What Is Novel? A Knowledge-Driven Framework for Bias-Aware Literature Originality Evaluation

Anonymous ACL submission

Abstract

Assessing research novelty is a core yet highly subjective aspect of peer review, typically based on implicit judgment and incomplete comparison to prior work. We introduce a literature-aware novelty assessment framework that explicitly learns how humans judge novelty from peer-review reports and grounds these judgments in structured comparison to existing research. Using nearly 80K novelty-annotated reviews from top-tier AI conferences, we fine-tune a large language model to capture reviewer-aligned novelty evaluation behavior. For a given manuscript, the system extracts structured representations of its ideas, methods, and claims, retrieves semantically related papers, and constructs a similarity graph that enables fine-grained, concept-level comparison to prior work. Conditioning on this structured evidence, the model produces calibrated novelty scores and human-like explanatory assessments, reducing overestimation and improving consistency relative to existing approaches.

1 Introduction and Related Work

Novelty plays a crucial role in peer review, influencing acceptance decisions and shaping research directions, yet it is among the most subjective and difficult criteria to assess reliably. Novelty evaluation of a scientific submission requires reviewers to implicitly compare it against previously published works under limited time and information constraints. As publication volume rapidly increases (Künzli et al., 2022) and reviewer workload intensifies, forming consistent and well-grounded novelty judgments becomes increasingly challenging. As a result, assessments vary across reviewers, and depend on incomplete coverage of related literature (Lin et al., 2023; Sizo et al., 2025).

Recently, large language models (LLMs) have shown promise in automating aspects in peer review (Dycke et al., 2023; Kuznetsov et al., 2024; Yu et al., 2024). Existing automatic review-generation systems can produce fluent, human-like reviews, yet their feedback is often overly positive and vague, lacking detailed critique and explicit reasoning about novelty (Du et al., 2024; Li et al., 2025; Idahl and Ahmadi, 2025). In particular, these systems do not generate structured novelty assessments grounded in systematic comparison to prior work. In parallel, prior approaches to novelty detection typically rely on embedding-based similarity (Shibayama et al., 2021; Shahid et al., 2025) or classification signals (Yan et al., 2025; Zhao and Zhang, 2025). While effective at identifying surface-level similarity, these methods do not produce human-like critique and explanations that reflect how novelty is assessed in practice, nor do they explicitly ground judgments in interpretable comparisons to related work.

This paper focuses on a reviewer-style novelty assessment approach that can generate human-style comments discussing originality, while explicitly grounding judgments in comparisons to relevant prior work. We argue that achieving this requires two key ingredients that are rarely integrated in existing approaches. First, models must be trained on supervision that reflects human novelty judgments as expressed in peer reviews. Second, novelty judgements must be conditioned on explicit evidence from related literature, enabling concept-level comparison rather opaque similarity scores.

To this end, we propose a automatic literature-aware framework for automated novelty assessment. We curate a large-scale dataset of nearly 80K review reports, extract and aggregate reviewer discussions of novelty into paper-level novelty assessments and normalized novelty scores. Then, critique-generation model is trained to produce structured novelty scores and textual assessments aligned with human reviewing behavior. To ground these judgments in concrete prior work, we introduce a graph-based retrieval module that identifies topically related papers based on semantic similarity across ideas, methods, and contributions, enabling explicit concept-level comparison to further refine the generated assessments. Our resulting novelty assessment helps the reviewer to highlight both original contributions and areas of overlap with existing research.

2 Benchmark Construction

Data collection. We follow the same crawling and data processing procedure as in prior work (Idahl and Ahmadi, 2025) to collect $79,973$ high-confidence review reports of $32,439$ submissions to top-tier AI conferences (NeurIPS and ICLR) on OpenReview[1] from 2022 onwards.

