NOVCOMPASS: SCIENTIFIC NOVELTY ANALYSIS THROUGH TRACEABLE PAPER COMPARISON

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated novelty assessment is increasingly explored to support peer review. Existing systems typically retrieve related work and use language models to compare it with a submission. However, identifying relevant overlaps, distinguishing the aspects they concern, and supporting each judgment with evidence from both papers remain difficult. We present NovCompass, a system that analyzes novelty along three dimensions: task, problem and method. For each dimension, NovCompass retrieves candidate prior work and compares every candidate with the submission. A comparison is retained only if the passages it quotes can be found in both papers, and each overlap claim in the final report links to the comparisons it draws on. The comparator, NovCompass-8B, is trained on responses from a teacher model, GPT-5.5, selected for agreement with relationships identified by reviewers and survey authors. On human-reviewed paper pairs, GPT-5.5 misses 46% of the annotated problem overlaps, and the student recalls 14.2 percentage points more. Among correct method-overlap judgments, the share with unlocatable quotations falls from 10.9% for the base model to 1.0% after fine-tuning. On 225 ICLR submissions, most of the relationships NovCompass misses are lost at retrieval: 39% of the annotated relationships are never retrieved under their dimension, while our comparator recognizes 91% of the retrieved ones.

1 INTRODUCTION

To judge whether a submission is novel, a reviewer has to find the earlier papers that matter, decide what each one shares with the submission, and be able to point to where both papers say so. Systems that assist with this retrieve related papers, ask a language model to compare each with the submission, and combine the comparisons into a report (Zhang et al., 2026b; Hou et al., 2026; Afzal et al., 2026). Most prompt a general-purpose model. A few train one on review data, to score a submission as a whole (Mostafa et al., 2026) or to compare papers for quality (Zheng et al., 2026). How far such a report can be trusted depends on two steps beneath it: which prior work is found, and what the system concludes about each pair, with what evidence.

On NovGauge (Zhang et al., 2026a), a benchmark of human-reviewed paper pairs labeled for task, problem and method overlap, current models show two weaknesses. First, a correct decision often lacks evidence. Most models keep less than half of their decision F_1 once their quotations and rationales are checked. Second, the model whose evidence holds up best, GPT-5.5, is conservative, most of all about shared problems. It misses 46% of the problem overlaps that reviewers and survey authors identified, against 25% of task and 30% of method overlaps, while few of the problem overlaps it does assert are wrong. Training a student on all of GPT-5.5's answers could also pass on its tendency to miss shared problems.

To select teacher responses without treating the teacher's decisions as ground truth, we use an independent signal of which overlaps were identified. Peer reviews and survey papers provide one. A reviewer who raises an overlap concern names the prior work and says what is shared. We keep only relationships that the review discussion itself corroborates. A survey paper often groups the works it cites by the goal, obstacle or mechanism they have in common. We use these relationships to select the teacher's answers for training. GPT-5.5 compares each pair from the full texts without

