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

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

[1] UKP Lab, TU Darmstadt and Hessian Center for AI (hessian.AI)
[2]MBZUAI, [3]Hebrew University of Jerusalem and The Allen Institute for Al (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: (*i*) content extraction from submissions, (*ii*) retrieval and synthesis of related work, and (*iii*) structured comparison for evidence-based assessment. Our method is informed by 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. It produces detailed, literature-aware analysis 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. The data and the code are available at https://ukplab.github.io/eacl2026-assessing-paper-novelty/

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), revealing a level of inconsistency that points to structural issues beyond simple reviewer overload. As manuscript submissions double roughly every 15 years (Larsen and von Ins, 2010) and individual reviewers now complete an average of 14 reviews per year (Díaz et al., 2024), the system expends an estimated 15 million hours annually on peer review alone (Aczel et al., 2021)—yet it delivers outcomes that are increasingly noisy, fragile, and difficult to justify.

Among peer review tasks, novelty assessment stands out as one of the most problematic ones (Ernst et al., 2021; Horbach and Halffman, 2019). Reviewers must determine whether a submission makes sufficiently original contributions. This them to identify what specific advances it makes beyond existing work, evaluating whether these advances are significant enough to warrant publication, and judging whether 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, which is becoming exponentially more difficult as publication rates accelerate and research domains specialize. Overwhelmed reviewers often resort to superficial analysis, 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 novelty assessment challenges at scale. These technologies have revolutionized text processing and demonstrated strong performance on knowledge-intensive tasks (Raiaan et al., 2024), such as literature understanding, synthesis, and comparative analysis across large corpora, with ongoing technical advances extending their capabilities toward specialized reasoning, improved factual grounding, and more efficient inference (Li et al., 2024; Zhang et al., 2025).

Despite this progress, novelty assessment has not yet been addressed as a dedicated task within the peer review process. Existing work instead incorporates novelty assessment indirectly.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2648–2671

Some approaches embed novelty assessment within research idea generation pipelines (Radensky et al., 2024; Lu et al., 2024; Li et al., 2025), while others generate peer reviews in which novelty judgments emerge implicitly from models trained on reviewer data (Idahl and Ahmadi, 2025; D'Arcy et al., 2024), or introduce novelty-related steps as part of broader review synthesis pipelines (Zhu et al., 2025). However, these approaches either operate on synthetic ideas rather than completed research contributions, limiting their applicability to real submissions, or do not evaluate novelty assessment capabilities in isolation, making it difficult to assess their reliability and generalizability. This leaves a critical gap, motivating the need for specialized methodologies for novelty assessment in peer review.

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 ideation stage of research (Radensky et al., 2024) (Shahid et al., 2025) (Li et al., 2025) (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.

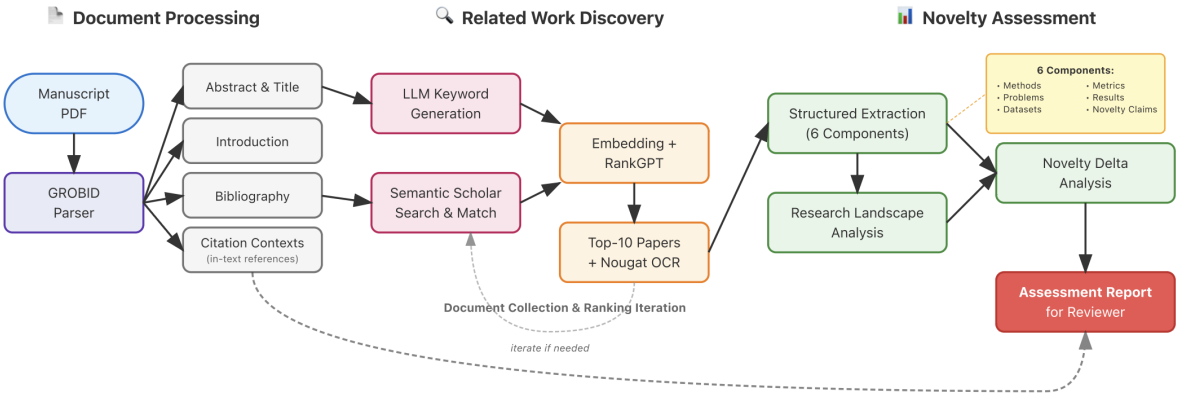

Figure 1: Automated novelty assessment pipeline. The system processes manuscripts through three stages: (1) Document Processing extracts content using GROBID, (2) Related Work Discovery identifies and ranks relevant papers via embedding similarity and LLM reranking, and (3) Novelty Assessment performs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

