arXiv:2508.10795v2 [cs.CL] 17 Aug 2025

Beyond "Not Novel Enough": Enriching Scholarly Critique with LLM-Assisted Feedback

Osama Mohammed Afzal[1], Preslav Nakov[2], Tom Hope[3], Iryna Gurevych[1]

[1] UKP Lab, TU Darmstadt and Hessian Center for AI (hessian.AI)
[2]MBZUAI, [3]The Allen Institute for AI (AI2),
www.ukp.tu-darmstadt.de

Abstract

Novelty assessment is a central yet understudied aspect of peer review, particularly in high-volume fields like NLP where reviewer capacity is increasingly strained. We present a structured approach for automated novelty evaluation that models expert reviewer behavior through three stages: content extraction from submissions, retrieval and synthesis of related work, and structured comparison for evidence-based assessment. Our method is informed by a large-scale analysis of human-written novelty reviews and captures key patterns such as independent claim verification and contextual reasoning. Evaluated on 182 ICLR 2025 submissions with human annotated reviewer novelty assessments, the approach achieves 86.5% alignment with human reasoning and 75.3% agreement on novelty conclusions—substantially outperforming existing LLM-based baselines. The method produces detailed, literature-aware analyses and improves consistency over ad hoc reviewer judgments. These results highlight the potential for structured LLM-assisted approaches to support more rigorous and transparent peer review without displacing human expertise. Data and code are made available.[1]

1 Introduction

The peer review system is collapsing under its own success. Two independent committees at NeurIPS 2021 disagreed on 23% of identical papers (Beygelzimer et al., 2023)—a breakdown in consistency that signals deeper problems than mere capacity constraints. With manuscript submissions doubling every 15 years (Larsen and Ins, 2010) and reviewers now handling 14 evaluations annually (Díaz et al., 2024), the system's 15 million annual reviewing hours (Aczel et al., 2021) are producing increasingly unreliable outcomes.

[1]https://ukplab.github.io/arxiv2025-assessing-paper-novelty

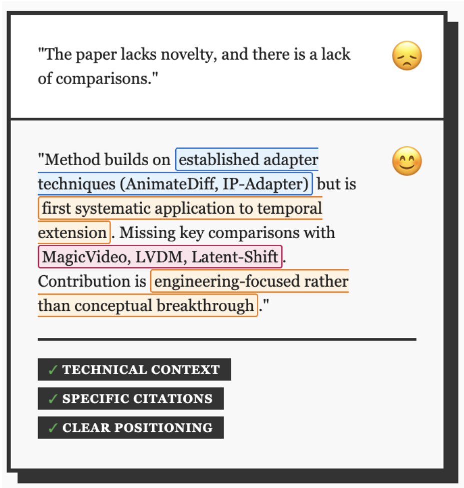

Figure 1: Comparison of a surface level human (Top) vs. LLM-written novelty assessment (Bottom).

Among peer review tasks, novelty assessment stands out as one of the most problematic (Ernst et al., 2020) (Horbach and Halffman, 2018). Novelty assessment requires reviewers to determine whether a submission makes sufficiently original contributions by identifying what specific advances it makes beyond existing work, evaluating whether these advances are significant enough to warrant publication, and verifying that the authors have accurately characterized their contributions relative to prior research. This knowledge-intensive process demands that reviewers maintain comprehensive awareness of related work across their field and can precisely distinguish between meaningful innovations and incremental modifications—a task that becomes exponentially more difficult as publication rates accelerate and research domains specialize. Overwhelmed reviewers often resort to superficial analyses, producing vague feedback like "not novel enough" without clear justification. The challenge compounds when reviewers encounter papers outside their specific expertise, leading to either overly conservative rejections or inadequate assessments that fail to catch incremental work (Kuznetsov et al., 2024).

Recent advances in large language models present an unprecedented opportunity to address these novelty assessment challenges at scale. These breakthrough technologies have revolutionized text processing and demonstrated remarkable performance across knowledge-intensive tasks (Raiaan et al., 2024), with recent technical advancements expanding capabilities to specialized reasoning and efficient inference (Li et al., 2024a; Zhang et al., 2025).

While recent LLM advances create this opportunity, no existing work specifically addresses novelty assessment as a dedicated task within the peer review process. Prior research incorporates novelty evaluation within idea generation pipelines (Radensky et al., 2025; Lu et al., 2024; Li et al., 2024b), generates peer reviews with novelty assessments occurring as a result of them existing in peer reviews from training data (Idahl and Ahmadi, 2025; D'Arcy et al., 2024), or adds novelty assessment steps to review synthesis pipelines for improvement (Zhu et al., 2025). However, these approaches either operate on synthetic ideas rather than real research contributions or fail to evaluate novelty assessment capabilities in isolation. This represents a critical gap requiring specialized methodologies for peer review novelty assessment.

To address this gap, we propose an end-to-end novelty assessment pipeline for peer review submissions. Our approach consists of three stages: document processing and content extraction, related work retrieval and ranking, and structured novelty assessment. The final stage implements four sequential steps: novelty related content selection from the submission pdf, building comprehensive understanding of related work from retrieved papers, comparing claimed novelty against the comprehensive analysis from the prior step, and generating a summary with cited evidence from the comparison. This pipeline operates on real research papers and directly evaluates novelty assessment capabilities, addressing the limitations of existing approaches. Importantly, we conduct the first evaluation of LLMs for novelty assessment using actual human data, including annotated novelty assessment statements, and provide comprehensive evaluation across multiple dimensions.

Research Questions and Contributions This work aims to address the following research questions:

1. How does our human-informed novelty assessment pipeline compare to existing approaches?

2. How well do our assessments align with human reviewer preferences across key evaluation dimensions?

3. Can automated evaluation reliably substitute for human judgment in assessing novelty assessment quality?

Our contributions are threefold:

- **Human Analysis Dataset and Insights**: A systematically curated dataset of 182 papers with annotated human novelty assessments from ICLR 2025, along with empirical insights into expert reviewer reasoning patterns, evaluation criteria, and argument structures that inform AI system design for novelty assessment.

- **Human-Informed Pipeline**: A literature-grounded pipeline that incorporates insights from human novelty assessment practices, featuring structured prompting strategies and targeted content extraction informed by observed expert reviewer behavior.

- **Comprehensive Evaluation and Analysis**: Systematic comparison of our human-informed approach against existing baselines and human reviewers, with fine-grained evaluation across multiple dimensions and validation of automated assessment methods.

2 Related Work

AI-Assisted Peer Review Systems Our work is positioned at the peer review stage of scientific research, where our system operates when a manuscript is submitted for evaluation. While previous works (D'Arcy et al., 2024) (Idahl and Ahmadi, 2025) (Zhu et al., 2025) (Chitale et al., 2025) (Chang et al., 2025) (Nemecek et al., 2025) have developed end-to-end peer review generation pipelines that may implicitly include novelty assessment steps, we are the first to focus specifically on building a dedicated pipeline for novelty assessment and the first to systematically evaluate LLMs on this task. A related line of work operates at the