Novelty-related text extraction and aggregation. For each review, we use an instruction-tuned LLM to extract textual segments that discuss novelty, either explicitly or implicitly through positive or negative assessment of ideas and contributions. Next, the LLM is prompted to aggregate the extracted novelty-related content across all reviews for a submission (Fig. 5, Appendix C). This process generates a single concise, paper-centric originality assessment for a given paper, based solely on human reviews. We consider three aggregation scenarios:

1. **Consensus:** all reviewers agree on the novelty assessment ($17,946$ papers). In this case, the model summarizes the shared opinion, incorporating all extracted novelty-related statements.
2. **Majority agreement:** reviewers express conflicting assessments, but a majority supports one stance ($8,891$ papers). Here, the model adopts the majority opinion while explicitly reflecting the overall novelty discussion.
3. **Tie:** an equal number of reviewers express positive and negative novelty assessments ($5,602$ papers). In this scenario, the model examines the strength of evidence and arguments provided by the reviewers; if no strong justification is present, it assigns a marginal novelty rating.

Beside the textual assessment, the model assigns a discrete, normalized novelty score in the range $[-1, 2]$, corresponding to not novel, limited novelty, moderate novelty, and high novelty. The labels are derived from the overall stance expressed in the reviews rather than from external annotations.

In this study, we use Llama 3.1-8B-Instruct (Meta, 2024) as a deterministic aggregation and rewriting component. In preliminary experiments, we observed that alternative prompt formulations produced qualitatively similar aggregated novelty assessments.

Quality check. To ensure reliability, we conduct basic a quality check to verify that the aggregated assessments preserve the core novelty-related content expressed by reviewers. Specifically, we conduct tests on 500 papers, compare sentence-level embeddings (Reimers and Gurevych, 2019) of the aggregated novelty assessments with those of the original human reviews. The average cosine similarity score is 0.78, indicating substantial semantic overlap between the aggregated assessments and the original reviews.

3 Methodology

Our system operates in two stages: Firstly, learning human novelty-evaluation pattern from peer-review dataset via supervision, and secondly, applying this learned signal to new manuscripts along with using retrieval-enhanced reasoning and structured comparison against related prior work. The full abstract pipeline is illustrated in Figure 1.

Structured Knowledge Extraction: Given a new manuscript, we first pass its full text to a dedicated knowledge-extraction agent. This module uses Llama-3.1-8B-Instruct model to transform unstructured scientific text into a structured representation containing: *core ideas*, *methods*, *contributions*, *data sources*, and *experimental components*. Formally, the manuscript is mapped to a tuple $K_{\text{ms}} = \{C, M, R, D, E\}$. The prompt used to generate the structured knowledge is in Figure 6. where each element represents a set of semantically meaningful descriptors.

Retrieval of Semantically Related Papers and KG Creation: To identify prior work most relevant to the manuscript, we perform semantic retrieval using the Semantic Scholar API (Fricke, 2018). Each element of K_{ms} is used as a query, and for each query we retrieve up to five candidate papers. Let $P = \{p_1, \dots, p_n\}$ denote the union of all retrieved documents. For each retrieved paper p_i, we extract its structured knowledge representation K_i using the same knowledge-extraction module applied to the manuscript. This produces a comparable set of semantic descriptors for every candidate paper.

Afterwards, we construct a knowledge graph $G = (V, E)$ in which each node corresponds to either the manuscript or a retrieved paper.

[1] https://openreview.net/

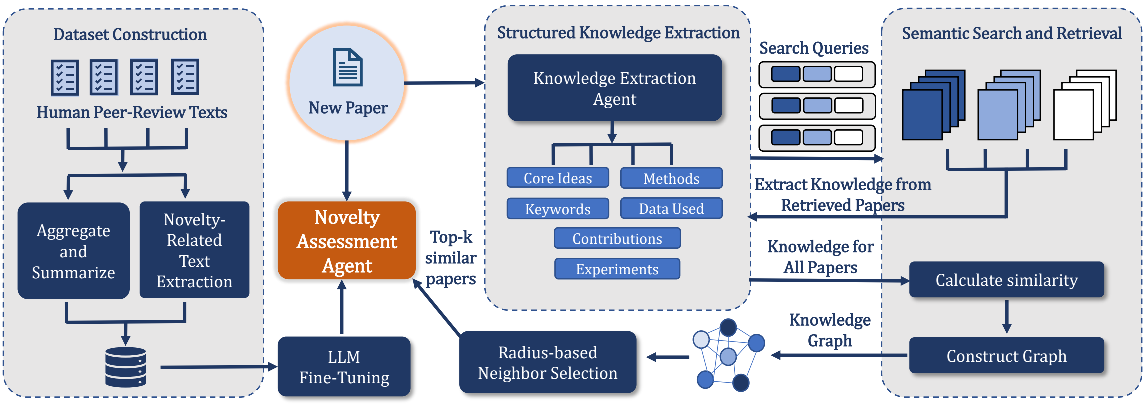

Figure 1: End-to-end Framework for Novelty Assessment. The human peer-review data is aggregated to provide a solid dataset reflecting novelty evaluations. Afterwards, semantic search and retrieval pipeline provides the model with top related papers for comparison.

