelty points. It treats the general decomposition of a complex task into few-shot prompted subtasks as moderately novel. This is because the idea aligns with existing prompting practices, but is applied to machine translation in a targeted way. The result is broadly consistent with the human evaluation, which highlights decomposed prompting and two-stage translation as the main contributions.

The case also reveals a limitation. The human reviewer gives an overall negative judgment that the paper is thin on innovation, while ClaimNOVA only captures a local negative point about code release being standard practice. This suggests that claim-level evaluation can support structured novelty assessment, but holistic paper-level novelty limitations remain difficult to capture.

5 Conclusion

We introduced ClaimNOVA, an evidence-grounded claim-level framework for assisting reviewers in novelty judgment. The framework preserves novelty claims in introduction, retrieves relevant background evidence, performs evidence-grounded claim-level judgment, and organizes the resulting evaluations into positive, neutral, and negative. It also incorporates novelty type as a structured prior for type-aware evidence selection and comparison.

Experiments on NovBench show that ClaimNOVA improves Relevance and Coverage over zero-shot, few-shot, and RAG baselines using the same backbone model. Further analyses highlight the roles of background evidence, novelty type, and claim-level evaluation, while revealing challenges in polarity calibration. Our findings suggest that LLM-based systems can provide useful structured support for reviewers, but should be positioned as reviewer-assistance tools rather than replacements for expert judgment.

Limitations

This work has several limitations. First, our experiments focus on the EMNLP 2023 of NovBench,

with background evidence constructed from ACL, EMNLP, and NAACL papers from 2019 to 2022. This provides a controlled NLP setting, but the findings may not directly generalize to other disciplines, venues, or review cultures.

Second, our framework relies on novelty claims in introduction. While this keeps the evaluation aligned with the authors' stated contributions, it may miss novelty-related information appearing in other parts of the paper, such as the method, experiments, or discussion sections. In some cases, the true novelty or limitation of a contribution can only be judged by reading the full paper in detail.

Third, the quality of the generated evaluations depends on retrieval and background card construction. If the retrieval stage misses important prior work, or if background cards fail to capture the most relevant contributions of prior papers, the judge agent may produce incomplete or biased novelty evaluations. Our results also suggest that more fine-grained evidence selection is still needed.

Finally, our evaluation mainly relies on automatic metrics, including Relevance, Coverage, Clarity, and DistAcc. Although these metrics allow scalable comparison across methods, they may not fully capture usefulness for real reviewers. For example, high Relevance may partly reflect extractive similarity to source claims rather than genuine evaluative quality. More extensive human evaluation and interactive reviewer studies are needed to assess whether the generated outputs genuinely support human novelty judgment.

Ethics Statement

ClaimNOVA is intended to assist reviewers, not replace them. Its outputs should be treated as structured, evidence-grounded support rather than final novelty judgments, since generated evaluations may still be incomplete, biased, or incorrect. Automated novelty support may also influence reviewers' perceptions if used without appropriate safeguards. Therefore, such systems should be deployed only in human-in-the-loop settings, with transparent disclosure of AI assistance and final responsibility remaining with human reviewers and editors. Our experiments use existing benchmark data and publicly available scholarly papers, and we do not use the framework to evaluate active submissions or make real decisions.

Acknowledgments

References

Osama Mohammed Afzal, Preslav Nakov, Tom Hope, and Iryna Gurevych. 2026. Beyond "not novel enough": Enriching scholarly critique with LLM-assisted feedback. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2648–2671, Rabat, Morocco. Association for Computational Linguistics.

Bruce Alberts, Brooks Hanson, and Katrina L. Kelner. 2008. Reviewing peer review. *Science*, 321(5885):15–15.

Tong Bao, Mir Tafseer Nayeem, Davood Rafiei, and Chengzhi Zhang. 2025. SurveyGen: Quality-aware scientific survey generation with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2712–2736, Suzhou, China. Association for Computational Linguistics.