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., 2024)(Shahid et al., 2025)(Li et al., 2025) 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 adopted an LLM-as-a-Judge approach for our evaluation setup, following recent best practices in LLM evaluation for complex reasoning tasks. 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.

Dataset Scale: Our dataset comprises 182 papers and 352 reviews, which is comparable to or larger than datasets used in related peer review research: (Du et al., 2024) use 100 papers with 380 reviews, (Kennard et al., 2022) label review data sourced for 188 papers, (Hua et al., 2019) work with 400 reviews, and (Chamoun et al., 2024) evaluate on 300 reviews. Novelty assessment requires careful manual annotation of scattered novelty discussions across reviews, making large-scale annotation resource-intensive. Following established practice in peer review analysis, we prioritize annotation quality over quantity, ensuring each example receives thorough annotation.

3 Methodology

3.1 Human Analysis for Prompt Design

To understand human novelty assessment, we analyzed ICLR 2025 reviews, which explicitly require novelty evaluation in dedicated sections. We sourced submissions from OpenReview and used keyword-based search for terms including "novel", "original", "research gap", "innovation", "incremental", "prior work", and "existing work". We selected the top 200 papers ranked by: (1) reviews with >4 novelty keywords, (2) consistent novelty discussion patterns, and (3) total review count.

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.

To expedite annotation, we used GPT-4o mini for sentence-level classification to identify novelty discussions, which a human annotator then verified and refined by selecting all sentences containing actual novelty assessments. We found that 18 papers (9%) contained limited genuine novelty assessments—often keyword matches referring to paper components rather than evaluation. The remaining 182 papers formed our analysis dataset, with statistics by decision category reported in Table 1. We identified recurring patterns in reviewer reasoning, evaluation criteria, and argument structures, focusing on how reviewers structure arguments, prioritize evidence, and compare submissions to prior work. The four key patterns were identified through an exploratory qualitative review, where the primary author examined novelty-related review segments, allowing patterns to emerge inductively through close reading and thematic analysis. 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 8 and 9.

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 9.)

Different analytical lenses: Some reviewers focused on methodological innovations while others evaluated the systems holistically, calibrating the expectations based on field's 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 8 and 9.

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 of the prompt explicitly instructs models to identify missing comparisons and assess whether improvements stem from "implementation details rather than conceptual advances," directly reflecting this reviewer behavior (Figure 9). 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 1): (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[1] 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.

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

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., 2023) on concatenated titles and abstracts. Papers are ranked by cosine similarity 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., 2024) 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 the retrieved papers as raw text causes context optimization challenges that degrade LLM performance. Recent research has demonstrated that model performance consistently degrades with 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., 2024).

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 6).

Landscape Analysis Expert reviewers possess comprehensive domain knowledge of established benchmarks, techniques, metrics, and recent developments in their areas. To approximate this foundation, we incorporate a landscape analysis step that systematically organizes structured components from retrieved related work. Using GPT-4.1, we perform cross-paper synthesis to identify methodological clusters, trace problem evolution, map evaluation ecosystems, and establish technical relationships (Figure 7). This produces a hierarchical organization of the research space with explicit connections between related, competing, and complementary approaches, providing contextual background for novelty assessment that mimics expert reviewers' organized domain understanding.

Novelty Delta Analysis This step performs comparative analysis between the submission and prior work using the following 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 8 and 9).

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 10).

4 Evaluation

4.1 Evaluation Data

The evaluation dataset comprises the same 182 annotated examples used during human prompt design. For each example, we prompt GPT-4.1 with the human review and its corresponding annotated novelty statements to generate a coherent assessment using the prompt in Figure 11.