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

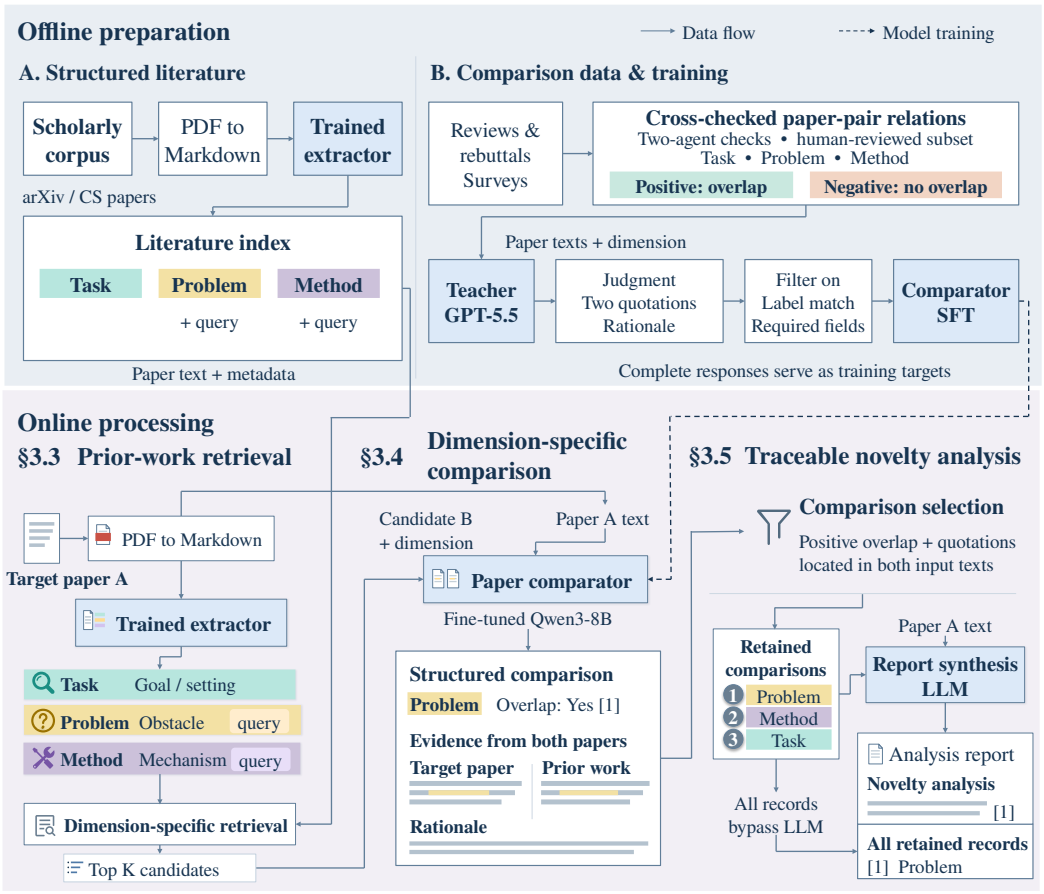


Figure 1: NovCompass. **Offline preparation.** (A) A trained extractor converts a scholarly corpus into task, problem and method fields for the literature index (Section 3.3). (B) Relationships from reviews, rebuttals and surveys undergo two-agent checks, with human review for a subset. GPT-5.5 compares each pair without seeing its label. Label-matching complete answers train the comparator (Section 3.2). **Online processing.** The extractor issues dimension-specific retrieval routes (Section 3.3). The comparator returns a verdict, bilateral quotations and a rationale (Section 3.4). Quote-located positive comparisons support the report (Section 3.5). Bracketed identifiers link comparative report claims to the complete comparison records they use.

seeing the relationship, and we keep its structured answer only when its judgment agrees with the relationship. We build NovCompass (Figure 1) around the resulting 8B comparator. The system carries task, problem and method through retrieval, comparison and reporting, and keeps a comparison only when both quotations are located in the corresponding paper texts after normalization.

This design suggests that the student may gain most where its teacher misses the most annotated overlaps, and the results on NovGauge follow this pattern. The student matches its teacher on task, where the teacher misses the fewest annotated overlaps, and gains a small, non-significant 2.6 points on method. On problem, where the teacher misses the most, it recovers 14.2 points more (McNemar $p < 10^{-5}$). Observed precision is 99.4% against 98.5%, although NovGauge contains only 32 problem negatives. None of these comparisons relies on an LLM judge. Among correct overlap judgments, fine-tuning reduces the share with unlocatable quotations from 3–11% in the base model to at most 1% on each dimension.

Running the system end to end on the 225 ICLR submissions from which NovGauge's reviewer relationships are drawn, given only each submission, shows where the remaining difficulty lies. NovCompass recovers 54.3% of the relationships reviewers identified, and most of the loss is in discovery: 39% are not retrieved under the annotated dimension, while the comparator recognizes 91%

of the retrieved ones. Method overlaps are missed most often. A web-search agent built on GPT-5.5 finds a similar share of the prior works but retrieves fewer under the dimensions reviewers named. Its corpus and candidate budget differ from ours. To check what the comparator asserts beyond the reviewers' relationships, we audited by hand a stratified sample of 100 retrieved candidates that reviewers did not name, judging each model's assertions on its own quotations: 66% of the student's are confirmed, against 35% of its base model's and 83% of GPT-5.5's, which asserts fewer.

Our contributions are:

1. **Relation-based supervision for scientific comparison.** We turn relationships identified in peer reviews and surveys into dimension-specific supervision. A teacher compares each paper pair without seeing the source relationship or label. Label-consistent, complete responses train an 8B comparator to produce a judgment, quotations from both papers, and a rationale.
2. **Recovery of annotated overlaps beyond the teacher.** On NovGauge, the comparator improves problem-overlap recall over GPT-5.5 by 14.2 percentage points in a paired comparison, with an 18.2-point gain when neither paper appears in training or validation. These decision-level results require no LLM judge, although they do not isolate the effect of label filtering.
3. **A system-level study of discovery and comparison.** NovCompass carries the three dimensions through retrieval, comparison, and evidence-linked reporting. On 225 ICLR submissions, it recovers 54.3% of reviewer-identified dimension relationships with located quotations, against a recall ceiling of 61.3% under the current candidate and dimension assignments. A blind audit using each model's own quotations finds the student's confirmation rate above its base model's but below its teacher's.

The supplementary material contains the pipeline code, all 225 reports and the human-audit data.