Joydeep Biswas, Sheila Schoepp, Gautham Vasan, Anthony Opipari, Arthur Zhang, Zichao Hu, Sebastian Joseph, Matthew Lease, Junyi Jessy Li, Peter Stone, Kiri L. Wagstaff, Matthew E. Taylor, and Odest Chadwicke Jenkins. 2026. Ai-assisted peer review at scale: The aaai-26 ai review pilot. *Preprint*, arXiv:2604.13940.

Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. 2025. TreeReview: A dynamic tree of questions framework for deep and efficient LLM-based scientific peer review. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15662–15693, Suzhou, China. Association for Computational Linguistics.

Italo Luis da Silva, Hanqi Yan, Lin Gui, and Yulan He. 2025. GraphMind: Interactive novelty assessment system for accelerating scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 286–294, Suzhou, China. Association for Computational Linguistics.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. Marg: Multi-agent review generation for scientific papers. *Preprint*, arXiv:2401.04259.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. 2024. Reviewer2: Optimizing review generation through prompt generation. *Preprint*, arXiv:2402.10886.

Google Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Tirthankar Ghosal, Rajeev Verma, Asif Ekbal, and Pushpak Bhattacharyya. 2019. DeepSentiPeer: Harnessing sentiment in review texts to recommend peer review decisions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1120–1130, Florence, Italy. Association for Computational Linguistics.

Jiajun Hou, Hexuan Deng, Wenxiang Jiao, Xuebo Liu, Xiaopeng Ke, and Min Zhang. 2026. Noveltyagent: Autonomous novelty reporting agent with pointwise novelty analysis and self-validation. *Preprint*, arXiv:2603.20884.

Shengzhi Huang, Yong Huang, Yinpeng Liu, Zhuoran Luo, and Wei Lu. 2025a. Are large language models qualified reviewers in originality evaluation? *Information Processing & Management*, 62(3):103973.

Xuanjing Huang, Shihan Dou, and Zhangyue Yin. 2025b. The dual-edged sword: artificial intelligence's evolving role in academic peer review. *Sci. China Inf. Sci.*, 68(11).

Maximilian Idahl and Zahra Ahmadi. 2025. OpenReviewer: A specialized large language model for generating critical scientific paper reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 550–562, Albuquerque, New Mexico. Association for Computational Linguistics.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. AgentReview: Exploring peer review dynamics with LLM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1208–1226, Miami, Florida, USA. Association for Computational Linguistics.

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. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Sandeep Kumar, Tirthankar Ghosal, Vinayak Goyal, and Asif Ekbal. 2025. Can large language models unlock novel scientific research ideas? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33551–33575, Suzhou, China. Association for Computational Linguistics.

Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K. Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, Sheng Lu, Mausam, Margot Mieskes, Aurélie Névéol, Danish Pruthi, Lizhen Qu, Roy Schwartz, Noah A. Smith, Thamar Solorio, and 5 others. 2024. What can natural language processing do for peer review? *Preprint*, arXiv:2405.06563.

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

Erin Leahey, Jina Lee, and Russell J. Funk. 2023. What types of novelty are most disruptive? *American Sociological Review*, 88(3):562–597.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. 2024. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196.

Ethan Lin, Zhiyuan Peng, and Yi Fang. 2025. Evaluating and enhancing large language models for novelty assessment in scholarly publications. In *Proceedings of the 1st Workshop on AI and Scientific Discovery: Directions and Opportunities*, pages 46–57, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

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

Yan Liu, Zonglin Yang, Soujanya Poria, Thanh-Son Nguyen, and Erik Cambria. 2025. Harnessing large language models for scientific novelty detection. *Preprint*, arXiv:2505.24615.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf.

Simra Shahid, Marissa Radensky, Raymond Fok, Pao Siangliulue, Daniel S Weld, and Tom Hope. 2025. Literature-grounded novelty assessment of scientific ideas. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 96–113, Vienna, Austria. Association for Computational Linguistics.

Moritz Staudinger, Wojciech Kusa, Florina Piroi, and Allan Hanbury. 2024. An analysis of tasks and datasets in peer reviewing. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 257–268, Bangkok, Thailand. Association for Computational Linguistics.