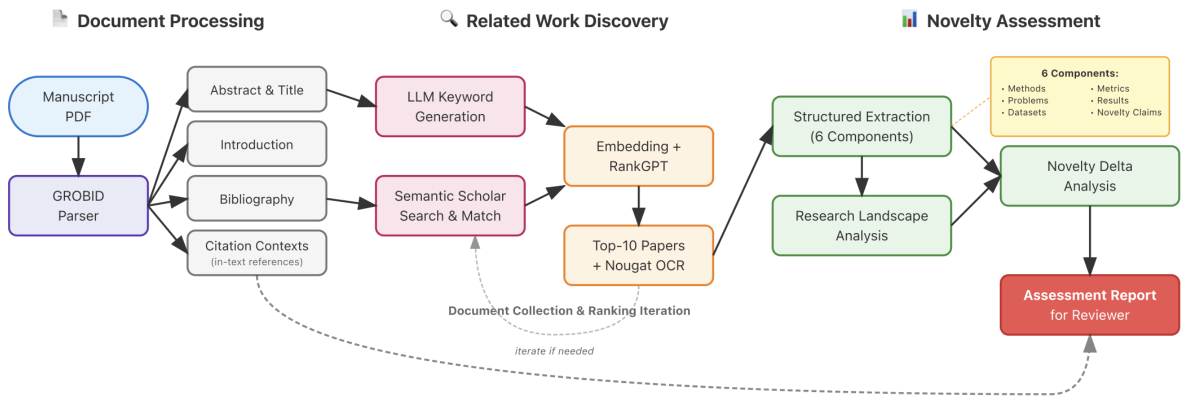

Figure 2: Automated novelty assessment pipeline. The system processes manuscripts through three stages: (1) Document Processing extracts content using GROBID, (2) Related Work Discovery identifies and ranks relevant papers via embedding similarity and LLM reranking, and (3) Novelty Assessment performs structured analysis to generate evidence-based novelty evaluations.

ideation stage of research (Radensky et al., 2025) (Shahid et al., 2025) (Li et al., 2024b) (Lu et al., 2024), developing pipelines for research idea generation that aim to improve novelty through feedback loops from a novelty assessor. In contrast, we operate at a more mature stage where ideas have been fully executed and comparative analyses are well-formulated. The evaluation in ideation-stage works focuses on synthetic ideas that are typically abstract and loosely defined, whereas we evaluate concrete, polished research contributions that have undergone the refinement process of execution and manuscript preparation.

Scientific Literature Analysis & Retrieval Our work employs an extensive related work discovery pipeline that collects papers cited within the submission and additionally retrieves related papers by querying with prompts generated by GPT-4.1. Papers are then ranked using an embedding-based method and reranked using RankGPT. We adapt this general approach from existing work (Radensky et al., 2025)(Shahid et al., 2025)(Li et al., 2024b) with modifications to ranking and filtering for our specific task. Similar retrieval-rank-rerank pipelines have been used for related work generation (Agarwal et al., 2025). Another retrieval approach is OpenScholar (Asai et al., 2024), which uses an LLM-RAG based approach to answer scientific queries by identifying relevant passages from 45 million open-access papers. Works like DeepReviewer (Zhu et al., 2025) incorporate OpenScholar for novelty validation. However, our primary criticism of OpenScholar for novelty assessment is that it provides only generic comparisons rather than the granular analysis across methodology, problem formulation, evaluation approaches, and novelty claims that our task requires.

Evaluation of LLM Generated Text Prior works evaluating generated peer reviews have adopted either quantitative evaluations, where they compare LLM-assigned scores (such as Overall Score, Soundness, etc.) against human-assigned scores on review forms, or qualitative evaluations using traditional metrics like BERTScore (Zhang et al., 2020), ROUGE (Lin, 2004), and BLEU (Papineni et al., 2002), or more recent approaches like LLM-as-Judge (Zheng et al., 2023). We adopt the LLM-as-Judge approach for our evaluation. Notably, no prior work has specifically evaluated LLM performance on novelty assessment as a dedicated task, making our evaluation framework the first of its kind.

3 Methodology

3.1 Human Analysis for Prompt Design

To understand how humans conduct novelty assessment, we analyzed reviews from ICLR 2025, which explicitly requires novelty evaluation with dedicated review sections, making novelty discussions more frequent than in other venues.

We sourced submissions from OpenReview and used keyword-based search for terms including "novel", "original", "research gap", "innovation", "incremental", "prior work", and "existing work". Papers were ranked using a scoring function prioritizing: (1) reviews with >4 novelty keywords, (2) consistent novelty discussion patterns across reviews, and (3) total review count. We selected the top 200 papers for analysis.

To speed up the annotation process, we employed multiple instances of GPT-4o mini to perform sentence-level classification, determining whether individual sentences discussed novelty. This classification helped human annotators identify general areas where novelty discussions take place, after which humans selected all sentences containing actual novelty assessments. This process revealed that 18 of the 200 sampled papers (9%) contained limited genuine novelty assessments, often triggered by keyword matches referring to paper components rather than novelty evaluation. The remaining 182 papers formed our final dataset for analysis. We systematically analyzed the selected assessments to identify recurring patterns in reviewer reasoning, evaluation criteria, and argument structures. This analysis focused on how reviewers structure their novelty arguments, what evidence they prioritize, and how they compare submissions to prior work.

This analysis revealed several key patterns in how expert reviewers assess novelty:

Verification over acceptance: Rather than accepting author claims at face value, reviewers independently verify relationships with prior work and critically examine how authors characterize related research, often distinguishing between author framing and actual technical relationships. Our prompt explicitly instructs models to "independently verify relationships" and "distinguish between author-claimed differences and independently observed differences," mirroring this critical verification approach, as shown in Figures 10 and 11.

Variable granularity: Reviewers assess contributions with varying detail—some providing global novelty assessments while others examine each contribution separately against relevant prior work. (We address this through the "Contribution Delta Analysis" section that systematically examines each claimed contribution individually against the most similar prior work, ensuring comprehensive coverage regardless of author presentation style, as detailed in Figure 11.)

Different analytical lenses: Some reviewers focus on methodological innovations while others evaluate systems holistically, calibrating expectations based on field maturity. Our prompt incorporates multiple analytical perspectives through separate sections for research positioning, methodological relationships, and field context considerations that help calibrate novelty expectations based on area maturity, shown across Figures 10 and 11.

Gap identification: Reviewers systematically identify gaps in related work discussions and distinguish between implementation-level improvements and genuine conceptual advances. (The "Related Work Considerations" section specifically instructs models to identify missing comparisons and assess whether improvements stem from "implementation details rather than conceptual advances," directly addressing this reviewer behavior in Figure 11.) These insights informed both our prompt task design and the input to the LLM.

3.2 Our Approach

Overview Our pipeline processes submission PDFs and generates structured novelty assessments through three stages (Figure 2): (i) Document Processing extracts key content from submissions, (ii) Related Work Discovery identifies and ranks relevant prior work, and (iii) Novelty Assessment performs comparative analysis to generate evidence-based novelty evaluations.

3.3 Stage 1: Document Processing

We extract structured content from submission PDFs using GROBID[2] to obtain titles, abstracts, bibliographies, and citation contexts required for subsequent stages.

3.4 Stage 2: Related Work Discovery

This stage identifies and ranks related work through a multi-step retrieval pipeline designed to capture both explicitly cited works and potentially relevant uncited research.

Cited Work Processing Bibliography entries are matched against Semantic Scholar to obtain standardized metadata (title, abstract, authors, publication date, venue) for consistent downstream processing.

Uncited Work Discovery To identify relevant work not cited by authors, we generate 5 keyword queries using GPT-4.1 and search Semantic Scholar. Results are filtered to remove exact title matches with the submission (avoiding potential preprints) and papers published after the submission date.

Embedding-based Ranking We generate embeddings for all collected papers using SPECTER v2 (Singh et al., 2022) on concatenated titles and abstracts. Papers are ranked by cosine similarity

[2]https://github.com/kermitt2/grobid

to the submission's embedding to identify semantically similar work.

LLM-based Reranking To prioritize papers with conceptual rather than purely semantic similarity, we employ LLM-based reranking (Sun et al., 2023b,a) with prompts emphasizing methodological approaches, novelty claims, and problem statements. We select the top-K (k=20) papers for novelty assessment.

Content Extraction For selected papers, we retrieve PDFs through a hierarchical search across Semantic Scholar, ACL Anthology, and arXiv. Retrieved papers are processed using MinerU (Wang et al., 2024; He et al., 2024) to extract introduction sections, with Nougat OCR (Blecher et al., 2023) as fallback for processing failures. We use these tools for OCRs here as they output more accurate OCRs and we will be using this paper content in the next stage.