Edges encode pairwise semantic similarity based on the overlap and embedding-level closeness of structured knowledge fields. For each pair (K_{ms}, K_i), we compute a cosine similarity score $s_i = \text{Sim}(K_{\text{ms}}, K_i)$ of embeddings. The top-k most similar papers, $P_{\text{top}} = \text{TopK}(P, s_i)$, are selected to form the comparative context for novelty estimation. For each (K_{ms}, K_j) where $p_j \in P_{\text{top}}$, we compute a structured overlap profile specifying which ideas, methods, or contributions exhibit similarity, and at what granularity. This overlap is used to detect idea-level redundancy and potential plagiarism, and is also considered an indicator of how novel the paper is.

Model Fine-Tuning for Novelty Assessment: To model human opinions about originality, we fine-tune Llama-3.1-8B-Instruct on our large-scale novelty dataset. Each training instance consists of the full paper text as input and the paper's aggregated novelty judgment, extracted novelty-relevant statements, and the normalized novelty score as output. The instruction format is designed to enforce paper-centric judgments rather than individual reviewer-centric summaries. Model configurations are detailed in Appendix D.

Finally, at inference time, we pass the manuscript and the structured knowledge of P_{top} to the fine-tuned novelty model. The model outputs a calibrated novelty score and a justification grounded in explicit contrasts between the manuscript and existing research. The system then compiles a comprehensive originality report consisting of: the computed similarity graph statistics, idea-level and method-level overlap analysis, the fine-tuned model's novelty score and its justification, and identification of closest prior work with highlighted conceptual intersections.

4 Experiments

We evaluate our framework on a held-out test set of 500 papers from our novelty benchmark. The evaluation focuses on both discrete novelty score prediction and free-text novelty assessment quality. We compare our approach (Novelty Reviewer) against a diverse set of strong baselines, including general-purpose LLMs of different sizes, and domain-adapted models published specifically for research review.

We show an example of the generated novelty report in Appendix B. Our proposed method provide a human-like assessment on originality aspect along with a calibrated novelty score. The judgments are grounded by a list of related research. Table 1 summarizes the quantitative performance of all models on discrete novelty score prediction over the test set. Our Novelty Reviewer outperforms both general LLMs and two fine-tuned models - Paper Reviewer (fine-tuned Qwen3-8B) and Open Reviewer (fine-tuned on the same Llama-3.1-8B-Instruct backbone) - across various metrics. The gains in precision and F1 demonstrate improved discriminative capability across novelty classes, suggesting that the model effectively separates truly novel contributions from incremental or non-novel work, even under label skew.

Beyond classification metrics, rank-based correlation measures provide a stricter test of alignment with human judgments. The Novelty Reviewer attains the highest Pearson correlation coefficient, giving stronger agreement between pre-

Table 1: Performance comparison of different models on novelty score prediction and free text evaluation.

Model	Accuracy	Precision	Recall	F1 Score	Pearson	(E−C) NLI	LLM Judge
GPT-OSS-20B (OpenAI, 2025)	0.53	0.2053	0.2950	0.2268	0.1600	0.6659	7.033
Llama-3.1-8B-Instruct (Meta, 2024)	0.15	0.1621	0.2797	0.0894	0.0828	0.4832	5.492
Mistral-7B-Instruct-v0.1 (Mistral AI, 2023)	0.07	0.1537	0.2574	0.0389	0.0746	0.5529	3.566
Qwen2.5-14B-Instruct-1M (Alibaba Cloud, 2024)	0.05	0.2620	0.2516	0.0261	-0.0321	0.4498	6.992
SciLlama (Senthil Kumar, 2025)	0.06	0.2121	0.2531	0.0294	-0.0321	0.5436	5.614
Paper Reviewer (Weathon, 2025)	0.33	0.1720	0.2893	0.1613	0.0915	0.6816	5.871
OpenReviewer (Idahl and Ahmadi, 2025)	0.08	0.1991	0.2026	0.0850	0.2412	0.5139	6.186
Novelty Reviewer	**0.62**	**0.3777**	**0.3187**	**0.3231**	**0.3121**	**0.7603**	**7.824**