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, Philip Torr, Bowen Zhou, and Nanqing Dong. 2025. Many heads are better than one: Improved scientific idea generation by a LLM-based multi-agent system. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28201–28240, Vienna, Austria. Association for Computational Linguistics.

Hongming Tan, Shaoxiong Zhan, Fengwei Jia, Hai-Tao Zheng, and Wai Kin (Victor) Chan. 2026. A hierarchical framework for measuring scientific paper innovation via large language models. *Information Sciences*, 728:122787.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. ReviewRobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 384–397, Dublin, Ireland. Association for Computational Linguistics.

Wenqing Wu, Chengzhi Zhang, Tong Bao, and Yi Zhao. 2025a. Sc4anm: Identifying optimal section combinations for automated novelty prediction in academic papers. *Expert Systems with Applications*, 273:126778.

Wenqing Wu, Chengzhi Zhang, and Yi Zhao. 2025b. Automated novelty evaluation of academic paper: A collaborative approach integrating human and large language model knowledge. *Journal of the Association for Information Science and Technology*, 76(11):1452–1469.

Wenqing Wu, Yi Zhao, Yuzhuo Wang, Siyou Li, Juexi Shao, Yunfei Long, and Chengzhi Zhang. 2026. Novbench: Evaluating large language models on academic paper novelty assessment. *Preprint*, arXiv:2604.11543.

Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, RenJing Cui, Chengcheng Han, Qiushi Sun, Zhiyong Wu, Yunshi Lan, and Xiang Li. 2024. Automated peer reviewing in paper SEA: Standardization, evaluation, and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10164–10184, Miami, Florida, USA. Association for Computational Linguistics.

Weizhe Yuan and Pengfei Liu. 2022. Kid-review: Knowledge-guided scientific review generation with oracle pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11639–11647.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212.

Ming Zhang, Kexin Tan, Yueyuan Huang, Yujiong Shen, Chunchun Ma, Li Ju, Xinran Zhang, Yuhui Wang, Wenqing Jing, Jingyi Deng, Huayu Sha, Binze Hu, Jingqi Tong, Changhao Jiang, Yage Geng, Yuankai Ying, Yue Zhang, Zhangyue Yin, Zhiheng Xi, and 4 others. 2026. Opennovelty: An llm-powered agentic system for verifiable scholarly novelty assessment. *Preprint*, arXiv:2601.01576.

Yi Zhao and Chengzhi Zhang. 2025. A review on the novelty measurements of academic papers. *Scientometrics*, 130(2):727–753.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, Torino, Italia. ELRA and ICCL.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. DeepReview: Improving LLM-based paper review with human-like deep thinking process. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29330–29355, Vienna, Austria. Association for Computational Linguistics.

A Novelty Type Construction and Definitions

We construct novelty type annotations in two steps. First, for each target paper, we provide GPT-4o with the paper's novelty descriptions and reviewer-written novelty evaluations, and ask it to generate preliminary novelty type descriptions in an open-ended manner. The prompt used in this step is shown in Figure 3. We then manually inspect and summarize the generated type descriptions. Based

Type	Definition
Method	New model, algorithm, framework, module, architecture, training strategy, or inference procedure.
Dataset	New dataset, corpus, benchmark, annotation scheme, data resource, or data collection process.
Task	New task formulation, problem setting, evaluation setting, or research objective.
Application	Applying an approach to a new domain, scenario, user group, language, or real-world use case.
Theory	New theoretical analysis, conceptual explanation, formal insight, or analytical perspective.
Result	New empirical finding, performance result, experimental observation, or analysis outcome.

Table 5: Operational definitions of the novelty types used in this work.

on this analysis, we consolidate the novelty types into six categories: *method*, *result*, *theory*, *application*, *task*, and *dataset*.