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., 2024)	23.7 ± 0.00	22.4 ± 0.00	0.0 ± 0.00	20.5 ± 0.00
Ours (GPT-4.1)	**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.

This synthesis step is necessary because novelty-related comments in reviews are typically scattered rather than consolidated. Direct concatenation of these fragments would introduce stylistic biases during evaluation, as the fragmented human annotations would differ substantially in structure and coherence from the unified assessments generated by our system. We therefore use the GPT-4.1-synthesized assessments as our ground truth in evaluation. To assess the risk of potential data leakage into GPT-4.1's pre-training corpus, we examined our evaluation set of 182 ICLR 2025 submissions. Only 11 of these papers had appeared on arXiv prior to the model's knowledge cutoff of June 1, 2024.

4.2 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 12 and 13 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 12. 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 13. 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 We validate our automated evaluation using three PhD students (two third-year and one first-year) in NLP and AI for Science, all with multiple publications. Annotators perform pairwise comparisons across the same four evaluation dimensions, viewing side-by-side novelty assessments produced by humans, our system, and the baselines. We collect 100 total judgments: 25 shared samples per annotator to measure agreement, and 25 unique samples per annotator, sampled randomly.

For each comparison, annotators choose A, B, Tie, or Unclear, and may optionally provide comments (Figures 4 and 5 illustrate the annotation interface). Table 7 reports moderate inter-rater agreement (0.493–0.560) and fair kappa scores (0.287–0.368), consistent with the subjective nature of novelty evaluation.

4.3 Baseline Methods

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

Scideator (Radensky et al., 2024) 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 iterative idea refinement, we adapt it by using paper titles and abstracts as input instead of nascent ideas.

OpenReviewer (Idahl and Ahmadi, 2025) OpenReviewer generates comprehensive peer reviews using Llama-OpenReviewer-8B, trained on 79,000 expert reviews from top conferences. We extract novelty-related content from its outputs using the same LLM-based approach applied to human reviews (Figure 11).

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 using the same approach as OpenReviewer. Notably, DeepReviewer was trained on ICLR 2025 data, which includes our **entire evaluation** dataset.

Adaptation of Baselines Our groundtruth novelty assessments are extracted from human written reviews from the ICLR 2025 set that we labeled earlier. These extracted novelty segments are then run via the style normalization prompt in figure 11 to ensure consistent structure, tone, and length across all evaluation examples. Similarly both the DeepReviewer and OpenReviewer produced peer reviews are run via this pipeline to extract novelty segments and compose a coherent novelty assessment and run through the style normalization module. This is a fair adaptation as given these models are trained on these peer review datasets and they should able to mimic the distribution of novelty assessments found in these peer reviews. Given deepreviewer is trained on the data we are evaluating so even reproducing their training data would be enough to score high on the evaluation set.

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 (in %).

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 (in %).

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. Appendix Tables 9, 10, and 11 present qualitative, side-by-side output comparisons between our system and the baselines.

5.1 Overall Performance

Our system outperforms all AI and human baselines. It improves Reasoning Alignment by 44.1 and 35.9 percentage points over OpenReviewer (Idahl and Ahmadi, 2025) and DeepReviewer (Zhu et al., 2025), respectively, and by 21.4 points over human–human agreement. It also leads on Conclusion Agreement, exceeding the closest human baseline by 13 percentage points.

Sentiment Shift Analysis We analyze Positive Shift (neutral/negative $\rightarrow$ positive sentiment vs. human reference) and Negative Shift (the opposite). We see optimistic bias, with DeepReviewer exhibiting high Positive Shift. Our system shows lower Positive Shift than DeepReviewer, though OpenReviewer aligns most closely with human rates. For Negative Shift, OpenReviewer mirrors humans' critical tendency, followed by DeepReviewer. Our system has the lowest Negative Shift rate.

Depth and Prior Work Engagement Our system is the best for both dimensions, producing no surface-level analysis unlike all baselines (tables 3 and 4). This is because we target novelty assessment directly, while other systems generate complete peer reviews where novelty is secondary.