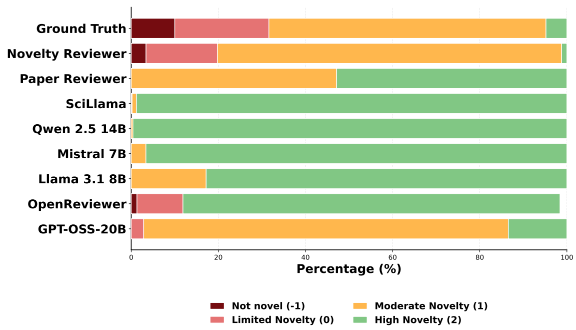

Figure 2: Distribution of predicted novelty scores across models.

dicted and ground-truth novelty rankings. This contrasts sharply with general-purpose and domain-adapted baselines, several of which exhibit weak or near-zero correlations.

Figure 2 compares the distribution of novelty predictions produced by different models against the human-derived ground truth. In contrast to most baselines, which overwhelmingly assign papers to the moderate or high novelty categories (1,2), our Novelty Reviewer exhibits a substantially higher sensitivity to low-novelty cases (-1,0), closely mirroring the empirical distribution observed in peer-review annotations. This indicates that existing models tend to systematically overestimate originality, collapsing predictions into the upper end of the novelty scale and failing to adequately penalize incremental or derivative contributions.

Beyond scalar scores, we assess the quality of generated novelty justifications using entailment–contradiction NLI metrics and Llama-3.1-8B-Instruct LLM-as-a-judge evaluation (Table 1). The Novelty Reviewer achieves the highest entailment-minus-contradiction score, indicating that its explanations are more logically consistent with ground-truth reviewer rationales. Notably, our model achieves higher LLM-as-a-judge scores than both Paper Reviewer and OpenReviewer with a clear margin, even though OpenReviewer is built on the same Llama-3.1-8B-Instruct backbone.

Case Study: Idea-level Plagiarism Detection We conduct an experiment to validate the usage of our framework in detecting concept-level plagiarized work. We randomly select a paper published in 2023, paraphrase its full abstract. We keep the ideas, methods, experiments described as exactly the same with only different word choices. We prompt our model and the baselines to evaluate the novelty described in the abstract and ask it to give a reference to a potential related article. Table 2 shows that other baselines fail to recognize ideas plagiarism, instead, tend to give high novelty scores, even the models which were able to retrieve the related paper.

Table 2: Evaluating model's ability to detect paraphrasing plagiarism.

Model	Paper Recognized	Novelty Score
GPT-OSS-20B (OpenAI, 2025)	✓	1
Llama-3.1-8B-Instruct (Meta, 2024)	×	2
Mistral-7B-Instruct-v0.1 (Mistral AI, 2023)	×	2
Qwen2.5-14B-Instruct-1M (Alibaba Cloud, 2024)	✓	2
SciLlama (Senthil Kumar, 2025)	×	2
Paper Reviewer (Weathon, 2025)	×	2
OpenReviewer (Idahl and Ahmadi, 2025)	×	1
Novelty Reviewer	✓	**0**

5 Conclusion

We present a human-aligned, literature-aware framework for automated novelty assessment. By constructing a large-scale benchmark and fine-tuning an LLM on human novelty judgments, our method captures reviewer-like evaluation behavior. Combined with structured contribution extraction and graph-based retrieval over related work, the model produces calibrated novelty scores with interpretable, grounded justifications. Extensive experiments show consistent improvements over strong baselines and enable reliable detection of idea-level plagiarism.

6 Limitations

While our framework provides a structured and human-aligned approach to assessing research novelty, it is subject to some limitations that should be acknowledged. The retrieval pipeline currently depends on the coverage and quality of external scholarly databases; which may lead to incomplete comparison set. Additionally, our framework is based on the fine-tuned model which was trained on data from NeurIPS and ICLR only. This makes the model more suitable for AI and ML research themes only.