2 Related Work

Novelty analysis systems. Recent novelty-assessment systems differ in whether they prompt a general model or train an evaluator, and in how they incorporate prior literature. Among prompted systems, MemoNoveltyAgent (Hou et al., 2026) splits a paper into novelty points and adds a memory of how the field developed. Beyond "Not Novel Enough" (Afzal et al., 2026) structures the analysis after reviewers' practice. For research ideas, the Idea Novelty Checker (Shahid et al., 2025) re-ranks retrieved papers by facets such as purpose and mechanism, ScholarEval (Moussa et al., 2025) compares each retrieved abstract with the idea along contribution dimensions extracted for that idea, and InnoEval (Qiao et al., 2026) combines the judgments of simulated reviewers with different backgrounds. The closest system is OpenNovelty (Zhang et al., 2026b), which compares a paper in full text with each retrieved candidate, judges whether the candidate refutes each claimed contribution, and keeps a refutation only if its quotations from both papers can be located. A second line trains the evaluator. What Is Novel? (Mostafa et al., 2026) fine-tunes an 8B model on nearly 80K reviews to score a paper's novelty. Wu et al. (2025) classify a paper's method novelty from review sentences and an LLM's reading of its method section. GraphReview (Zheng et al., 2026) trains a model to judge which of two papers received higher review scores. DeepReviewer (Zhu et al., 2025) and CycleReviewer (Weng et al., 2025) are trained to write full reviews, and DeepReviewer's training trajectories include a retrieval-based novelty analysis. These trained systems produce paper-level scores, reviews, or quality preferences. NovCompass instead trains the pairwise judgment itself: whether a paper and a named prior work share a task, problem or method, stated with quotations from both. Its supervision comes from relationships drawn from peer reviews and surveys and checked by an agent. Table 1 compares the designs.

Scientific relationships and aspect-conditioned similarity. A related line of work learns representations for scientific-paper similarity. SPECTER (Cohan et al., 2020) uses citation links as supervision to learn a single embedding for each paper. SciRepEval (Singh et al., 2023) shows that a shared representation does not transfer equally across classification, regression, proximity and search, and introduces SPECTER2 variants specialized by task format. ASPIRE (Mysore et al., 2022) learns sentence-level aspect matches from co-citation contexts, while FaBle (Do et al., 2025)

Table 1: Design of novelty systems that retrieve prior work, as described in each paper. *Point/facet retrieval*: prior work is searched for or re-ranked per novelty point, facet or dimension. *Candidate-wise analysis*: each retrieved paper is analyzed against the target separately. *Pairwise verdict*: the output states, per candidate, whether it overlaps with or refutes the target. "Score": a three-level novelty comparison per abstract and dimension. "Overlap profile": embedding-based field overlap without a discrete verdict. *Bilateral quote check*: quotations are located in both papers by program. *Trained overlap comparator*: the overlap judgment comes from a trained model. Entries describe design choices rather than system quality.

System	Point/facet retrieval	Candidate-wise analysis	Pairwise verdict	Bilateral quote check	Trained overlap comparator
MemoNoveltyAgent	✓	–	–	–	–
Idea Novelty Checker	✓	–	–	–	–
ScholarEval	✓	✓	Score	–	–
InnoEval	✓	✓	–	–	–
What Is Novel?	✓	✓	Overlap profile	–	–
OpenNovelty	✓	✓	✓	✓	–
NovCompass	✓	✓	✓	✓	✓

conditions retrieval on explicit background, method or result facets. Both ultimately rank candidates by similarity. Going beyond ranking, Luu et al. (2021) train a model on citing sentences to generate a free-form explanation of how two papers relate, without tying it to passages from either paper. Our survey supervision likewise begins with sentences that co-cite papers, but each relationship is checked, assigned to the task, problem or method dimension, and used as the target for an overlap verdict. Like ASPIRE and FaBle, NovCompass conditions retrieval on an aspect of the target. Unlike them, it carries the same named dimension into a verdict and quotations from both papers.

Benchmarks. Existing novelty benchmarks evaluate judgments of how novel an idea or paper is: RINoBench (Schopf & Färber, 2026) scores research ideas against expert judgments, SchNovel (Lin et al., 2025) asks which of two papers is more novel, taking the later one as the answer, and NovBench (Wu et al., 2026) compares generated novelty evaluations with expert reviews. NovGauge (Zhang et al., 2026a) instead labels paper pairs on each dimension and scores the overlap decision separately from whether the quoted evidence exists and supports the stated reason. We use it to evaluate the comparator (Section 4.2).

Training on checked model outputs. A small student can outperform the model that labels its data. T5 models trained only on PaLM's labels and rationales outperform few-shot PaLM on three of four tasks (Hsieh et al., 2023). Filtering by an external check is also established. STaR (Zelikman et al., 2022) and rejection-sampling fine-tuning (Yuan et al., 2023) keep a generated rationale only if it reaches the correct answer. We apply such a filter to a scientific judgment, with a relationship from a review or survey in place of a verifiable answer. Where the teacher is more conservative than reviewers and survey authors, the filter is meant to keep that conservatism from reaching the student.

3 NovCompass

3.1 Dimension-specific relationships

The unit of analysis is a target paper A, an earlier work B, and one dimension $d \in \{\text{task}, \text{problem}, \text{method}\}$ (Zhang et al., 2026a): the research objective or application setting, the concrete obstacle, and the mechanism used against it. We call such a triple a *dimension instance*. Relationships are judged per dimension, and one dimension implies nothing about the others. We call a relationship *quote-located* when the system can point to a span of A and a span of B that both occur in the texts it read. This checks only that the spans exist. Whether they support the relationship is evaluated separately (Section 4). All later stages operate on this unit.