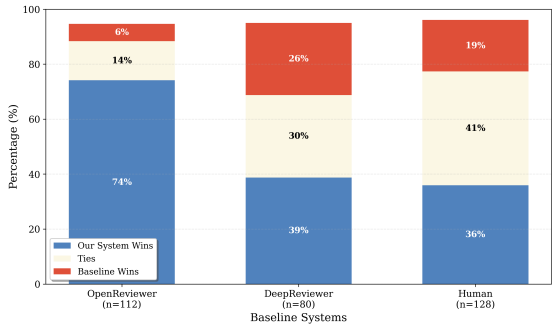

Figure 2: Overall performance comparison between our system and three baseline systems based on human evaluation (n values indicates number of comparisons).

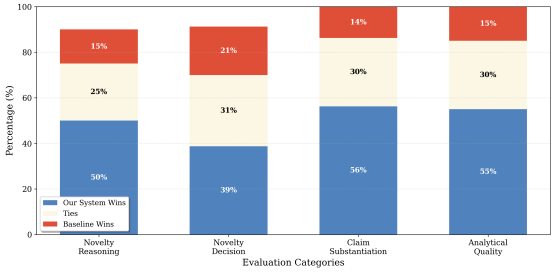

Figure 3: Performance breakdown across evaluation categories, aggregated across all baseline comparisons.

OpenReviewer performs worst, lacking retrieval capabilities. DeepReviewer uses OpenScholar retrieval but fails at comparative analysis. Human reviewers show high variance in engagement depth.

Human Evaluation Validation Human evaluations validate our LLM-as-a-Judge framework. Our system wins 74% of comparisons against OpenReviewer (Figures 2 and 3). Against DeepReviewer and human reviewers, win rates are lower (39% and 36%), but high tie rates (30% and 41%) indicate comparable quality, with low loss rates (16-26%). By dimension, Claim Substantiation and Analytical Quality achieve the highest win rates (56% and 55%), while Novelty Decision shows the most ties (31%), suggesting similar conclusions across approaches. This aligns with automated results, supporting our evaluation framework's validity.

5.2 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. We detail the analysis methodology in Appendix D.

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. In a protein design paper, reviewers familiar with the field's history correctly identified prior work on recombination techniques, while others viewed them as novelty.
- **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.

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.

5.3 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) while reducing computation cost and time, whereas landscape analysis contributes minimally (+3.2% reasoning, -0.7% conclusion). The major improvements come from prompt design, showing that careful prompting can outperform more complex methods requiring extensive training.

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.

We evaluate component contributions in our retrieval pipeline on 100 ICLR submissions (Table 6). The full pipeline combines keyword search with cited papers, ranks results by SPECTER2 embedding similarity, and applies GPT-3.5 reranking. Without LLM reranking (embeddings only), we achieve 71% overlap at top-10 with the full pipeline. When considering only keyword search (excluding citations), overlap drops to 32%, indicating that author citations provide crucial relevance signals beyond keyword matching.

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.

Our approach demonstrates that careful prompt design can achieve strong performance without requiring extensive model training. This is a strength of our method. While methods that attempt training (Zhu et al., 2025; Idahl and Ahmadi, 2025) require substantial computational resources (e.g., 8× H100 80G GPUs for 23,500 steps at 256K context), whereas our prompt-based approach achieves strong performance while offering greater computational efficiency.

Limitations

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

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

Method	Top-5	Top-10	Top-15	Top-20
Full Pipeline	1.000	1.000	1.000	1.000
Dense Only	0.612 ± 0.20	0.710 ± 0.18	0.748 ± 0.12	0.766 ± 0.10
KW Only	0.300 ± 0.27	0.317 ± 0.24	0.345 ± 0.22	0.349 ± 0.19

Table 6: Retrieval pipeline ablation study.

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

Nuanced Novelty: Breakthrough ideas often challenge conventional evaluation criteria. Our 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 manuscripts and reviews. Thus, we cannot claim that the approach generalizes to submissions written in other languages or rooted in different academic conventions; assessing cross-lingual performance is left for future work.