3.5 Stage 3: Novelty Assessment

We use GPT-4.1 (OpenAI, 2024) for its improved instruction-following capabilities. This stage consists of four sequential steps.

Structured Extraction Processing retrieved papers as raw text creates context optimization challenges that degrade LLM performance. Recent research demonstrates that model performance consistently degrades with increasing input length, even when task complexity remains constant (Hong et al., 2025). This occurs because either overwhelming models with unrelated information reduces accuracy (Zhu et al., 2025; Idahl and Ahmadi, 2025) or insufficient context through heavy truncation limits understanding (Radensky et al., 2025).

We extract six structured components aligned with novelty assessment requirements from each paper's title, abstract, introduction: (i) Methods, (ii) Problems addressed, (iii) Datasets, (iv) Results, (v) Evaluation approaches, and (vi) Novelty Claims. This preserves essential information while reducing context length to mitigate the performance degradation observed with longer, unstructured inputs (Figure 8).

Landscape Analysis Expert reviewers are typically assigned papers within their areas of expertise, providing them with comprehensive domain knowledge of established benchmarks, common techniques, evaluation metrics, and recent developments. To approximate this contextual foundation, we incorporate a landscape analysis step that systematically organizes the previously extracted structured components from retrieved related work.

Decision	Papers	Reviews	Words/rev	Rev/paper
No Decision / Withdrawn	51	110	1002	2.16
Reject	81	195	919	2.41
Accept (Poster)	45	102	962	2.27
Accept (Spotlight)	4	10	997	2.50
Accept (Oral)	1	2	1182	2.00
Total	182	419	959	2.30

Table 1: Distribution of papers and reviews with novelty discussions by ICLR 2025 decision outcomes

Using GPT-4.1, we perform cross-paper synthesis to identify methodological clusters, trace problem evolution over time, map evaluation ecosystems, and establish technical relationships between approaches (Figure 9). The landscape analysis produces a hierarchical organization of the research space with explicit connections between related approaches, competing methods, and complementary techniques.

This structured representation serves as contextual background for subsequent novelty assessment, mimicking the organized domain understanding that expert reviewers naturally possess when evaluating papers in their field.

Novelty Delta Analysis This step performs comparative analysis between the submission and prior work using three inputs: (1) the research landscape, (2) the submission's claimed contributions, and (3) citation contexts—sentences where the submission cites related work. Citation contexts reveal how authors position their contributions, enabling verification of claimed distinctions versus rhetorical framing.

Using GPT-4.1 with prompts informed by our human analysis (Section 3.1), the system implements key reviewer patterns: independent verification of author claims, variable granularity examination of contributions, and identification of gaps in related work discussions (Figures 10 and 11).

Assessment Report Generation The final step generates a concise paragraph long summary that appears similar to actual peer review novelty assessments, enabling direct comparison with human-written assessments (Figure 12).

System	Reasoning Alignment (%↑)	Conclusion Agreement (%↑)	Positive Shift (%↓)	Negative Shift (%↓)
OpenReviewer (Idahl and Ahmadi, 2025)	42.4 ± 0.39	46.8 ± 0.71	6.3 ± 0.27	15.3 ± 0.40
DeepReviewer (Zhu et al., 2025)	50.6 ± 0.67	51.5 ± 1.24	21.7 ± 1.89	9.1 ± 0.00
Human vs. Human	65.1 ± 1.05	62.8 ± 0.40	6.7 ± 0.79	15.0 ± 0.40
Scideator (Radensky et al., 2025)	23.7 ± 0.00	22.4 ± 0.00	0.0 ± 0.00	20.5 ± 0.00
Ours	**86.5** ± **0.20**	**75.3** ± **0.85**	16.3 ± 1.28	3.0 ± 0.43

Table 2: Summary of Reasoning Alignment, Conclusion Agreement, Positive Shift, and Negative Shift Metrics

4 Evaluation

We use the data we annotated before the prompt design stage when studying human patterns. We prompt GPT-4.1 with each human review and its corresponding annotated novelty assessment statements to generate a coherent novelty assessment using the prompt in Figure 13. This step is necessary because novelty-related comments typically appear scattered throughout reviews rather than as unified assessments. Simply concatenating these fragments would introduce stylistic biases during evaluation, as the disjointed human comments would differ markedly from the coherent assessments our system generates. The GPT-4.1 synthesized assessments serve as our evaluation ground truth.

4.1 Evaluation Methods

Automated Evaluation Evaluating novelty assessment systems presents significant challenges due to the subjective and knowledge-intensive nature of the task. What constitutes "novel" depends heavily on the evaluator's familiarity with the surrounding research landscape. Even when human reviewers reach similar novelty conclusions, they may arrive at these decisions through different reasoning paths and evidence bases.

Given these challenges, we employ an LLM-as-Judge framework using our style-normalized human novelty assessments as ground truth. We evaluate AI-generated assessments across four key dimensions using the prompts in Figures 14 and 15 with GPT-4.1 as our Judge:

Novelty Conclusion Alignment: Whether the AI assessment reaches similar novelty conclusions as human reviewers.

Novelty Reasoning Alignment: Whether the AI's reasoning process and justifications align with human reviewer logic.

Prior Work Engagement: Whether the assessment demonstrates adequate engagement with relevant literature rather than superficial analysis.

Depth of Analysis: Whether the assessment provides substantive, detailed evaluation rather than surface-level observations.

These dimensions ensure that AI assessments not only align with human judgments but also meet quality standards for thorough, evidence-based novelty evaluation. Our evaluation employs a two-stage process to ensure consistency. First, we extract core judgments (key novelty strengths and weaknesses) from human reviews using GPT-4.1 with the prompt in Figure 14. We perform this extraction separately to establish stable reference judgments, as combining extraction with evaluation would risk the LLM identifying different claims across comparisons. In the second stage, we evaluate AI-generated assessments against these pre-extracted judgments using the prompt in Figure 15. This evaluation quantifies four aspects: (1) *judgment similarity*, measuring whether the AI identifies the same specific novelty aspects with confidence scores; (2) *conclusion alignment*, checking whether bottom-line novelty sufficiency verdicts match; (3) *prior work engagement*, categorized as None, Limited (1-2 citations), or Extensive (3+); and (4) *depth of analysis*, rated as Surface Level, Moderate (1-2 aspects), or Deep (3+ detailed comparisons). Table 2 reports the resulting alignment scores across these dimensions.

Human Evaluation To validate our automated evaluation, we conduct human evaluation using three PhD students (two third-year, one first-year) specializing in NLP and AI for Science, all with multiple conference publications. The evaluation uses pairwise comparison across the same four dimensions.

Evaluators compare side-by-side novelty assessments from different systems (human ground truth, our approach, or baselines) with each pair representing different system types. We collected 100 total comparisons: 25 overlapping samples per evaluator (for inter-annotator agreement) and 25 unique samples each.

For each dimension, evaluators select: Candidate A wins, Candidate B wins, Tie, or Unclear. A comment box captures specific observations and assumptions, enabling both quantitative preference measurement and qualitative insights. The web interface used for human preference collection can be seen in Figures 6 and 7.

Inter-Rater Reliability Table 6 shows moderate inter-rater agreement (0.493-0.560) with fair kappa scores (0.287-0.368), reflecting the inherent subjectivity in evaluating novelty assessment quality.

4.2 Baseline Methods

We compare our approach against three existing systems, adapting each for novelty assessment evaluation.

Scideator (Radensky et al., 2025) Scideator includes a novelty classification module that uses GPT-4o with few-shot examples and task definition to classify ideas as 'novel' or 'not novel'. Originally designed for idea synthesis pipelines where LLMs iteratively refine ideas based on novelty feedback, we adapt it by translating Scideator's "idea" input to "title and abstract" input—recognizing these as the crystallized form of a scientific idea. This preserves Scideator's classification approach while shifting from assessing nascent ideas to evaluating completed scientific contributions.