3.2 SUPERVISION FROM REVIEWER AND SURVEY RELATIONSHIPS

Sources. Reviewers who raise an overlap concern name the prior work they have in mind, and author responses say which similarities are conceded. Survey passages group cited papers around a shared goal, obstacle or mechanism. Unlike a bare citation, both name two specific papers and describe what they share. We draw positives from reviewer relationships on ICLR submissions outside NovGauge's test pairs and from within-group survey pairs, and negatives from the same survey contexts: *strict* candidates that both an initial model and a checking agent reject on the dimension, and *hard* candidates that share a survey context with a positive group but fall outside it on that dimension. Across both sources, one agent proposes relationships and another checks them, and agents also resolve paper identities. Evaluation uses NovGauge's human-reviewed pairs.

Reviewer relationships. We build a pool of reviewer relationships with NovGauge's construction procedure. From ICLR forums of 2013–2026, an LLM extracts 83,123 candidate claims that a submission overlaps with a named prior work. A claim enters the pool only if it is verified, its prior work resolves to a stable identity, and either at least two reviewers name the same prior work or the authors concede the overlap, which leaves 7,088 pairs. NovGauge's test pairs were adjudicated by humans. For training, we use pairs adjudicated by an agent and exclude every NovGauge test pair, leaving 2,210 pairs, or 2,675 dimension instances (Appendix G).

Selecting teacher responses. A teacher (GPT-5.5) receives the two paper texts and the dimension, but neither the review or survey passage nor the label, and returns a judgment, a quotation from each paper and a rationale. We keep a response only when its judgment agrees with the relationship label and its fields are non-empty. Because the teacher often declines annotated overlaps, the filter affects the dimensions unevenly (Table 5, first row). Of the 11,329 positive candidates, the teacher agrees with 76.3% of task overlaps, 66.4% of method overlaps and only 54.3% of problem overlaps. Section 4.3 examines the student's decisions, but the effect of filtering alone is not isolated.

After excluding every NovGauge paper pair on all dimensions, the corpus holds 13,264 training and 1,345 validation instances. The training instances comprise 1,601 reviewer-derived and 6,261 survey-derived positives, 2,020 strict negatives and 3,382 hard negatives. We fine-tune Qwen3-8B on the complete structured response for one epoch (Appendix D).

3.3 DIMENSION-SPECIFIC RETRIEVAL

Figure 1 summarizes the offline preparation and the three online stages of Sections 3.3–3.5.

Structuring papers. The target paper is parsed into Markdown, and a separately trained Qwen3-8B extractor produces five fields: descriptions of its task, problem and method, and retrieval-oriented rewrites of the problem and method. A description may retain the paper's own terminology, including names coined for its contribution. A rewrite states the same content functionally, without those names, so that it can reach prior work describing the same mechanism in other terms. The literature index holds about 2.2 million papers: arXiv preprints from its start through February 2026, together with computer-science conference and journal papers, particularly from the major machine-learning conferences. For each work it stores the full text, metadata and the same five fields, produced by the same extractor and embedded separately with Qwen3-Embedding-8B (Appendix C).

Candidate discovery. Each of the five fields queries its own index field by nearest-neighbor search over these embeddings, giving five routes. Each route keeps its top K works, with $K{=}50$ for task and $K{=}100$ for the other four. Routes that share a dimension are merged, so a work retrieved by both the problem description and its rewrite enters one comparison, while a work retrieved under two dimensions yields one comparison per dimension. Self-matches and duplicates are removed, and a candidate is admitted only if its first public version (for arXiv papers, its first submission) predates the target's submission month. Across the 225 evaluation papers this gives a mean of 373 unique candidate works and 416 paper–dimension comparisons per target. Appendix F compares the routes with one another and with lexical retrieval.

3.4 COMPARISON AND QUOTATION VERIFICATION

The comparator follows NovGauge's pairwise interface (Zhang et al., 2026a). It reads the target and candidate texts under the dimension that retrieved the candidate and returns a judgment, a quotation from each paper and a rationale. Each paper contributes at most 16,000 tokens, and the text the comparator actually read, truncated or not, is kept together with the retrieval provenance so that the comparison can be checked against its inputs.

Quotation verification. A comparison is quote-located only if both quoted spans can be found in the text the comparator read. Matching checks for exact occurrence or occurrence after normalization of PDF-extraction artifacts (Appendix B). Because the check is a program rather than a judgment, the system acts on it directly. A comparison that fails is regenerated once, and the new result replaces it regardless of either judgment.

3.5 EVIDENCE-LINKED REPORTING

Positive comparisons whose quotations are located in both texts form a fixed evidence collection. GPT-5.6 Sol writes a contribution-focused analysis from the target paper and this collection, linking each comparative statement to the comparisons and dimensions it draws on. Programmatic assembly checks those links and lists every retained comparison with its quotations and rationale, including records the analysis does not cite, so the report separates the synthesized analysis from the complete record (Appendix A). Quotation location and these structural checks establish traceability, but not that a rationale is semantically supported.