Second, using the consolidated type schema, we design another GPT-4o prompt to assign novelty type labels to the novelty descriptions of each target paper. The prompt for type annotation is shown in Figure 4. A paper may be associated with multiple novelty types. The resulting type labels are used as auxiliary priors for evidence selection and type-aware comparison in ClaimNOVA, rather than as final prediction targets.

We use this taxonomy as a practical schema for NLP papers rather than a universal taxonomy for all scientific domains. For example, claims about new architectures or algorithms are assigned to *method*; claims about newly constructed resources are assigned to *dataset*; claims about new problem formulations are assigned to *task*; and claims about empirical findings or performance improvements are assigned to *result*.

B Details of Type-aware Background Evidence Selection

For each novelty claim c_i, ClaimNOVA first applies an automatic prefiltering step to reduce the target paper's background card pool to a manageable set of candidate evidence cards. Each background card is scored with respect to the target claim using four complementary signals: semantic similarity, token overlap, retrieval score, and novelty type matching.

Semantic similarity. We compute semantic similarity between the target claim and each background card using embedding cosine similarity. Let $g(\cdot)$ denote the text encoder, and let x_j be the textual representation of background card b_j, obtained by concatenating its title, summary, and key contribution sentences. The semantic similarity score is computed as:

$$\mathrm{Sim}(c_i, b_j) = \frac{g(c_i)^\top g(x_j)}{\|g(c_i)\|\|g(x_j)\|}. \quad (2)$$

This signal captures whether the background paper discusses a similar research problem, method, resource, or empirical finding.

Token overlap. We compute lexical overlap between the target claim and the background card after lowercasing, removing punctuation, and applying basic normalization. Let $V(c_i)$ and $V(b_j)$ denote the normalized token sets of the claim and background card, respectively. We use a claim-centric overlap score:

$$\mathrm{Tok}(c_i, b_j) = \frac{|V(c_i) \cap V(b_j)|}{|V(c_i)|}. \quad (3)$$

This score measures how many important terms in the target claim are covered by the background card. It helps retain cards that share technical terms, such as method names, dataset names, task names, or domain-specific expressions.

Identify the broad novelty type(s) of the given paper based on:
(1) novelty-related descriptions extracted from the introduction, and
(2) reviewer comments about the paper's novelty.

The two sources refer to the same paper.

Instructions:
- Infer the novelty type(s) from each source separately, then give the final novelty type(s).
- Use broad, high-level, and reusable type names only.
- Do not use overly specific or paper-specific descriptions.
- Each type name should be short, ideally 1–3 words.
- Do not include sentiment or quality judgments.
- Return the output as a json dictionary.

Output format:
{{
"intro_novelty_type": ["..."],
"review_novelty_type": ["..."],
"final_novelty_type": ["..."]
}}

Introduction novelty descriptions:
{intro_novelty_text}

Reviewer novelty comments:
{review_novelty_text}

Figure 3: Prompt for open-ended novelty type discovery.

You are an expert NLP researcher.

Your task is to identify the MAIN types of novelty in a paper.

IMPORTANT REQUIREMENTS:

You MUST choose types ONLY from the following list:
["method", "dataset", "task", "application", "theory", "result"]
Each output type must be:
High-level (NOT too specific)
Conceptually distinct
Output 1–3 types ONLY.
If multiple types apply, include all major ones, but do NOT exceed 3.
Do NOT generate new labels or synonyms.
For example, use "method" instead of "model", "framework", or "approach"
Use "dataset" instead of "data" or "corpus"
If unsure, prefer:
"method" for modeling or algorithmic contributions
"dataset" for data construction or resources
"task" for new problem formulations
"application" for real-world or domain-specific usage
"theory" for analytical or explanatory contributions
"result" for performance improvements or empirical findings
Do NOT provide any explanation.

OUTPUT FORMAT (strict JSON):
{{
"novelty_types": ["type1", "type2"]
}}

Introduction novelty descriptions:
{intro_novelty_text}

Figure 4: Prompt for assigning novelty type labels to target papers.