OpenReviewer (Idahl and Ahmadi, 2025) OpenReviewer generates comprehensive peer reviews using Llama-OpenReviewer-8B, trained on 79,000 expert reviews from top conferences. Since it generates complete reviews rather than targeted novelty assessments, we extract novelty-related content from its outputs using the same LLM-based approach applied to human review normalization as shown in prompt in figure 13.

DeepReviewer (Zhu et al., 2025) DeepReviewer is a multi-stage review framework that combines literature retrieval done with OpenScholar (Asai et al., 2024) with evidence-based argumentation, powered by DeepReviewer-14B trained on structured review annotations. We extract novelty assessments from its outputs using the same LLM-based extraction approach employed for OpenReviewer and human reviews. Notably, DeepReviewer was trained on ICLR 2025 data, which encompasses our entire evaluation dataset—a critical consideration when interpreting its performance.

5 Results and Analysis

We evaluated each system by comparing its novelty assessments against human novelty assessments as reference. For papers with multiple human reviewers, we also conducted human-vs-human comparisons to establish a baseline. Table 2 presents the overall results.

System	Surface-Level (%)	Moderate (%)	Deep (%)
OpenReviewer	67.4	31.3	1.2
DeepReviewer	43.4	56.6	0.0
Human vs. Human	22.3	66.2	11.5
Scideator	44.9	54.5	0.6
Ours	0.0	47.9	52.1

Table 3: Reasoning Depth Distribution (Percentages)

System	None (%)	Limited (%)	Extensive (%)
OpenReviewer	39.9	53.1	7.0
DeepReviewer	24.7	75.3	0.0
Human vs. Human	19.6	65.2	15.2
Scideator	0.0	75.9	24.1
Ours	0.0	39.1	60.9

Table 4: Prior Work Engagement Distribution (Percentages)

5.1 Overall Performance

Our system significantly outperforms both AI baselines and the human-vs-human baseline across key metrics. For Reasoning Alignment, our system achieves scores that are 44.1 and 35.9 percentage points higher than OpenReviewer (Idahl and Ahmadi, 2025) and DeepReviewer (Zhu et al., 2025), respectively, and 21.4 percentage points above the human baseline. For Conclusion Agreement, our system again leads all three baselines, with the human baseline performing closest at approximately 13 percentage points below our system. Outputs of our pipeline in comparison to the baselines can be seen in Tables 7, 8 and 9.

5.2 Sentiment Shift Analysis

We analyze two derived metrics from novelty conclusions. Positive Shift measures assessments changing from neutral/negative to positive sentiment compared to human reference, while Negative Shift measures the opposite direction. Higher Positive Shift indicates overly optimistic assessment, while higher Negative Shift suggests excessive criticism. AI systems generally demonstrate optimistic bias, as evidenced by DeepReviewer's high Positive Shift score. Our system shows lower Positive Shift than DeepReviewer, though OpenReviewer aligns most closely with human rates. For Negative Shift, humans tend to be more critical—a pattern mirrored by OpenReviewer, followed by DeepReviewer. Our approach achieves the lowest Negative

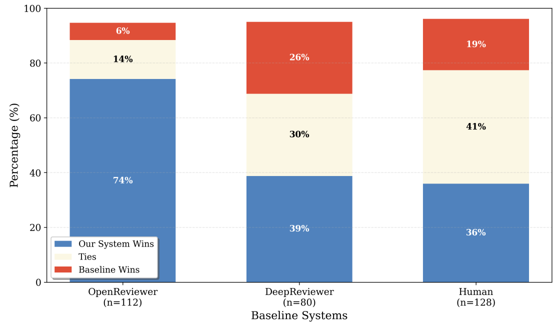

Figure 3: Overall performance comparison between our system and three baseline systems based on human evaluation (n values indicates number of comparisons)

System Configuration	Reasoning (%)	Conclusion (%)
Naive Prompt	39.3	24.7
+ Our Prompt Design	80.0 (+40.7)	71.5 (+46.8)
+ Structured Extraction	83.3 (+3.3)	76.0 (+4.5)
+ Landscape Analysis (Full)	86.5 (+3.2)	75.3 (–0.7)

Table 5: Component Analysis: Incremental Contribution of Pipeline Components

Shift rate.

5.3 Depth and Prior Work Engagement

Tables 3 & 4 show our system achieves the highest scores for both dimensions, producing no surface-level analyses unlike all baselines. This stems from our specialized multi-step pipeline targeting novelty assessment, while other systems generate complete peer reviews where novelty is a minor component. OpenReviewer performs worst, lacking retrieval and relying on parametric knowledge. DeepReviewer uses OpenScholar retrieval but fails at comparative analysis. Human reviewers show high variance, with some engaging extensively while others provide minimal analysis.

5.4 Human Evaluation Validation

We conducted human evaluations to validate our LLM-as-Judge evaluation framework. Figures 3 and 4 show the pairwise comparison results. Against OpenReviewer, our system wins 74% of the time. Performance against DeepReviewer and human reviewers is more mixed (39% and 36% win rates), but high tie rates (30% and 41%) indicate many assessments were judged comparable. Loss rates remain low across all comparisons (16-26%). By dimension (Figure 4), Claim Substantiation and Analytical Quality perform best (56% and 55% win rates). Novelty Decision shows the most ties (31%), suggesting different approaches often yield similar conclusions. These patterns align with our automated results, supporting our evaluation approach's validity.

5.5 Analysis: Understanding Human Alignment Patterns

Our system's higher agreement scores compared to human-human baselines warrant careful examination. To investigate this, we analyzed papers with multiple human reviewers to understand the sources of disagreement.

5.5.1 Sources of Human Reviewer Variability

Qualitative analysis reveals several factors contributing to reviewer disagreement: **Different Evaluation Lenses**: Reviewers often focus on different aspects of novelty. In submission Ipe4fMCBXk, half the reviewers emphasized methodological contributions while others focused on application novelty, leading to opposite conclusions from the same paper. **Varying Domain Expertise**: Reviewers' background knowledge affects assessments. For instance, in a protein design paper, reviewers familiar with the field's history correctly identified prior work on recombination techniques, while others assessed these as novel contributions. **Assessment Granularity**: Some reviewers provide high-level judgments ("innovative approach") while others focus on specific technical details. This variation in granularity contributes to disagreement even when reviewers might agree on underlying facts.

5.5.2 The Role of Systematic Evaluation

Our system's approach differs from human review in applying consistent evaluation criteria. It evaluates multiple dimensions (methodology, application, prior work) for every paper, maintains uniform depth of analysis across assessments, and applies consistent thresholds for novelty judgments. This systematic approach may explain the alignment patterns: when human reviewers disagree due to focusing on different aspects, our system's comprehensive evaluation can align partially with each perspective.

5.6 Component Analysis

Table 5 shows the incremental contribution of each pipeline component. Our human-informed prompt design provides the largest gains (+40.7% reasoning, +46.8% conclusion), reflecting the importance of structured evaluation criteria derived from our human analysis. Structured extraction adds moderate improvements (+3.3% reasoning, +4.5% conclusion) but reduces overall computation costs and time by a lot, while landscape analysis contributes minimally (+3.2% reasoning, -0.7% conclusion).

6 Conclusion

We present a human-informed pipeline for automated novelty assessment in peer review, addressing a critical gap in AI-assisted review systems. Our approach combines systematic related work retrieval with structured evaluation criteria derived from analysis of expert reviewer patterns. Experimental results demonstrate that our system outperforms existing AI baselines and achieves higher agreement rates than human-human comparisons across key evaluation dimensions. The system produces consistently deep analyses with strong prior work engagement, while human evaluations show our assessments are comparable to human reviews in 41% of cases with low loss rates (14-21%) across all dimensions. These findings highlight both the potential for systematic AI assistance in novelty assessment and the inherent subjectivity in human novelty judgments. Our work provides a foundation for more reliable, transparent novelty evaluation in scientific peer review.