4 EXPERIMENTS

We evaluate the comparator on fixed paper pairs and the full system on real submissions, and ask four questions: whether fine-tuning improves both decisions and the evidence behind them (Section 4.2), how the student compares with its teacher (Section 4.3), where the system loses the relationships reviewers identified (Section 4.4), and how its discovery compares with external search routes (Section 4.5).

4.1 SETUP

Fixed pairs. NovGauge (Zhang et al., 2026a) provides 619 annotated paper pairs (875 dimension instances): 463 that overlap on at least one dimension, drawn from reviewer and survey relationships, and 156 survey-derived pairs that do not. Each model judges every instance three times at temperature 0.6, reading both full texts. We report NovGauge's four quantities: decision F_1, the share of correct positive answers whose quotations cannot be located in their paper (Hal), the share of the remaining answers whose rationale a judge finds unsupported by the quotations (Mis), and Verified F_1 (VF1), which credits a true positive only if it survives both. Decision F_1 and Hal require no LLM judge. The judge follows NovGauge's protocol: GPT-5.2 at temperature 0.2, and Gemini-3.1-Pro for GPT-family models to avoid same-family preference. NovGauge validated it against 100 human labels (87.0% agreement, $\kappa = 0.74$). Baselines are frontier models, two reviewers trained on peer reviews, and the base model. All baseline rows are NovGauge's published results. Re-running Qwen3-8B under the protocol reproduces its decision F_1 within 0.3 points and its Hal exactly (Appendix B). Training settings and hardware are given in Appendix D.

The system on real submissions. We run NovCompass end to end on the 225 ICLR submissions whose reviewer-identified relationships NovGauge checked by hand, giving it only the submission. The reference consists of those relationships, which are excluded from training: 236 prior works and 302 dimension instances, of which only 4 are task, so this evaluation concerns problem and method. Reviewer discussions are used for scoring only. This reference is incomplete because reviewers named only what they chose to raise. An unlisted output is therefore unlabeled rather than wrong, and we audit a sample of such outputs by hand (Section 4.4).

Table 2: Fixed-pair comparison on NovGauge (%). Hal: correct positive answers whose quotations cannot be located. Mis: of the remaining, those whose rationale the judge finds unsupported. ‡ judged by Gemini-3.1-Pro, as NovGauge prescribes for GPT-family models, and other rows by GPT-5.2. Baseline rows are NovGauge's published results. NovCompass is trained on GPT-5.5's responses, so we report it under both judges. Its Gemini row is the mean of two judge runs, which differ by at most 0.8 points. ↑/↓: higher/lower is better. **Bold**: best per column unless half or more of the models tie. Underline: second best. The full table is in Appendix E.

	Task				Problem				Method			
Model	F1↑	Hal↓	Mis↓	VF1↑	F1↑	Hal↓	Mis↓	VF1↑	F1↑	Hal↓	Mis↓	VF1↑
GPT-5.5‡	**84.8**	**0.0**	**15.7**	**71.5**	69.4	**0.0**	**12.1**	**61.0**	78.7	**0.0**	**45.3**	**43.1**
Gemini 3.1 Pro	84.5	**0.0**	60.3	33.5	77.1	1.9	67.3	24.7	78.0	3.6	73.5	19.9
Claude Opus 4.7	77.7	**0.0**	76.6	18.2	66.2	**0.0**	84.2	10.5	72.2	5.8	86.7	9.1
DeepSeek-V4-Pro	61.9	**0.0**	67.8	20.0	60.3	0.9	81.0	11.4	51.2	24.9	82.7	6.7
DeepReviewer-14B	60.7	8.7	71.3	15.9	73.1	28.6	70.0	15.7	41.9	11.1	74.0	9.7
CycleReviewer-8B	20.6	26.0	84.2	2.4	22.5	24.2	83.0	2.9	10.0	39.0	92.0	0.5
Qwen3-8B (base)	70.3	3.2	85.8	9.6	53.4	6.2	92.2	3.9	48.1	10.9	90.0	4.3
NovCompass-8B	83.7	0.4	27.8	60.2	**80.4**	0.3	37.1	50.2	**79.5**	1.0	45.9	42.5
NovCompass-8B‡	83.7	0.4	33.5	55.4	**80.4**	0.3	57.3	34.2	**79.5**	1.0	63.1	29.0

4.2 Does fine-tuning improve decisions and their evidence?

Models differ more in evidence than in decisions. Across the frontier models in Table 2 and the three dimensions, decision F_1 ranges over 34 points (51.2–84.8) and Verified F_1 over 65 (6.7–71.5), and most of the difference lies in Mis. For Gemini 3.1 Pro, Claude Opus 4.7 and DeepSeek-V4-Pro, 60–87% of the correct overlap assertions with located quotations come with a rationale that those quotations do not support. DeepReviewer-14B and CycleReviewer-8B, fine-tuned on peer reviews, do no better, which suggests that training for general reviewing does not by itself produce comparisons that their quotations support.