Retrieval score. We also use the initial retrieval score obtained when constructing the target paper's background pool. Let $r(p, b_j)$ denote the original retrieval score between target paper p and background paper b_j. We normalize this score within the background pool of p:

$$\mathrm{Ret}(b_j) = \frac{r(p, b_j) - r_{\min}}{r_{\max} - r_{\min} + \epsilon}, \tag{4}$$

Where

$$r_{\min} = \min_{b_k \in \mathcal{B}_p} r(p, b_k), r_{\max} = \max_{b_k \in \mathcal{B}_p} r(p, b_k) \tag{5}$$

Here, $\mathcal{B}_p$ is the background pool of target paper p, and ϵ is a small constant used to avoid division by zero. This score reflects the overall relevance of the background paper to the target paper before claim-level evidence selection.

Novelty type matching. We incorporate novelty type matching as a structured prior. A background card receives a type-matching signal when its novelty type overlaps with the novelty type of the target claim. This helps compare different claims under appropriate dimensions. For example, method-type claims are compared with prior methodological contributions, dataset-type claims with existing resources, task-type claims with prior task formulations, and result-type claims with related empirical findings.

After computing these signals, we rank background cards and keep the top candidates for agent-based selection. In our implementation, we keep the top 12 cards after prefiltering, and the selector finally retains up to 5 cards for each claim. Unlike prefiltering, which relies on lightweight similarity and type-matching signals, the selector agent performs a more contextual judgment. It is instructed to choose evidence cards that are directly comparable to the target claim and useful for determining whether the claim is substantially novel, moderately novel, or mainly incremental.

C Details of Evidence-grounded Claim-level Novelty Judgment

The judge agent is responsible for converting each novelty claim into an evidence-grounded novelty judgment. For each claim c_i, the judge receives three types of input: the original novelty claim, its novelty type, and the selected background evidence $\mathcal{B}_i$. The selected evidence consists of background cards chosen by the selector agent. Each card contains structured information such as title, year, venue, novelty type, summary, and key contribution sentences.

Input to the judge. The input to the judge contains:

- **Novelty claim:** the original novelty claim in introduction to be evaluated;
- **Novelty type:** the contribution type of the claim, such as method, dataset, task, application, theory, or result;
- **Selected background evidence:** a small set of background cards that are relevant to the claim.

Polarity label definitions. The judge predicts one of three polarity labels:

- POSITIVE: the claim appears substantially novel relative to the selected background evidence. This label is used when the claim introduces a clearly new method, dataset, task

You are an expert NLP researcher helping construct background evidence for academic novelty evaluation.

You will be given:
- a target paper with title, abstract, introduction-level novelty claims, and novelty types;
- a background paper with title, abstract, venue, year, and parsed main text.

Your task is to create a concise background card that captures the parts of the background paper most relevant to evaluating the novelty of the target paper.

Focus on:
- the main research problem or task addressed by the background paper;
- the key method, dataset, task, application, theory, or result contribution;
- whether the background paper is comparable to the target paper;
- contribution sentences that may help judge whether the target paper's novelty claims are new, moderately new, or incremental.

Do not summarize unrelated implementation details, experimental settings, or general background unless they are directly useful for novelty comparison.

Output strictly in JSON:

```
{
 "title": "...",
 "year": "...",
 "venue": "...",
 "novelty_type": ["method", "dataset", "task", "application", "theory",
"result"],
 "summary": "...",
 "key_contributions": [
  "...",
  "..."
 ],
 "relevance_to_target": "..."
}
```

Figure 5: Prompt template for constructing target-aware background cards.

formulation, application, theoretical perspective, or empirical finding that is not directly covered by the background evidence.

- NEUTRAL: the claim shows moderate, partial, or incremental novelty. This label is used when the claim extends existing work, combines known ideas in a somewhat new way, or presents a contribution that is useful but not strongly distinct from prior work.

- NEGATIVE: the claim appears to have limited novelty or is mostly covered by prior work. This label is used when the selected background evidence already contains highly similar ideas, methods, resources, tasks, or findings.