Limitations

Despite achieving strong performance, our system has several important limitations:

Evaluation Scope: Our evaluation focuses on computer science papers from ICLR 2025. The system's performance on other scientific domains remains untested and likely requires domain-specific adaptations.

Consistency vs. Diversity: While our analysis shows that systematic evaluation reduces reviewer disagreement, this consistency might eliminate valuable diversity in perspectives. The 35-40% human-human disagreement rate may reflect legitimate differences in expertise and viewpoint rather than mere inconsistency.

Nuanced Novelty: Breakthrough ideas often challenge conventional evaluation criteria. Our system's consistent approach might miss paradigm-shifting contributions that human experts would recognize through intuition or deep domain expertise.

Language Scope: Our study evaluates the system only on English-language manuscripts and reviews. As a result, we cannot claim that the approach generalizes to submissions written in other languages or rooted in different academic conventions; assessing cross-lingual performance remains future work.

References

Balazs Aczel, Barnabas Szaszi, and Alex O. Holcombe. 2021. A billion-dollar donation: estimating the cost of researchers' time spent on peer review. *Research Integrity and Peer Review*, 6(1):14.

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2025. Litllm: A toolkit for scientific literature review. *Preprint*, arXiv:2402.01788.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, and 6 others. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *Preprint*, arXiv:2411.14199.

Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2023. Has the machine learning review process become more arbitrary as the field has grown? the neurips 2021 consistency experiment. *Preprint*, arXiv:2306.03262.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *Preprint*, arXiv:2308.13418.

Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Zhijiang Guo, and Ngai Wong. 2025. Treereview: A dynamic tree of questions framework for deep and efficient llm-based scientific peer review. *ArXiv*, abs/2506.07642.

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

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

Oscar Díaz, Xabier Garmendia, and Juanan Pereira. 2024. Streamlining the review process: Ai-generated annotations in research manuscripts. Preprint available on arXiv.

Neil A. Ernst, Jeffrey C. Carver, Daniel Méndez, and Marco Torchiano. 2020. Understanding peer review of software engineering papers. *Empirical Software Engineering*, 26.

Conghui He, Wei Li, Zhenjiang Jin, Chao Xu, Bin Wang, and Dahua Lin. 2024. Opendatalab: Empowering general artificial intelligence with open datasets. *arXiv preprint arXiv:2407.13773*.

Kelly Hong, Anton Troynikov, and Jeff Huber. 2025. Context rot: How increasing input tokens impacts llm performance. Technical report, Chroma.

Serge P.J.M. Horbach and Willem Halffman. 2018. The ability of different peer review procedures to flag problematic publications. *Scientometrics*, 118:339 – 373.

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.

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.

Peder Olesen Larsen and Markus Ins. 2010. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, 84(3):575–603.

Baolin Li, Yankai Jiang, Vijay Gadepally, and Devesh Tiwari. 2024a. Llm inference serving: Survey of recent advances and opportunities. *2024 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–8.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Deli Zhao, Yu Rong, Tian Feng, and Lidong Bing. 2024b. Chain of ideas: Revolutionizing research via novel idea development with llm agents. *Preprint*, arXiv:2410.13185.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Alexander Nemecek, Yuzhou Jiang, and Erman Ayday. 2025. The feasibility of topic-based watermarking on academic peer reviews. *Preprint*, arXiv:2505.21636.

OpenAI. 2024. Gpt-4.1 (june 2024 version). `https://platform.openai.com/`. Large language model.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. 2025. Scideator: Human-llm scientific idea generation grounded in research-paper facet recombination. *Preprint*, arXiv:2409.14634.

Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Jashim Sakib, Most. Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839–26874.

Simra Shahid, Marissa Radensky, Raymond Fok, Pao Siangliulue, Daniel S Weld, and Tom Hope. 2025. Literature-grounded novelty assessment of scientific ideas. *arXiv preprint arXiv:2506.22026*.

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

Weiwei Sun, Zheng Chen, Xinyu Ma, Lingyong Yan, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023a. Instruction distillation makes large language models efficient zero-shot rankers. *ArXiv*, abs/2311.01555.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023b. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024. Mineru: An open-source solution for precise document content extraction. *Preprint*, arXiv:2409.18839.

Dalong Zhang, Jun Xu, Jun Zhou, Lei Liang, Lin Yuan, Ling Zhong, Mengshu Sun, Peilong Zhao, QiWei Wang, Xiaorui Wang, Xinkai Du, Yangyang Hou, Yu Ao, ZhaoYang Wang, Zhengke Gui, ZhiYing Yi, Zhongpu Bo, Haofen Wang, and Huajun Chen. 2025. Kag-thinker: Interactive thinking and deep reasoning in llms via knowledge-augmented generation. *ArXiv*, abs/2506.17728.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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 Human Evaluation Protocol: Novelty Assessment Comparison

A.1 Task Design

We conducted a comparative evaluation where human evaluators assessed the quality of AI-generated novelty assessments against expert-written reference assessments. Each evaluator compared pairs of AI-generated assessments (labeled A and B) to a human expert's gold-standard novelty review of the same research paper.

A.2 Evaluation Framework

A.2.1 Materials Provided

For each evaluation, evaluators received: (1) an expert-written gold-standard novelty review as reference, (2) two novelty assessments (A and B) with system identities hidden.

A.2.2 Evaluation Dimensions

Evaluators assessed each pair across four dimensions:

1. Reasoning Alignment: Which assessment better captures the key novelty reasoning from the reference? Evaluators considered similarity of novelty claims, logical arguments, and focus areas.

2. Decision Alignment: Which assessment reaches a novelty verdict most consistent with the reference? This included agreement on overall judgment (novel/incremental/mixed) and similar weighting of novelty factors.

3. Claim Substantiation: Which assessment better supports its novelty claims with evidence? Evaluators looked for specific citations, concrete examples from the paper, and absence of unsupported generalizations.

4. Analytical Quality: Which assessment provides more insightful technical analysis of novelty? This considered depth of technical discussion, specificity of analysis, and balanced consideration of strengths and limitations.

For each dimension, evaluators selected one of four options: *A wins*, *B wins*, *Tie* (both equally good/poor), or *Unclear* (cannot determine).

A.3 Evaluation Guidelines

A.3.1 Instructions for Evaluators

Evaluators were instructed to read the reference assessment thoroughly before evaluating A and B, evaluate each dimension independently, and base judgments on substantive content rather than stylistic differences. They allocated 4–7 minutes per example to ensure thorough evaluation and flagged ambiguous cases with explanatory comments when necessary.

A.3.2 Evaluation Focus

Evaluators were instructed to **prioritize** substance and accuracy of novelty reasoning, alignment with reference judgments (particularly for Dimensions 1–2), quality and depth of technical analysis (particularly for Dimensions 3–4), and specific evidence and citations supporting claims. They were instructed to **disregard** writing style, grammar, or formatting differences; suggestions for paper improvement unrelated to novelty; minor phrasing variations with equivalent meaning; and length differences if content quality was comparable.

A.4 Implementation Details

A.4.1 Evaluation Platform

The evaluation was conducted through a custom web interface presenting materials in a standardized format (see Figures 6 and 7). Each evaluator received a unique evaluator ID, 50 randomly assigned paper-assessment pairs, and the ability to save progress and flag unclear cases.

A.4.2 Quality Control

We calculated inter-evaluator agreement using Cohen's kappa reported in Table 6.