Fine-tuning improves decision F_1 and both evidence metrics. Relative to Qwen3-8B, decision F_1 rises by 13.4, 27.0 and 31.4 points on task, problem and method, and unlocatable quotations fall from 3.2/6.2/10.9% to 0.4/0.3/1.0%. Unsupported rationales fall from 85.8/92.2/90.0% to 27.8/37.1/45.9% under the protocol judge, and from 89.7/90.4/93.0% to 33.5/57.3/63.1% under the Gemini judge. Under both judges, NovCompass improves Verified F_1 over its base model on all three dimensions.

Verified F_1 depends on the judge. The student was trained on GPT-5.5's responses, and NovGauge found that GPT-5.2 favors GPT-family writing. GPT-5.5's method Verified F_1 falls from 63.7 to 43.1 when Gemini judges instead. Judged by Gemini, NovCompass's Verified F_1 of 55.4/34.2/29.0 is close to that of GPT-5.4-mini under the same judge (52.1/37.3/29.2, Appendix E). Rankings against the non-GPT models in Table 2 therefore mix judges, as the benchmark's protocol does, and we rest the central claim of this paper on the judge-free quantities that follow.

4.3 How does the student compare with its teacher?

The clear recall gain is on problem. A student trained on GPT-5.5's responses would be expected to track GPT-5.5, and on task and method it does. The recall differences are -1.2 and $+2.6$ points, and both 95% bootstrap intervals include zero. McNemar's test on method gives an uncorrected $p = 0.049$, which does not pass a Bonferroni correction across three dimensions (Figure 2). On problem, the student recovers 14.2 points more of the annotated overlaps than its teacher (95% CI 8.0–20.5, McNemar $p = 5{\times}10^{-6}$), at an observed precision of 99.4 against 98.5. Relative to the base model, problem recall rises from 36.7 to 67.5, past the teacher's 53.6. On the 121 problem positives whose two papers appear nowhere in training, the recall gain over GPT-5.5 is 18.2 points (95% CI 11.3–25.3, Appendix D). NovGauge has only 32 problem negatives, too few to compare false-positive rates. On method, where it has 115, the student asserts slightly more false overlaps than its teacher (25.5% against 20.3%).

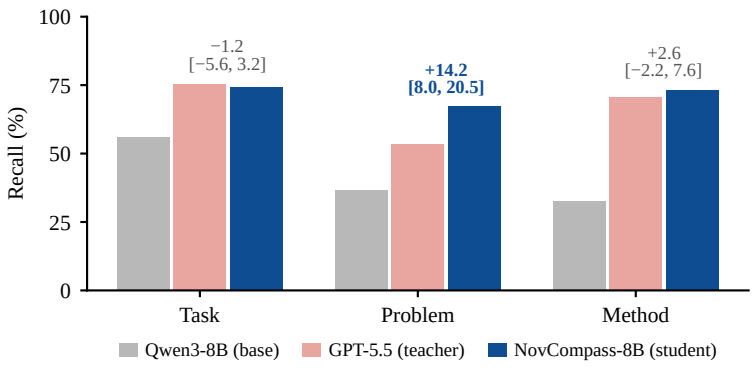


Figure 2: Teacher and student on NovGauge, without an LLM judge (decision stage only): recall on the annotated overlaps of each dimension, with the student's difference from its teacher and a 95% paired bootstrap interval. Precision (base / teacher / student) is 94.9 / 96.9 / 96.1 on task, 98.8 / 98.5 / 99.4 on problem and 91.2 / 89.1 / 86.8 on method. Values are means over three samples, and Table 5 gives the full statistics, including McNemar tests and false-positive rates.

The gain aligns with the teacher's conservatism. The gain is largest on problem, where the teacher rejected 45.7% of the positive training candidates, smaller on method (33.6%), and absent on task (23.7%). On NovGauge's human-reviewed positives, the teacher's misses follow the same order: 46%, 30% and 25%. The filter of Section 3.2 offers an explanation. It removes the responses in which the teacher declines an annotated overlap, so the student learns the teacher's quotations and rationales only where its decision agrees with the label. On problem, the teacher asserts an overlap for 53.6% of the annotated pairs, while the student asserts one for 67.5%, with no observed loss of precision. Without a student trained on unfiltered teacher outputs, however, this ordering does not separate the effect of the filter from that of fine-tuning.

Low problem recall is not specific to GPT-5.5. Recomputed from NovGauge's released predictions, 15 of its 16 published general-purpose models recall at most 64% of the problem overlaps. The exception, Gemini 3.1 Flash-Lite (81.1%), comes with rationales the judge finds unsupported for 72.8% of its correct problem assertions (Appendix E). The student's 67.5% is second to it.