Auxiliary judgment fields. In addition to the polarity label, the judge produces several auxiliary fields to make the judgment more traceable:

- **Overlap level:** whether the claim has low, medium, or high overlap with the selected background evidence;

- **Incrementality:** whether the claim mainly represents an incremental extension of prior work;

- **Rationale:** a short explanation of why the claim receives the predicted polarity label;

- **Supporting evidence:** relevant evidence snippets or summaries from the selected background cards;

- **Evaluation sentence:** a concise sentence that expresses the novelty judgment for the claim.

Evaluation sentence generation. The evaluation sentence is the main content passed to the writer agent. It is required to satisfy three constraints. First, it should preserve key technical terms from the original claim, such as method names, dataset names, task names, or core mechanisms. Second, it should be explicitly evaluative rather than merely descriptive. For example, instead of repeating that a paper "proposes a new module," the sentence should indicate whether the module appears substantially novel, moderately novel, or limited in novelty relative to prior work. Third, it should avoid introducing unsupported entities or claims that do not appear in either the original claim or the selected background evidence.

Output schema. The judge output follows a structured JSON schema:

```
{
  "claim_id": "...",
  "claim": "...",
  "novelty_type": ["method"],
 "label": "positive | neutral | negative",
  "overlap_level": "low | medium | high",
  "is_incremental": true,
```

You are an expert peer reviewer specializing in evidence-based novelty evaluation.

You will be given:
- a target paper.
- one novelty claim from the target paper's introduction.
- the novelty type of the claim.
- a list of candidate background cards.

Your task is to select the background cards that are most useful for judging the novelty of the given claim.

Selection criteria:
- Prefer background cards that are directly comparable to the target claim.
- Prefer evidence that addresses the same contribution type, such as method, dataset, task, application, theory, or result.
- Select cards that can help determine whether the claim is substantially novel, moderately novel, or mainly incremental.
- Do not select cards only because they share broad topic words if they are not useful for novelty comparison.
- Select at most 5 background cards.

Output strictly in JSON:
```
{
 "claim_id": "...",
 "claim": "...",
 "selected_background_cards": [
   {
    "card_id": "...",
    "reason": "Why this background card is useful for evaluating the claim."
   }
 ]
}
```

Figure 6: Prompt template for the selector agent.

```
  "rationale": "...",
  "supporting_evidence": [
    {
      "card_id": "...",
      "evidence": "..."
    }
  ],
  "evaluation_sentence": "..."
}
```

Role in the full pipeline. The judge does not generate the final structured review directly. Instead, it produces claim-level judgments and evaluation sentences. The writer agent then groups these sentences into Positive, Neutral, and Negative Novelty Evaluations according to the predicted labels. This separation helps keep the final output aligned with claim-level evidence while avoiding uncontrolled free-form review generation. The prompt template used by the judge agent is shown in Figure 7.

D Full Baseline Results

Table 6 reports the full results of all prompting baselines under zero-shot, few-shot, and RAG settings. We include both general-purpose LLMs and review-specialized LLMs to provide a broader comparison of model behavior on novelty evaluation. For each setting, models are evaluated using the same four automatic metrics: Relevance, Coverage, Clarity, and DistAcc.

The results show that general-purpose LLMs, such as GPT-4o (OpenAI et al., 2024), GPt-5 (OpenAI, 2025), Gemini-2.5-flash (Gemini Team, 2025), gpt-oss (OpenAI et al., 2025) and DeepSeek-R1 (DeepSeek-AI et al., 2025), generally achieve strong performance across prompting settings, while review-specialized models such as SEA-S (Yu et al., 2024), Reviewer2 (Gao et al., 2024) and SEA-E are also competitive, especially in DistAcc and Coverage. However, model size, training objective, and specialization differ substantially across these baselines. Therefore, in the main text, we focus on the controlled comparison using Qwen3-14B as the shared backbone, while this appendix provides the complete baseline results for reference.