A.4.3 Data Collection

Completed evaluations were submitted as structured JSON files containing dimension-wise selections (A/B/Tie/Unclear), time spent per evaluation, and comments for flagged cases.

Category	Agreement	Kappa	Comparisons
Novelty Reasoning Alignment	0.520	0.341	75
Novelty Decision Alignment	0.533	0.346	75
Claim Substantiation	0.493	0.287	75
Analytical Quality	0.560	0.368	75

Table 6: Inter-Rater Reliability Metrics Across Categories

B Output Examples

Output of our pipeline can be seen in Tables 7, 8 and 9. It is quite evident that our system aligns better with the human as compared to the baselines across all four dimensions.

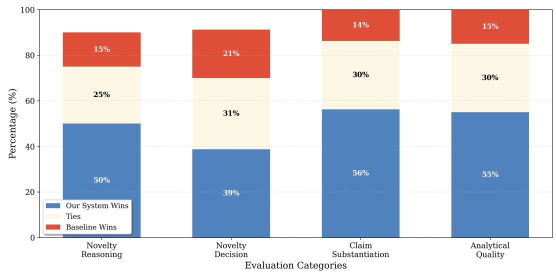

Figure 4: Performance breakdown across evaluation categories, aggregated across all baseline comparisons.

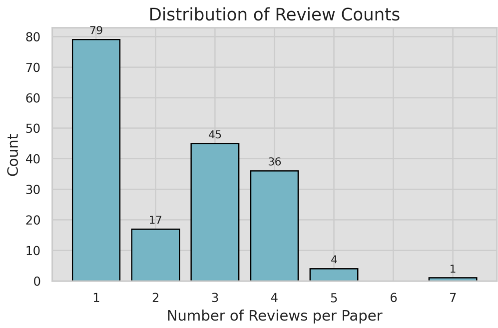

Figure 5: Distribution of the number of reviews per paper. Most papers received 1 to 4 reviews.

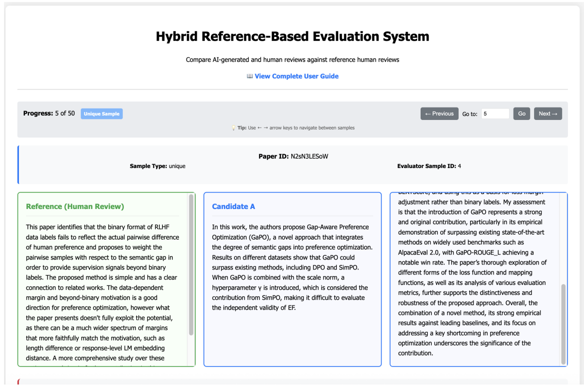

Figure 6: Screenshot of the custom-built interface used for human evaluation. Annotators compared AI-generated and human-written novelty assessments across multiple dimensions, including reasoning depth, prior work engagement, and conclusion alignment.

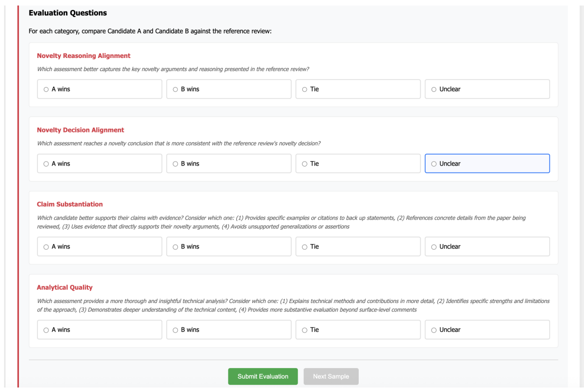

Figure 7: Screenshot (2) of the custom-built interface used for human evaluation. Annotators compared AI-generated and human-written novelty assessments across multiple dimensions, including reasoning depth, prior work engagement, and conclusion alignment.

Human (Reference)	Scideator (Baseline)	Ours (Proposed)
The proposed approach aims to significantly reduce memory consumption during long-sequence inference while maintaining model performance and requiring minimal modifications to existing LLM frameworks. While the integration of CGE and RGL is presented as a novel solution to the KV cache memory challenge, the overall novelty of the work is limited. Sparse attention mechanisms have already been extensively explored in prior literature, such as [1] and [2], which diminishes the originality of the proposed methods. Furthermore, the CGE component closely resembles previous approaches like H20 [3] and SnapKV [4], with only incremental differences. The feedback mechanism using sliding windows has also been well-studied in H20 [3]. As such, the main contribution appears to be a specific combination and implementation of existing ideas rather than a fundamentally new technique. The paper would benefit from a clearer articulation of how its methods differ from these established approaches to better establish its unique contribution.	`- Class: novel` The idea is novel because IntelLLM introduces a distinct approach to key-value (KV) cache compression by using strategies like center of gravity eviction (CGE) and remote gap localization (RGL), which are not seen in existing works such as RazorAttention[o] and LeanKV[1]. These methods prioritize retaining essential tokens and leveraging positional features to enhance compression efficiency, offering a dedicated approach with unique techniques for balancing compression rate and performance.	IntelLLM introduces new heuristics—Center of Gravity Eviction (CGE) and Remote Gap Localization (RGL)—for token selection and long-range dependency preservation in KV cache compression for LLM inference, but these are incremental variants of established token eviction approaches. The submission overstates its novelty, as the core ideas (token selection, attention sparsity, training-free deployment) are already well-explored, and similar methods (e.g., RazorAttention, PyramidKV, L_2 Norm) achieve comparable goals without model changes or fine-tuning. Several highly relevant recent works are omitted from the discussion, and the claims of being the first to balance compression and performance or to preserve long-range dependencies are not substantiated by the literature. The main technical delta lies in the specific heuristics (CGE, RGL) and their empirical performance, rather than in a conceptual advance. Reviewers should view IntelLLM as a routine, incremental contribution and may wish to request more comprehensive comparisons and a more accurate positioning within the current research landscape.

Table 7: Full novelty assessments from the human reviewer (reference), the Scideator baseline, and our proposed system. Key phrases are highlighted to show verdict alignment: positive novelty claims, limited/incremental novelty, comparative analysis, and critical issues.

Research Paper Information Extraction Prompt

```
You are tasked with extracting key information from a research paper
for building a knowledge representation.
Paper title: {title}
Based on the paper content provided below, extract the following information:
- "methods": [List of methods/approaches proposed in the paper],
- "problems": [List of problems the paper addresses],
- "datasets": [List of datasets used for evaluation],
- "metrics": [List of evaluation metrics used],
- "results": [List of objects with 'metric' and 'value' fields
  representing key quantitative results],
- "novelty_claims": [Claims about what is novel in this work]
Be precise and specific.
Paper content:
{abstract}
{introduction}
```