Human Analysis for Prompt Design: Our approach lacks formal inter-rater reliability measures, but we argue it was appropriate for this initial investigation into an understudied phenomenon, with the patterns' effectiveness validated by our results.

Human Study: Our human evaluation is based on 100 pairwise comparisons with three expert annotators, comparable to related work in peer review analysis ((Yuan and Liu, 2022): 40 papers; (Chamoun et al., 2024) 100 examples; (Dycke et al., 2025): 20+ papers). While a larger sample would provide additional confidence, novelty assessment requires in-depth domain expert annotators with deep familiarity with the relevant literature, making extensive human evaluation resource-prohibitive.

Acknowledgments

This work has been funded by the European Union (ERC, InterText, 101054961). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work has also been co-funded by the LOEWE Distinguished Chair "Ubiquitous Knowledge Processing", LOEWE initiative, Hesse, Germany (Grant Number: LOEWE/4a//519/05/00.002(0002)/81).

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. 2024. Nougat: Neural optical understanding for academic documents. In *Proceedings of the Twelfth International Conference on Learning Representations*, ICLR '24, Vienna, Austria.

Eric Chamoun, Michael Schlichtkrull, and Andreas Vlachos. 2024. Automated focused feedback generation for scientific writing assistance. In *Findings of the Association for Computational Linguistics: ACL 2024*, ACL '24 Findings, pages 9742–9763, Bangkok, Thailand. Association for Computational Linguistics.

Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. 2025. TreeReview: A dynamic tree of questions framework for deep and efficient LLM-based scientific peer review. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, EMNLP '25, pages 15662–15693, Suzhou, China. Association for Computational Linguistics.

Maitreya Prafulla Chitale, Ketaki Mangesh Shetye, Harshit Gupta, Manav Chaudhary, Manish Shrivastava, and Vasudeva Varma. 2025. AutoRev: Multi-modal graph retrieval for automated peer-review generation. *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.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, and 21 others. 2024. LLMs assist NLP researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, EMNLP '24, pages 5081–5099, Miami, Florida, USA. Association for Computational Linguistics.

Nils Dycke, Matej Zečević, Ilia Kuznetsov, Beatrix Suess, Kristian Kersting, and Iryna Gurevych. 2025. STRICTA: Structured reasoning in critical text assessment for peer review and beyond. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '25, pages 22687–22727, Vienna, Austria. Association for Computational Linguistics.

Oscar Díaz, Xabier Garmendia, and Juanan Pereira. 2024. Streamlining the review process: AI-generated annotations in research manuscripts. *Preprint*, arXiv:2412.00281.

Neil A. Ernst, Jeffrey C. Carver, Daniel Mendez, and Marco Torchiano. 2021. Understanding peer review of software engineering papers. *Empirical Software Engineering*, 26(5):103.

Conghui He, Wei Li, Zhenjiang Jin, Chao Xu, Bin Wang, and Dahua Lin. 2024. OpenDataLab: Empowering general artificial intelligence with open datasets. *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 PJM Horbach and Willem Halffman. 2019. The ability of different peer review procedures to flag problematic publications. *Scientometrics*, 118(1):339–373. Epub 2018 Nov 29.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, NAACL '19, pages 2131–2137, Minneapolis, Minnesota. Association for Computational Linguistics.

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)*, NAACL '25, pages 550–562, Albuquerque, New Mexico. Association for Computational Linguistics.

Neha Nayak Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2022. DISAPERE: A dataset for discourse structure in peer review discussions. In *Proceedings of the 2022 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL '22, pages 1234–1249, Seattle, United States. 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 von 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. 2024. LLM inference serving: Survey of recent advances and opportunities. In *Proceedings of the 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, Yu Rong, Deli Zhao, Tian Feng, and Lidong Bing. 2025. Chain of ideas: Revolutionizing research via novel idea development with LLM agents. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, EMNLP '25 Findings, pages 8971–9004, Suzhou, China. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, ACL '04 Workshop, 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. *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. 2024. 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. In *Proceedings of the Fifth Workshop on Scholarly Document Processing*, SDP '25, pages 96–113, Vienna, Austria. Association for Computational Linguistics.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, EMNLP '23, pages 5548–5566, Singapore. Association for Computational Linguistics.