7 Ethical Considerations

This work addresses a sensitive task of assessing research novelty and originality, which has potential implications for scientific evaluation and decision-making. Our system is designed to support, rather than replace, human judgment. Rather than acting as an automated final decision-maker, the framework serves as a complementary analytical tool that promotes consistency and transparency in novelty evaluation.

The benchmark dataset is constructed exclusively from publicly available peer-review reports, and no personally identifiable information beyond what is already disclosed in the source data is intentionally used. We would like to emphasize that the system evaluates novelty relative to accessible prior work and learned reviewer patterns in which the review process was completely double-blinded. No author names or organizational info were taken into account when fine-tuning the model.

References

Alibaba Cloud. 2024. Qwen2.5-14b-instruct-1m. Hugging Face model repository.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, and 1 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.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. Nlpeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, pages 5049–5073.

Suzanne Fricke. 2018. Semantic scholar. *Journal of the Medical Library Association: JMLA*, 106(1):145.

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.

Nino Künzli, Anke Berger, Katarzyna Czabanowska, Raquel Lucas, Andrea Madarasova Geckova, Sarah Mantwill, and Olaf von Dem Knesebeck. 2022. «i do not have time»—is this the end of peer review in public health sciences? *Public health reviews*, 43:1605407.

Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, and 1 others. 2024. What can natural language processing do for peer review? *arXiv preprint arXiv:2405.06563*.

Ruochi Li, Haoxuan Zhang, Edward Gehringer, Ting Xiao, Junhua Ding, and Haihua Chen. 2025. Unveiling the merits and defects of llms in automatic review generation for scientific papers. In *the 25th IEEE International Conference on Data Mining*.

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

Meta. 2024. Llama 3.1-8b-instruct. Hugging Face model repository.

Mistral AI. 2023. Mistral-7b-instruct-v0.1. Hugging Face model repository.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

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.

N. Senthil Kumar. 2025. SciLLaMA-3.2-3B. `https://huggingface.co/SenthilKumarN/SciLlama-3.2-3B`. Hugging Face model repository.

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)*.

Sotaro Shibayama, Deyun Yin, and Kuniko Matsumoto. 2021. Measuring novelty in science with word embedding. *PloS one*, 16(7):e0254034.

Amanda Sizo, Adriano Lino, Álvaro Rocha, and Luís Paulo Reis. 2025. Defining quality in peer review reports: a scoping review. *Knowledge and Information Systems*, pages 1–48.

Weathon. 2025. paper_reviewer. `https://huggingface.co/weathon/paper_reviewer`. Hugging Face model repository.

Zhengxu Yan, Han Li, and Yuming Feng. 2025. Noveltyrank: Estimating conceptual novelty of ai papers. *arXiv preprint arXiv:2512.14738*.

Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, Renjing Cui, Chengcheng Han, Qiushi Sun, and 1 others. 2024. Automated peer reviewing in paper sea: Standardization, evaluation, and analysis. In *Findings of the Association for Computational Linguistics: EMNLP*.

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

A Further Details

A.1 Evaluation Metrics

Entailment–Contradiction (E-C) NLI Evaluation Metric: To assess the semantic alignment between model-generated novelty assessments and reference novelty summaries, we adopt a discriminative Natural Language Inference (NLI)–based metric. Given a model output as the *premise* and the corresponding reference novelty summary as the *hypothesis*, we use a pretrained NLI classifier (Roberta) to estimate the probabilities of *entailment* and *contradiction*.

Formally, for each pair (p_i, h_i), the NLI model produces a probability distribution $\mathbf{q}_i = [q_i^{\text{contra}}, q_i^{\text{neutral}}, q_i^{\text{entail}}]$. We aggregate these scores across the evaluation set and compute the E–C score as

$$\text{E-C} = \frac{1}{2}\left(\frac{1}{N}\sum_{i=1}^{N}\left(q_i^{\text{entail}} - q_i^{\text{contra}}\right) + 1\right) \quad (1)$$

where N denotes the number of evaluated samples. The affine transformation scales the metric to the range $[0, 1]$, with higher values indicating stronger entailment and lower contradiction between the model output and the reference. Intuitively, this metric rewards outputs that are semantically supported by the reference novelty summary while penalizing statements that contradict it.