Figure 8: Research Paper Information Extraction Prompt

Human (Reference)	**DeepReviewer (Baseline)**	**Ours (Proposed)**
The paper "DuRND: Rewarding from Novelty to Contribution for Reinforcement Learning via Dual Random Networks Distillation" proposes an extension to classical RND by introducing two distinct random network modules—one for states deemed "successful" and another for states associated with "failure." This innovation allows for the derivation of both a "novelty" and a "contribution" reward signal, striking a balance between exploratory and exploitative behavior. While the additional novelty introduced by DuRND is incremental compared to classical RND, I still believe the contribution is valuable and fills a gap in the current literature.	This paper proposes a novel framework, Dual Random Networks Distillation (DuRND), to address the exploration-exploitation dilemma in reinforcement learning by combining novelty-based exploration and value-based exploitation in a unified framework. The novelty and contribution rewards in DuRND are based on prediction errors from random network modules, which may not capture complex state relationships in certain environments, and the reliance on prediction error as a proxy for novelty might be insufficient where the state space has a complex structure or the prediction error does not align with actual information gain. The approach introduces an interesting combination of novelty and contribution rewards but relies on computationally efficient heuristics that might not capture the true novelty in complex environments. While the framework demonstrates effectiveness and low computational overhead, the linear adjustment of weights for novelty and contribution rewards is a heuristic that may not be optimal or universally applicable. Overall, the originality lies in the integration of novelty and value-based signals, but there remain significant limitations in the robustness and generalizability of the contribution due to dependency on specific proxy measures and heuristic schedules.	DuRND is a direct extension of Random Network Distillation (RND), introducing a dual random network architecture to compute both novelty (exploration) and "contribution" (exploitation) rewards, with a dynamic scheduling mechanism that shifts emphasis over time. While the use of random networks for both signals is a novel implementation detail, the underlying concepts—combining exploration and exploitation signals and dynamically adjusting their balance—are well-established in the literature (e.g., SORS, ROSA, Sibling Rivalry, SERENE). The authors accurately position DuRND as a low-overhead, scalable method, but somewhat overstate its conceptual novelty and the dichotomy between prior approaches. Empirical results show improved performance on standard sparse-reward benchmarks, though such gains may be attributable to implementation choices rather than fundamental innovation. Overall, DuRND offers a practical and well-engineered architectural tweak to RND, with its main contribution being a simple, dual-network design for joint intrinsic reward computation, rather than a conceptual breakthrough.

Table 8: Full novelty assessments from the human reviewer (reference), the DeepReviewer baseline, and our proposed system for the DuRND paper. Key phrases are highlighted to show verdict alignment: novel/valuable aspects, limited/incremental novelty, technical details, and overstated claims.

Human (Reference)	**OpenReviewer (Baseline)**	**Ours (Proposed)**
This paper introduces a method for generating image inputs to Vision Language Models (VLMs) that embed "meta-instructions," enabling the model to respond to any user query appended to the image with a specified "spin," such as a particular sentiment or language. While the concept of providing meta-instructions through image inputs is not entirely novel—prior work, notably Bailey et al. (2023), has explored related prompt injection attacks via images, this paper offers the most thorough and comprehensive treatment of the subject to date. The general algorithm employed, which involves creating a dataset of input-output pairs and training an image via projected gradient descent to maximize the likelihood over this dataset, closely aligns with the "Behavior Matching" approach described by Bailey et al. (2023). As such, the methodological contribution is incremental rather than fundamentally new. However, the paper distinguishes itself through its in-depth analysis and experimental results, particularly the study of semantic changes in images resulting from various attacks, with a focus on how meta-instruction attacks can preserve the original image meaning. The transferability experiments and the breadth of evaluation across multiple VLMs further strengthen the contribution. Overall, the novelty of this work lies less in the method itself and more in the scope, depth, and clarity of its empirical investigation, as well as the new insights it provides into the vulnerabilities of VLMs to image-based prompt injection. The paper would benefit from more explicitly acknowledging the overlap with prior methods and more clearly articulating how its results advance the understanding of this threat model.	This paper introduces a novel method of embedding hidden meta-instructions in images as a way to attack visual language models, which is distinct from traditional jailbreaking or adversarial example attacks. While the approach is new, the paper could provide more insight into how the proposed attacks compare to existing forms of adversarial attacks on VLMs, highlighting the unique aspects of meta-instruction attacks.	This assessment finds that the submission addresses indirect, cross-modal prompt injection in Visual Language Models (VLMs) by embedding hidden meta-instructions in images, aiming to steer model outputs while preserving image semantics. The work is most closely related to recent studies on adversarial image prompting (e.g., Qi et al. 2024, Bagdasaryan et al. 2023), but distinguishes itself through more systematic optimization for semantic preservation and a broader range of meta-instructions beyond jailbreaking. The main substantive contributions are a rigorous, multi-metric evaluation of attack effectiveness and semantic preservation, and empirical evidence that image-based meta-instructions can be more effective than explicit text prompts. However, the assessment notes that the conceptual advances are incremental, as the core idea of cross-modal prompt injection and semantic preservation has been explored in prior work and some novelty claims (e.g., being the first to frame VLM users as victims) are somewhat overstated. Overall, the submission's primary strengths lie in evaluation rigor and empirical findings, while its conceptual contributions represent a natural progression of the field rather than a fundamental shift.

Table 9: Full novelty assessments from the human reviewer (reference), the OpenReviewer baseline, and our proposed system for the Meta-Instructions in VLMs paper. Key phrases are highlighted to show verdict alignment: novel/strength claims, limited/incremental novelty, prior work comparison, and overstated claims.

Research Landscape Analysis

```
# Research Landscape Analysis

## Task
Analyze the collection of research papers provided below to create a
comprehensive map of the research landscape they represent. The submission
paper is the focus of our analysis, and the related papers provide context.

## Input Format
You will be provided with structured information extracted from multiple
research papers including:
- A submission paper that is the focus of our analysis
- Multiple related papers that form the research context

Each paper contains:
- methods: List of methods/approaches proposed
- problems: List of problems addressed
- datasets: List of datasets used
- metrics: List of evaluation metrics
- results: Key quantitative results
- novelty_claims: Claims about what is novel in the work

## Output Format
Provide a comprehensive analysis with the following sections:

1. METHODOLOGICAL LANDSCAPE
   - Identify and describe the main methodological approaches across the papers
   - Group similar or related methods into clusters
   - Highlight methodological trends or patterns
   - Describe relationships between different methodological approaches

2. PROBLEM SPACE MAPPING
   - Identify the key problems being addressed across the papers
   - Analyze how different papers approach similar problems
   - Highlight patterns in problem formulation

3. EVALUATION LANDSCAPE
   - Analyze the common datasets and evaluation methods
   - Identify patterns in how performance is measured
   - Compare evaluation approaches across papers

4. RESEARCH CLUSTERS
   - Identify groups of papers that appear closely related
   - Describe the key characteristics of each cluster
   - Analyze relationships between clusters

5. TECHNICAL EVOLUTION
   - Identify any visible progression or evolution of ideas
   - Highlight building blocks and their extensions
   - Note any competing or complementary approaches

## Example Output Format
METHODOLOGICAL LANDSCAPE
- Cluster 1: [Description of similar methods across papers]
  - Papers X, Y, Z employ transformer-based approaches with variations in...
  - These methods share characteristics such as...
  - They differ primarily in...

PROBLEM SPACE MAPPING
- Problem Area 1: [Description of a common problem addressed]
  - Papers A, B, C all address this problem but differ in...
  - The problem is formulated differently in Paper D which focuses on...

... [additional sections] ...

Ensure your analysis is comprehensive, identifying significant patterns
and relationships across the collection of papers.

## Papers:
{papers}
```

Figure 9: Research Landscape Analysis Prompt

Novelty Delta Analysis for Reviewer Support - Part 1

```
# Novelty Delta Analysis for Reviewer Support

## Task
Independently analyze how the submission paper's contributions relate to existing
work in the field, critically examining both author claims and actual relationships.
This analysis should help reviewers assess novelty by providing objective comparisons
with prior work.

## Input Format
You will be provided with:
1. The structured information from the submission paper
2. A comprehensive research landscape analysis
3. Citation sentences for key related papers (how authors cite and characterize these works)

## Key Analysis Principles
- Independently verify relationships between submission and prior work
- Critically examine how authors characterize and compare with prior work
- Identify discrepancies between author characterizations and actual relationships
- Present evidence-based observations without making final judgments
- Distinguish between author-claimed differences and independently observed differences
- Provide context about field maturity and related work

## Output Format
Provide a detailed analysis with the following sections:

1. RESEARCH CONTEXT POSITIONING
   - Situate the submission within the identified research landscape
   - Identify the most closely related prior works
   - Independently assess how the submission relates to existing methodological clusters
   - Analyze its place within the problem space and evaluation approaches
   - Note: Do not accept author positioning claims without verification

2. AUTHOR CITATION ANALYSIS
   - Analyze how authors characterize and compare with each cited related work
   - Identify patterns in how authors position their contributions relative to others
   - Assess whether characterizations of prior work are accurate and balanced
   - Note discrepancies between how authors describe prior work and independent assessment
   - Evaluate whether claimed improvements or differences are substantiated
   - Identify rhetoric that may overstate differences or understate similarities

3. CONTRIBUTION DELTA ANALYSIS
   For each main contribution claimed in the submission:
   - Identify the most similar prior work for this specific contribution
   - Critically examine whether claimed differences actually exist
   - Detail exactly how this contribution differs from prior work, based on evidence
   - Compare author characterizations with independently verified relationships
   - Distinguish between substantive differences and superficial variations
   - Note when author claims about novelty or extension may be overstated
   - Consider whether improvements might be due to implementation details rather than
     conceptual advances
   - Note: Present factual observations about deltas without accepting author framing

4. FIELD CONTEXT CONSIDERATIONS
   - Provide information about how active/mature this research area is
   - Identify recent survey papers or literature reviews in this space
   - Note trends in how the field has been evolving
   - Present context about typical incremental advances in this field
   - Note: Offer context that helps reviewers calibrate their expectations
```