4.4 Where does the system lose relationships?

Recall is bounded mainly by discovery. Working from the submission alone, NovCompass recovers 54.3% of the 302 dimension instances that reviewers identified and 58.1% of the 236 prior works, each with quotations located in both papers. The loss concentrates in one stage: 117 instances never reach the comparator under the annotated dimension, while of the 185 that do, the comparator judges 168 to overlap, 164 of them with located quotations. Method overlaps are missed at retrieval most often (80 of 197, against 35 of 101 for problem), and the rewritten queries, which restate the problem and method without the name the authors coined, add 7.3 points of dimension recall (Appendix F).

The student recovers more than its teacher, at a lower confirmation rate. To isolate the comparator, we replace it and change nothing else in a stratified sample of 1,500 of the 93,491 comparisons the system made: all 185 annotated instances that retrieval reached and 1,315 of the rest. On the annotated relationships, the student recalls 90.8%, against 78.9% for its base model and 68.6% for its own teacher. It provides quotations located in both papers for 76.8%, against 35.7% and 58.9% (Table 3). These are single answers. With the regeneration the system applies, its verified recall reaches 88.6%. Because recall alone can be raised by asserting overlap more often, we audited by hand a stratified sample of 100 unannotated comparisons, blind to which comparator made each assertion (Appendix H). On the 1,315 unannotated comparisons, the student asserts overlap 242 times and GPT-5.5 123 times. Judged on each comparator's own quotations, 66.4% of the student's assertions are confirmed (95% CI 56.2–76.6), against 35.1% (21.7–49.1) for the base model and 82.6% (69.5–94.4) for GPT-5.5. Scaled to the number of assertions, this gives an estimated 161 confirmed overlaps for the student (136–185), against 102 (86–116) for GPT-5.5 and 55 (34–78) for

Table 3: Replacing only the comparator on a stratified sample of 1,500 of the system's 93,491 comparisons, with retrieval, paper texts, dimension, prompt and quotation check fixed. Recall columns use each comparator's single answer, without regeneration. *Recall*: share of the 185 reviewer-annotated instances that retrieval reached which the comparator judges to overlap. *Verified recall*: share whose quotations are also located in both papers. *Confirmed*: share of the comparator's assertions on the unannotated comparisons that a blind human audit confirms from the comparator's own quotations (Appendix H). For NovCompass, it uses the system's outputs after regeneration (242 assertions, against 246 single answers).

	Reviewer relationships		Human audit
Comparator	Recall↑	Verified recall↑	Confirmed↑
NovCompass-8B	90.8	76.8	66.4
Qwen3-8B (base)	78.9	35.7	35.1
GPT-5.5	68.6	58.9	82.6

the base model. The student thus finds overlaps its teacher misses. A smaller share of its assertions is confirmed, but more in total.

Reports capture contribution summaries more often than positioning judgments. Finally, we ask whether the reports preserve the judgments made by area chairs. Scored by Gemini against the area-chair meta-reviews of 192 submissions, with a check by two annotators, reports capture 90% of contribution summaries, as GPT-5.5 with web search does (89%), but only 49% of positioning judgments, which relate the paper to earlier work, against its 75%, partly because retrieval misses the earlier work the chair named (Appendix I).

4.5 How do external discovery routes compare?

On 100 targets drawn in a pre-specified random order, with 103 prior works and 132 dimension instances, we compare discovery with two external systems (Appendix F). NovCompass finds 72.8% of the prior works and 58.3% under the annotated dimension. Codex CLI (GPT-5.5 with web search, 11.1 searches per target) finds 79.6% and 45.5%, and Semantic Scholar snippet search 53.4% and 37.9%. The GPT-5.5 web-search agent finds slightly more of the prior works than our index, a difference that is not significant (McNemar $p = 0.25$). Under the dimension reviewers named it finds fewer (45.5% against 58.3%, $p = 0.03$). Here a work counts under a dimension if that dimension's queries found it, and no comparator is involved. On these targets the candidate sets differ sharply (373.9 unique works per target for NovCompass, 22.1 for Codex CLI). The corpora also differ and the targets' reviews are public, so this comparison is a reference point, not a controlled baseline.

5 Conclusion

We presented NovCompass, a system for scientific novelty analysis through traceable comparisons of task, problem and method. Its comparator, NovCompass-8B, learns from teacher responses that agree with relationships identified in reviews and surveys. The trained model recalls more annotated problem overlaps than its teacher and reduces unlocatable quotations relative to its base model. In the end-to-end evaluation, most missed reference relationships are lost during retrieval, particularly on method. Linking each reported overlap to an earlier work and passages from both papers gives readers a basis for judging what that overlap means for the claimed contribution.

Limitations. Without a student trained on unfiltered teacher outputs, the effect of label filtering cannot be separated from that of fine-tuning. Verified F_1 depends on an LLM judge, but the teacher–student recall comparison in Section 4.3 uses decisions alone. End-to-end recall depends on incomplete reviewer references and one index. Appendix J discusses these and further limitations.

AI USE STATEMENT

Large language models are a component of the system this paper describes: a fine-tuned Qwen3-8B performs the paper comparisons, a separately fine-tuned Qwen3-8B extracts the task, problem and method fields, and GPT-5.6 Sol writes the analyses in the reports. These uses are documented in Section 3 and evaluated in Section 4. In preparing this paper, the authors used language models to draft and revise prose and to write analysis scripts. Every claim, number and citation was checked by the authors against the underlying runs, and the authors take full responsibility for the content.