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. *Preprint*, arXiv: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*, EMNLP '23, 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.

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

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. *Preprint*, arXiv:2506.17728.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning*

Representations, ICLR '20, Addis Ababa, Ethiopia. OpenReview.net.

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 *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems, Datasets and Benchmarks Track*, NeurIPS '23, New Orleans, LA, USA.

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)*, ACL '25, pages 29330–29355, Vienna, Austria. Association for Computational Linguistics.

A Data Analysis

A.1 Sampling Methodology

Our sampling is sentiment-agnostic. We sample for the presence of novelty discussions (both positive and negative), rather than targeting novelty issues specifically. The selected keywords (e.g., "novel", "contribution", "prior work") occur in both types of assessments.

Empirical analysis of sentiment distribution We analyzed the sentiment of novelty discussions in our dataset (352 reviews total) and found: 45 Positive, 234 Negative, 73 Mixed. While negative discussions are more prevalent (as expected, since issues receive more attention in reviews), our data includes substantial coverage of positive and mixed novelty assessments, demonstrating that our sampling captures the full spectrum of novelty discussions rather than being biased toward issues only.

B Human Evaluation Protocol: Novelty Assessment Comparison

B.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.

B.2 Evaluation Framework

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.

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

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.

B.3 Evaluation Guidelines

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.

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 quality was comparable.

B.4 Implementation Details

Evaluation Platform The evaluation was conducted through a custom web interface presenting materials in a standardized format (see Figures 4 and 5). Each evaluator received a unique evaluator ID, 50 randomly assigned paper-assessment pairs, and the ability to save progress and flag unclear cases.

Quality Control We calculated inter-evaluator agreement using Cohen's kappa reported in Table 7.

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 7: Inter-rater reliability metrics across categories.

C Output Examples

Output of our pipeline can be seen in Tables 9, 10 and 11. It is quite evident that our system aligns better with the human as compared to the baselines across all four dimensions.

D Understanding Human Alignment Patterns

Pattern Analysis We analyzed 45 papers where multiple human reviewers reached different novelty conclusions. This was determined from the LLM as judge results we received during evaluation. In a human-AI collaborative setup, we first iteratively examined review pairs to identify recurring disagreement patterns, then developed categories along two observable dimensions: (1) focus divergence - what aspects reviewers discussed (methods, applications, results, prior work, etc.), and (2) assessment granularity - their level of analytical detail (high-level vs detailed).

We then used Claude Code to perform side-by-side comparative reading and categorization of all 45 review pairs according to these predefined categories. Our analysis revealed the patterns: 62.2% of cases (28/45) showed granularity differences with one reviewer providing detailed component-level analysis while another gave high-level assessment, and 75.6% of cases (34/45) showed focus differences with reviewers evaluating different aspects of the work.

Misrepresentation of Novelty Additionally, we manually reviewed ten random generated outputs from the novelty-delta-analysis stage in order to interpret where does the misrepresentation of novelty arise from, we analyzed the structure and reasoning patterns. We found that the system's primary mode of analysis involves evaluating how authors characterize their contributions relative to cited works, using the citation contexts from the paper. When relevant uncited work is identified, the system flags it for additional comparison but bases its core novelty assessment on the cited literature where authors' explicit positioning is available. This pattern reveals that most identifiable novelty overstatement arises from inadequate differentiation from already-cited work. Because we have access to authors' own citation contexts, we can directly evaluate whether their novelty claims hold up against how they characterized prior work. In contrast, for uncited works, we can identify potential gaps but lack the authors' explicit framing of the novelty relationship, making these better suited for clarification during rebuttal rather than definitive assessment.