LLM Judge: Figure 3 shows the usage of LLM as a judge for evaluating how the model's response align with the ground truth text.

A.2 Scores Distribution

Figure 4 compares the distribution of novelty scores produced by different automated reviewers against the ground-truth annotations. Most baseline LLMs exhibit a strong bias toward high novelty scores, with distributions sharply concentrated near the upper end of the scale and medians close to 2. This indicates a systematic tendency to overestimate novelty and collapse predictions into a narrow, overly optimistic range. In contrast, the Novelty Reviewer produces a broader and more balanced distribution, closely matching the shape, mean, and median of the ground-truth scores. The alignment suggests better calibration across the full novelty spectrum and a higher sensitivity to distinguishing incremental contributions from genuinely novel work.

```
Rate how well the Model Output aligns with the Reference Answer on a scale from 0 to 10,
taking into account both the correctness and coverage of the novelty description and
consistency with the provided Novelty Score.

Reference Answer: {}

Model Output: {}

Respond with ONLY a single number from 0 to 10.
```

Figure 3: Prompt for free text evaluation using LLM as judge.

B Output Report Example

We show an example of the manuscript novelty report below.

Generated: 2026–01–03 00:02:27

Analyzed Submitted Manuscript

Novelty Score: 1

Short Novelty Review: The paper addresses an important and timely problem in open-world semantic segmentation by explicitly distinguishing semantic-level (class) shifts from domain-level (covariate) shifts, which is a recognized challenge in OOD detection and domain generalization. The core ideas—generative data augmentation to simulate anomalies and domain shifts, uncertainty-aware training, and feature alignment for domain invariance—are all grounded in well-established research directions. While the proposed combination is thoughtful and well-motivated, each component closely aligns with existing paradigms such as synthetic anomaly generation, uncertainty calibration for OOD detection, and domain-aligned representation learning. The primary novelty lies in integrating these components into a unified framework explicitly designed to disentangle semantic and domain shifts within segmentation models. This integration and the specific training strategy tailored to recalibrate uncertainty for semantic shifts provide some originality beyond straightforward application of prior techniques. However, the paper does not introduce a fundamentally new problem formulation, theoretical insight, or learning principle; instead, it refines and combines known ideas in a systematic way.

Core Ideas. Semantic segmentation under multiple distribution shifts; jointly handling anomaly segmentation and domain generalization.

Methods. Coherent generative-based data augmentation; learnable uncertainty function; relative contrastive loss; two-stage noise-aware training.

Keywords. Semantic segmentation; anomaly detection; domain generalization; generative-based data augmentation; learnable uncertainty function; relative contrastive loss; two-stage training.

Most Similar Papers

Rank #1 Similarity: 27.9%
Title: Show or Tell? Effectively Prompting Vision–Language Models for Semantic Segmentation
Year: 2025 **Citations:** 2
Their Ideas. Effectively prompting vision–language models for semantic segmentation; comparison between textual and visual prompts.
Their Methods. Few-shot prompted semantic segmentation; PromptMatcher, a training-free baseline combining text and visual prompts.

Rank #2 Similarity: 25.6%
Title: Confidence-aware Training of Smoothed Classifiers for Certified Robustness
Year: 2022 **Citations:** 10
Their Ideas. Use of smoothed classifiers to construct models with provable robustness against ℓ_2 adversarial perturbations.
Their Methods. Randomized smoothing; sample-wise control of robustness during training.

Rank #3 Similarity: 24.8%
Title: Improved Stability and Generalization Guarantees of the Decentralized SGD Algorithm
Year: 2023 **Citations:** 8
Their Ideas. Generalization error analysis of decentralized stochastic gradient descent based on algorithmic stability.
Their Methods. Decentralized stochastic gradient descent (D-SGD).