E Agent Prompt Templates

To support reproducibility, we provide the prompt templates used by each agent in ClaimNOVA, as shown in Figures 5,6,7,8,9. Each prompt specifies the agent role, input information, task instructions, and expected output format.

Model	Zero-shot				Few-shot				RAG			
	Rel.	Cov.	Clar.	DistAcc	Rel.	Cov.	Clar.	DistAcc	Rel.	Cov.	Clar.	DistAcc
General LLMs												
DeepSeek-R1-70B	3.4885	0.2074	0.6470	0.6572	3.4452	0.2112	0.6455	0.6274	3.0376	0.1500	0.6626	0.6260
DeepSeek-R1-14B	3.4058	0.2053	0.6404	0.6436	3.3844	0.2252	0.6483	0.6606	3.0715	0.1713	0.6628	0.6417
DeepSeek-R1-8B	2.9475	0.1603	0.4949	0.5143	3.4703	0.2190	0.6160	0.6517	2.4988	0.1284	0.5459	0.5220
Qwen3-32B	3.4747	0.2065	0.6609	0.6555	3.4175	0.2242	0.6497	0.6944	3.2555	0.1692	0.6679	0.6604
Qwen3-14B	3.3892	0.1974	**0.6634**	0.6627	3.3656	0.2238	0.6487	0.6734	3.2015	0.1673	0.6681	0.6245
Qwen3-8B	3.5769	0.1996	0.6471	0.6595	3.4061	0.2240	0.6487	0.6784	3.3042	0.1732	0.6593	0.6737
GPT-4o	**3.6983**	**0.2332**	0.6595	**0.6979**	**3.5609**	**0.2391**	0.6587	**0.7091**	3.4481	0.2237	0.6668	**0.6965**
GPT-5	3.2772	0.1591	0.6209	0.4830	3.3124	0.1806	0.6164	0.5411	3.2300	0.1673	0.6666	0.6453
gpt-oss-120b	3.2586	0.1787	0.6535	0.4376	3.1897	0.1830	**0.6613**	0.5107	3.1424	0.1644	0.6656	0.6027
gpt-oss-20b	3.3158	0.1843	0.6270	0.5785	3.3298	0.2098	0.6364	0.6461	3.1372	0.1676	0.6269	0.6430
Gemini-2.5-flash	3.4711	0.2120	0.6414	0.6011	3.4726	0.2364	0.6573	0.6590	**3.5089**	**0.2270**	**0.6682**	0.5923
Specialized LLMs												
CycleReviewer-70B	3.4632	0.2198	0.6598	0.6326	3.3426	0.2209	0.6494	0.6522	3.0292	0.1522	0.6587	0.4892
CycleReviewer-8B	3.0712	0.1577	0.6377	0.3837	3.0833	0.1892	0.6206	0.4088	2.8853	0.1336	0.5817	0.2785
DeepReviewer-14B	2.7402	0.1173	0.6425	0.6356	2.7161	0.1140	0.6130	0.6301	2.7288	0.1193	0.6161	0.5246
DeepReviewer-7B	2.4654	0.0745	0.5992	0.6134	2.6589	0.1051	0.6160	0.6308	2.5465	0.1004	0.5697	0.3907
Llama-OpenReviewer-8B	2.1293	0.0604	0.3339	0.2216	2.3047	0.0823	0.5381	0.6205	1.2235	0.0317	0.2749	0.1031
Reviewer2	1.8377	0.0408	0.4496	0.4517	0.9993	0.0013	0.3402	0.4083	0.1556	0.0000	0.0184	0.0709
SEA-E	3.4259	0.2610	0.6497	0.6834	3.3356	0.2483	0.6395	0.6609	3.3807	0.2712	0.6585	0.5965
SEA-S	3.6304	0.2576	0.6630	0.7162	3.4091	0.2454	0.6519	0.7149	3.5170	0.2474	0.6662	0.6740
Human	2.7899	–	–	–	2.7899	–	–	–	2.7899	–	–	–

Table 6: Full baseline results under zero-shot, few-shot, and RAG prompting strategies. For each metric and prompting setting, the best-performing general LLM is highlighted in bold, and the best-performing specialized LLM is underlined. Rel., Cov., Clar., and DistAcc denote Relevance, Coverage, Clarity, and DistAcc, respectively. Human refers to human reviewer novelty comments evaluated under the same automatic metrics; "–" indicates not applicable.