Factuality Analysis LLMs are known to hallucinate references, which is precisely why our pipeline is specifically designed to be grounded through an extensive multi-step retrieval pipeline with multiple reranking stages. Each related paper alongside the submission is also used in our pipeline. To directly address whether our system's citations accurately support the conclusions or occasionally introduce factual inconsistencies, we conducted a systematic manual verification study.

We randomly sampled 50 factual claims from our novelty delta analysis outputs. Each claim was manually verified against the corresponding paper and categorized as ACCURATE, PARTIALLY ACCURATE, INACCURATE, or CANNOT VERIFY. As shown in Table 8, about 96% of the claims were accurate or partially accurate.

The 14 partially accurate claims typically involved minor discrepancies, such as incorrect author attribution (e.g., "Sun et al." instead of "Xu et al.") or year errors (e.g., 2024 instead of 2023), while the core method descriptions remained correct. Only two claims (4%) contained substantive factual errors: one misattributed a paper's domain (text vs. image attacks), and one overstated a method's historical significance.

Verification Status	**Count**	**Percentage**
Accurate	34	68%
Partially Accurate	14	28%
Inaccurate	2	4%

Table 8: Verification status summary.

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 9: 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.

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 10: 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 11: 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.

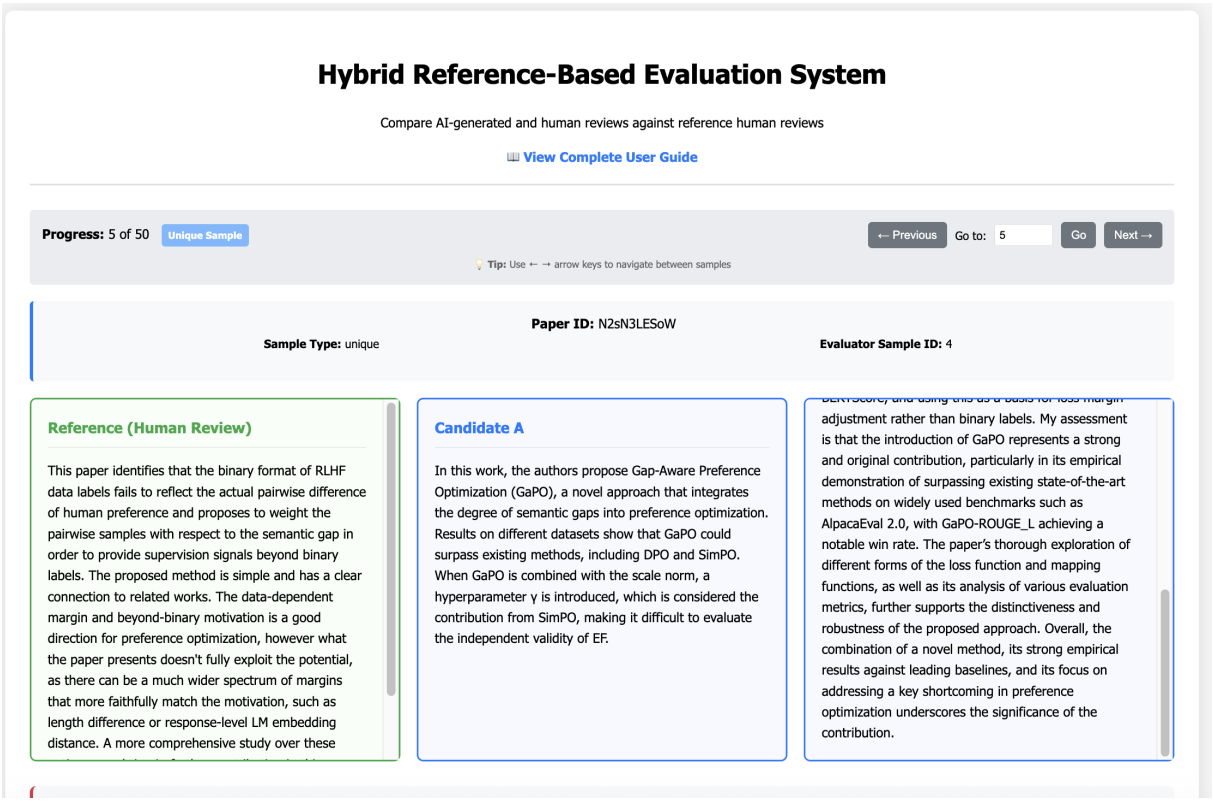

Figure 4: Screenshot of the custom-built interface used for human evaluation. Annotators compared AI-generated and human-written novelty assessments across multiple dimensions, including reasoning depth, prior work engagement, and conclusion alignment.