Rank #4 Similarity: 24.8%
Title: DockGame: Cooperative Games for Multimeric Rigid Protein Docking
Year: 2023 **Citations:** 2

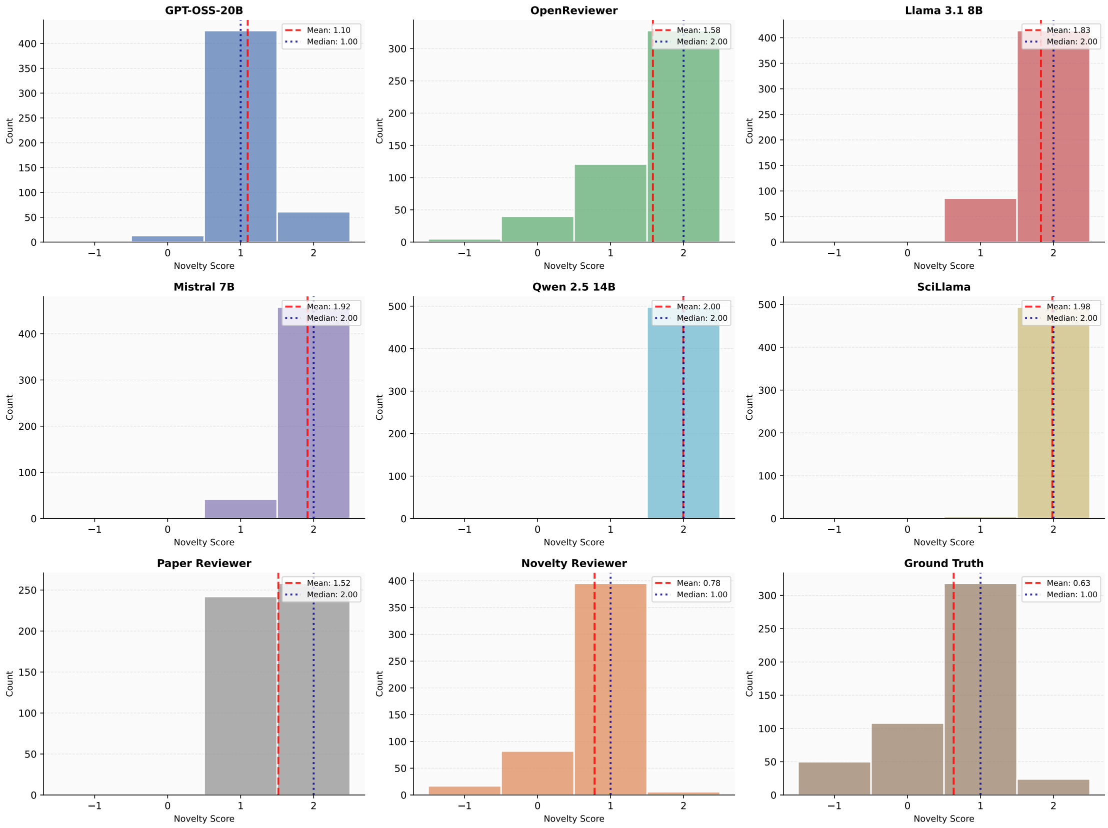

Figure 4: Distribution of predicted novelty scores across different models compared to ground truth. Histograms show the frequency of novelty scores assigned by various LLM-based reviewers and baselines. Red dashed lines indicate the mean score, while blue dotted lines denote the median for each model.

Their Ideas. Modeling multimeric rigid protein docking as a cooperative game between proteins.
Their Methods. Gradient-based learning with surrogate potentials; diffusion-based generative models over protein action spaces.

This report is automatically generated and intended to provide a structured overview of semantic similarity between the analyzed manuscript and existing literature as a helper tool for reviewers.

C System Prompts

We provide the prompts used for designing the LLM agents in this appendix. Figure 5 shows the prompt for constructing the dataset. Figure 6 shows the prompt used for the knowledge extraction agent.

D LLM Fine-Tuning Setup

The Novelty Reviewer model is obtained by fine-tuning meta-llama/Llama-3.1-8B-Instruct on our novelty benchmark dataset. It will be publicly available after the blind review phase. Training is performed on the full tokenized dataset, which contains aggregated peer-review supervision instances derived from our dataset construction pipeline. The model is trained to generate reviewer-aligned novelty scores and structured justifications conditioned on manuscript content and retrieved contextual evidence. The training corpus consists exclusively of peer-review–derived supervision, where each example includes the manuscript text, extracted novelty-related review statements, and normalized novelty labels. All evaluations are conducted on a held-out test set.