Figure 10: Novelty Delta Analysis for Reviewer Support - Part 1

Novelty Delta Analysis for Reviewer Support - Part 2

```
5. CRITICAL ASSESSMENT CONSIDERATIONS
   - Identify aspects where claimed novelty may be overstated
   - Analyze whether authors' characterizations of their own novelty align with evidence
   - Consider whether empirical improvements might result from factors other than claimed
     innovations
   - Assess whether terminology differences might mask conceptual similarities
   - Identify instances where "extensions" might be routine adaptations
   - Note: Frame these as considerations rather than definitive judgments

6. RELATED WORK CONSIDERATIONS
   - Identify potentially relevant work not addressed in the submission
   - Highlight areas where additional comparisons are necessary
   - Note incomplete or potentially misleading characterizations of prior work
   - Identify when claimed "limitations" of prior work may be exaggerated
   - Compare how authors cite specific works versus how they actually relate
   - Note: Present these as information that might help complete the picture

7. KEY OBSERVATION SUMMARY
   - Highlight the most significant independently verified differences from prior work
   - Summarize the main relationships to existing research
   - Identify which claimed contributions have the strongest and weakest differentiation
   - Note the most important discrepancies between author characterizations and
     independent assessment
   - Note: Frame as observations to inform the reviewer's independent judgment

## Evidence Standards
For each observation, provide:
- Specific references to prior work
- Clear distinction between author claims and independently verified differences
- Explicit identification of similarities and differences based on technical details
- Assessment of whether differences appear substantive or superficial
- Analysis of accuracy in how authors characterize related work

## Example Format for Citation Analysis
"For [Paper X], the authors characterize it as 'limited to simple datasets' and claim
their work 'extends X to complex scenarios.' The citation sentences appear in the
following contexts:
- 'Unlike X, which only works on simple datasets, our approach handles complex
  scenarios' (Introduction)
- 'X proposed the basic framework, but did not address challenge Y' (Related Work)

Independent analysis suggests that Paper X actually did address complex scenarios
in Section 3.2, though using different terminology. The authors' characterization
appears to understate X's capabilities to emphasize their contribution. The actual
primary difference appears to be [specific technical difference] rather than the
complexity of supported scenarios."

Remember that your role is to provide objective analysis that helps reviewers make
informed judgments about novelty. Carefully examine both what authors explicitly
claim and how they implicitly position their work through their characterizations
of prior research.

{structured_representation}

## Papers from related work not cited
{not_cited_paper_titles}

##Citation Context
{citation_contexts}

## Research Landscape
{research_landscape}
```

Figure 11: Novelty Delta Analysis for Reviewer Support - Part 2

Reviewer Summary Prompt

```
Summarize the following assessment in 5 sentences for a reviewer reviewing at an AI conference.


## Delta Assessment
{novelty_assessment}
```

Figure 12: Reviewer Summary Prompt

Novelty Assessment Normalization Prompt

```
I'll provide you with a novelty assessment extracted from an academic peer review,
along with the full review for context. Please reformat the novelty assessment into
a standardized paragraph that begins with a brief description of the paper's contribution
before analyzing its novelty.

Example of desired format:
"This paper presents a method for neural network compression using knowledge distillation
with a focus on mobile applications. The approach has limited novelty, as it largely
builds upon existing techniques in the literature. While the authors claim their technique
is the first to combine layerwise distillation with quantization-aware training, similar
combinations have been explored in prior work by Smith et al. (2022) and Jones et al. (2023).
The main contribution appears to be a specific implementation detail in how gradient flows
are managed during the distillation process, but this incremental advance does not
significantly push the boundaries of the field. The paper would benefit from more clearly
articulating the specific differences from existing approaches to better establish its
    contribution."

Full review (for context):
{full_review}

Extracted novelty assessment to be reformatted:
{novelty_statements}

Important guidelines:
1. Begin with a clear description of what the paper presents/proposes (drawn from the full
    review if needed)
2. Create a cohesive paragraph that flows from describing the contribution to analyzing its
    novelty
3. Maintain all novelty claims and critiques from the original assessment
4. Preserve references to prior work and comparisons
5. Keep the reviewer's judgment of novelty level
6. Incorporate relevant context from the full review to provide a complete picture of the
    novelty assessment
7. Follow the structure of the example paragraph: description first, then novelty analysis
8. Preserve all critical analysis regarding limitations or strengths of novelty claims

Provide the reformatted novelty assessment:
```

Figure 13: Novelty Assessment Normalization Prompt

Core Novelty Judgment Extraction Prompt

```
Extract 2-3 core novelty judgments from this assessment:

{reference_assessment}

Focus on statements that directly assess:
- How novel/original the contribution is
- How work relates to prior research
- Specific novelty limitations
- Whether advance is incremental/fundamental

Exclude general recommendations or writing suggestions.

For each judgment, explain why it's considered a core novelty assessment.
Provide rationale for your selection of these specific judgments.
```

Figure 14: Core Novelty Judgment Extraction Prompt

Reviewer Novelty Evaluation Prompt

```
Compare reviewer assessment against reference using these core judgments:

Core Judgments: {extracted_core_judgments}
Reference: {reference_assessment}
Reviewer: {reviewer_assessment}

Evaluate three dimensions:

1. JUDGMENT SIMILARITY: Do they identify same novelty strengths/weaknesses?
   - For each core judgment, find corresponding judgment in reviewer assessment
   - Assess similarity and provide detailed explanation of alignment/differences
   - Include confidence score for each comparison
   - If the core judgement is referring to a very specific aspect of the methodology
     and the reviewer assessment does not mention it, then the core judgment is
     not similar to the reviewer assessment.

2. CONCLUSION ALIGNMENT: Same bottom-line about novelty sufficiency?
   - Determine overall conclusions (SUFFICIENT / INSUFFICIENT / MIXED)
   - Explain whether conclusions align and why

3. PRIOR_WORK_ENGAGEMENT:
   - How does the reviewer engage with prior work?
   - Does the reviewer mention prior work?
   - Does the reviewer compare the current work to prior work?
   - Does the reviewer provide evidence for their claims?
   - Does the reviewer use prior work to support or critique the work?
   - Evaluate number and relevance of citations to prior work
     (NONE: no citations; LIMITED: 1 to 2; EXTENSIVE: 3+ relevant citations).

4. DEPTH_OF_ANALYSIS:
   - Assesses how deeply specific novelty aspects are compared to prior work
     (SURFACE LEVEL: vague; MODERATE: 1 to 2 aspects; DEEP: 3+ or highly detailed comparisons)

Provide explanations for all assessments to support reasoning.
```

Figure 15: Reviewer Novelty Evaluation Prompt