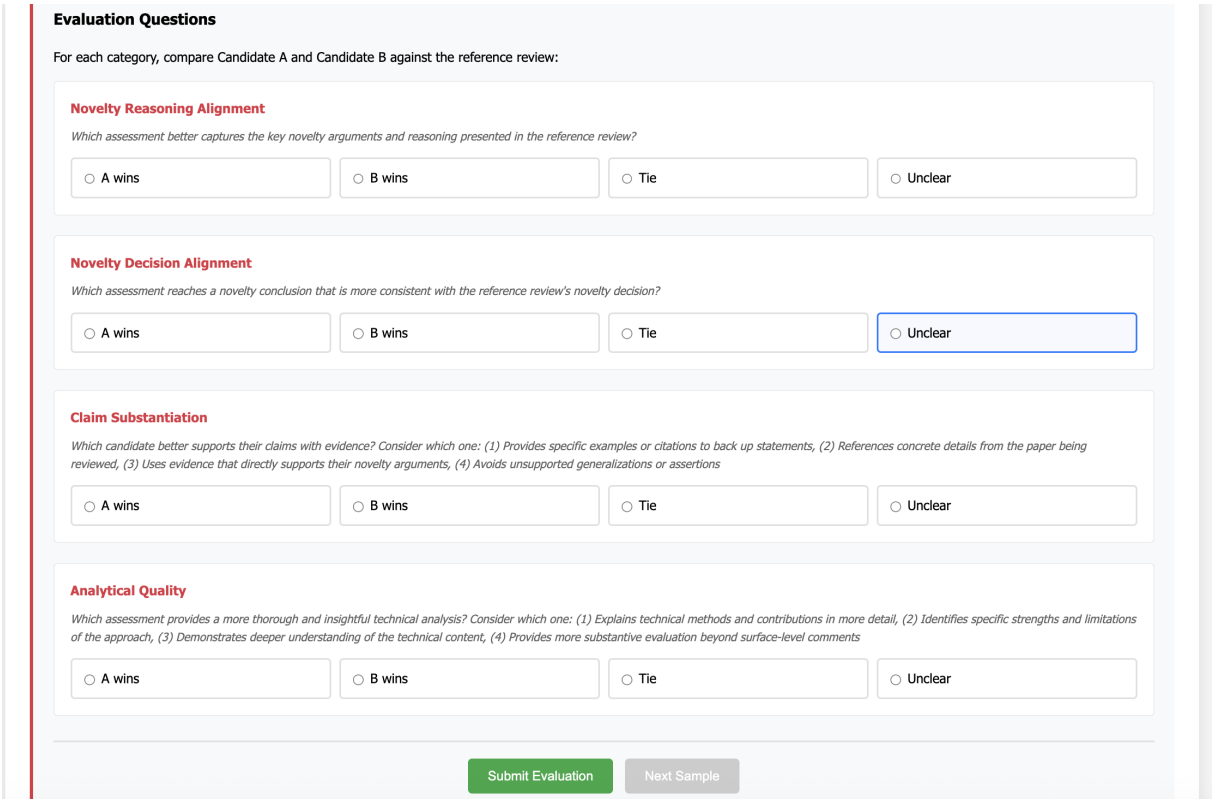

Figure 5: Screenshot (2) of the custom-built interface used for human evaluation. Annotators compared AI-generated and human-written novelty assessments across multiple dimensions, including reasoning depth, prior work engagement, and conclusion alignment.

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 6: Research paper information extraction prompt.

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 7: 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 8: 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 9: 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 10: 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 11: 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 12: 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 13: Reviewer novelty evaluation prompt.