You are a peer-review expert specializing in academic novelty evaluation.

You will be given:
- one introduction-level novelty claim from a target paper.
- the novelty type of the claim.
- selected background evidence from related prior papers.

Your task is to select the background cards that are most useful for judging the novelty of the given claim.

Your task is to evaluate the novelty of the claim with respect to the provided background evidence.

Use the following label definitions:
- **positive**: the claim appears substantially novel relative to the background evidence.
- **neutral**: the claim shows moderate, partial, or incremental novelty.
- **negative**: the claim appears to have limited novelty or is mostly covered by prior work.

Important requirements:
- Keep the judgment grounded in the provided background evidence.
- Preserve key technical terms from the original claim.
- Do not simply copy the original claim.
- Generate one concise evaluation sentence that is close to the claim but explicitly evaluative.
- Avoid introducing entities or claims that are not supported by the given evidence.

Output strictly in JSON:
```
{
 "claim_id": "...",
 "claim": "...",
 "novelty_type": ["..."],
 "label": "positive | neutral | negative",
 "overlap_level": "low | medium | high",
 "is_incremental": true,
 "rationale": "Brief explanation of the novelty judgment.",
 "supporting_evidence": [
   {
    "card_id": "...",
    "evidence": "Relevant evidence from the background card."
   }
 ],
 "evaluation_sentence": "A concise claim-grounded novelty evaluation sentence."
}
```

Figure 7: Prompt template for the judge agent.

You are a peer-review writing assistant.

You will be given a list of claim-level novelty judgments.
Each judgment contains:
- the original novelty claim.
- a predicted novelty label.
- a judge-generated evaluation sentence.

Your task is to organize the judge-generated evaluation sentences into the following three sections:

Positive Novelty Evaluations
-

Neutral Novelty Evaluations
-

Negative Novelty Evaluations
-

Rules:
- Use the judge-generated evaluation sentences as the main content.
- Do not copy original novelty claims unless they are already part of the judge-generated evaluation sentence.
- Do not introduce new judgments not supported by the judge outputs.
- Do not change the polarity label assigned by the judge.
- If a category has no evaluation, output only a dash under that category.
- Keep the output concise and academic in tone.

Output strictly in JSON:
```
{
  "positive_novelty_evaluations": [
    "..."
  ],
  "neutral_novelty_evaluations": [
    "..."
  ],
  "negative_novelty_evaluations": [
    "..."
  ],
  "final_review_text": "Positive Novelty Evaluations\n- ...\n\nNeutral Novelty Evaluations\n- ...\n\nNegative Novelty Evaluations\n- ..."
}
```

Figure 8: Prompt template for the writer agent.

You are a critical peer-review analyst.

You will be given:
- the original introduction-level novelty claims.
- the claim-level judge outputs.
- the final grouped novelty evaluations.

Your task is to analyze whether the final grouped evaluations are consistent with the claim-level judgments.

Check the following:
- whether important novelty claims are missing from the final output.
- whether any evaluation sentence is assigned to the wrong polarity category.
- whether any sentence drifts too far from the original novelty claim.
- whether the final output is overly extractive or overly abstract.
- whether the generated evaluations miss paper-level novelty limitations.

Do not rewrite the final output.
Only provide diagnostic analysis.

Output strictly in JSON:
```
{
  "missing_claims": [
    "..."
  ],
  "polarity_inconsistencies": [
    {
      "sentence": "...",
      "issue": "..."
    }
  ],
  "semantic_drift_issues": [
    {
      "sentence": "...",
      "issue": "..."
    }
  ],
  "overall_comments": "...",
  "does_not_modify_final_output": true
}
```

Figure 9: Prompt template for the critic agent.