Training Procedure: Fine-tuning is performed for three epochs using the AdamW optimizer with a cosine learning rate schedule. The learning rate is set to 2×10^{-5} with a warmup of 50 steps. Training proceeds for a total of 2,433 optimization steps with gradient accumulation disabled. Each micro-batch contains a single sequence, and distributed training is used to achieve an effective batch size of 32 across 32 GPUs.

To support long-context modeling, we set the maximum sequence length to 32,120 tokens, enabling the model to process full manuscripts and

```
You are an expert reviewer for top-tier AI and ML conferences. Your task is to analyze paper
reviews focusing on novelty aspects and synthesize a single, cohesive novelty assessment.

Given multiple peer reviews of a research paper, distill the collective perspective on novelty
and originality into a unified evaluation written as direct statements about the work itself.
Be factual, concise, and balanced.

Do not add your own judgment; only reflect the points expressed in the given reviews. Write
in a direct voice about the paper (e.g., "The paper introduces...", "The work extends...",
"The approach combines...").

You will be provided with full review texts and pre-extracted novelty-related quotes and
assessments from each review.

Read and analyze carefully the following reviews for a research paper:
Full Reviews: {reviews}
The following segments were identified as novelty discussions for each review:
{novelty_excerpts}
Assign a novelty score [-1:2] reflecting how reviewers collectively perceive the
originality of the paper.

Base the score only on their aggregated opinions.
-1 = Not novel (Reviewers consistently describe work as incremental, derivative, or
replicating existing approaches with minimal innovation.)

0 = Limited novelty (Reviewers find the work somewhat standard, showing minor variations or
applications of known methods without substantial conceptual or technical innovation.)

1 = Moderately novel (Reviewers acknowledge some originality but note significant overlap
with prior work, or see it as a competent extension/combination of existing ideas.)

2 = Highly novel (Reviewers recognize fundamentally new ideas, approaches, problem
formulations, or insights that significantly advance the field.)


If reviewers disagree significantly on specific aspects, present both perspectives as
contrasting assessments of the work itself (e.g., "The method extends X, though the extensions
remain largely incremental").


The output should be in the following format:

Novelty Score: [-1, 0, 1, 2]

Score Justification:

[2-3 sentences explaining the novelty level as direct statements about the paper's
contributions and their originality]

Detailed Assessment:

[4-6 sentences written as direct statements about the paper covering:]

- The main novel contributions or new ideas introduced

- Limitations in originality or areas of incremental advancement

- Specific aspects of novelty: problem formulation, methodological innovations, experimental
insights, or theoretical advances
```

Figure 5: Prompt for human peer-reviews aggregation.

```
Extract information from this research paper and return ONLY valid JSON.
Title: {title}
Content: {text}
Return this JSON structure:
{{"core_ideas": ["list of main ideas"],
"methods": ["list of methods"],
"contributions": ["list of contributions"],
"keywords": ["list of keywords"],
"data_sources": ["list of data sources"],
"experiments": ["list of experiments"]}}
```

Figure 6: Prompt for extracting structured knowledge from paper text.

aggregated review contexts without truncation, and sample packing is disabled to preserve document boundaries. We enable gradient checkpointing to reduce memory usage, along with Flash Attention for improved computational efficiency. Training is conducted in mixed precision with automatic bfloat16 support and TensorFloat-32 acceleration. Weight decay is disabled, following standard practice for instruction fine-tuning of large language models.

System Configuration: Training is executed using multi-GPU distributed execution (32 A100 GPUs with 80 GB memory each) with DeepSpeed ZeRO Stage-3 optimization. Additional efficiency improvements are provided via Liger kernel integrations, including optimized rotary embeddings, RMS normalization, GLU activations, and fused linear cross-entropy operations. The model follows the Llama-3.1-8B-Instruct chat template and uses custom padding and end-of-sequence tokens. All experiments are conducted using Transformers 4.55.2, PyTorch 2.6.0, Datasets 4.0.0, and Tokenizers 0.21.4. Checkpoints are saved twice per epoch, and only model weights are retained to reduce storage overhead.