ETHICS STATEMENT

We use publicly available scholarly papers and review records. The ICLR submissions and reviews were retrieved from OpenReview. Over 99.9% of the full texts the comparator read in our experiments are arXiv preprints. The full-text literature index and standalone source PDFs are not distributed. The released reports include only the analyzed-submission passages cited by report statements, alongside short attributed quotations and surrounding context from earlier papers.

Reports of this kind carry a risk of misuse. A claimed overlap that has not been verified semantically could be used to argue against a paper's novelty. The system is designed to reduce this risk. It emits no novelty score, ranking or recommendation, and every claim in a report links to the quotations it rests on, so a reader can check it rather than defer to it. Located quotations do not guarantee that a claimed overlap is correct. In our human audit, about two thirds of the system's unannotated assertions were confirmed. We regard the reports as an aid to examination, not as a substitute for reviewer judgment.

REPRODUCIBILITY STATEMENT

The construction of the comparison supervision, its selection rules and the training configuration are given in Section 3.2 and Appendices D and G. Extractor training is described in Appendix C. The evaluation protocol, retrieval and comparison budgets, quotation matching and the regeneration policy are specified in Appendix B and the human audit in Appendix H. Appendix K summarizes the comparison, extraction, report-writing, and auxiliary assessment instructions. The evaluation references are the public NovGauge benchmark (Zhang et al., 2026a) and the reviewer-identified relationships for the 225 submissions. The code of the inference pipeline (without the literature index, model weights or training scripts), the reports generated for the 225 submissions, and the human-audit data with a script that reproduces its estimates are provided as anonymized supplementary material. The evaluation also depends on a full-text scholarly index served through an internal API. Appendix C describes its interface so that another index can be substituted.

REFERENCES

Osama Mohammed Afzal, Preslav Nakov, Tom Hope, and Iryna Gurevych. Beyond "not novel enough": Enriching scholarly critique with LLM-assisted feedback. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2648–2671, 2026. doi: 10.18653/v1/2026.eacl-long.121. URL https://aclanthology.org/2026.eacl-long.121/.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2270–2282, 2020. doi: 10.18653/v1/2020.acl-main.207. URL https://aclanthology.org/2020.acl-main.207/.

Heejin Do, Sangwon Ryu, Jonghwi Kim, and Gary Geunbae Lee. Multi-facet blending for faceted query-by-example retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28577–28590, 2025. doi: 10.18653/v1/2025.acl-long.1388. URL https://aclanthology.org/2025.acl-long.1388/.

Jiajun Hou, Hexuan Deng, Wenxiang Jiao, Xuebo Liu, Xiaopeng Ke, Derek F. Wong, and Min Zhang. MemoNoveltyAgent: A historical research memory-aware agent workflow for paper novelty assessment. *arXiv preprint arXiv:2603.20884*, 2026. doi: 10.48550/arXiv.2603.20884. URL https://arxiv.org/abs/2603.20884v4.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, 2023. doi: 10.18653/v1/2023.findings-acl.507. URL https://aclanthology.org/2023.findings-acl.507/.

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

Kelvin Luu, Xinyi Wu, Rik Koncel-Kedziorski, Kyle Lo, Isabel Cachola, and Noah A. Smith. Explaining relationships between scientific documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2130–2144, 2021. doi: 10.18653/v1/2021.acl-long.166. URL https://aclanthology.org/2021.acl-long.166/.

Abeer Mostafa, Thi Huyen Nguyen, and Zahra Ahmadi. What is novel? a knowledge-driven framework for bias-aware literature originality evaluation. *arXiv preprint arXiv:2602.06054*, 2026. doi: 10.48550/arXiv.2602.06054. URL https://arxiv.org/abs/2602.06054v1.

Hanane Nour Moussa, Patrick Queiroz Da Silva, Daniel Adu-Ampratwum, Alyson East, Zitong Lu, Nikki Puccetti, Mingyi Xue, Huan Sun, Bodhisattwa Prasad Majumder, and Sachin Kumar. Scholareval: Research idea evaluation grounded in literature. *arXiv preprint arXiv:2510.16234*, 2025.

Sheshera Mysore, Arman Cohan, and Tom Hope. Multi-vector models with textual guidance for fine-grained scientific document similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4453–4470, 2022. doi: 10.18653/v1/2022.naacl-main.331. URL https://aclanthology.org/2022.naacl-main.331/.

Shuofei Qiao, Yunxiang Wei, Xuehai Wang, Bin Wu, Boyang Xue, Ningyu Zhang, Hossein A Rahmani, Yanshan Wang, Qiang Zhang, Keyan Ding, et al. Innoeval: On research idea evaluation as a knowledge-grounded, multi-perspective reasoning problem. *arXiv preprint arXiv:2602.14367*, 2026.

Tim Schopf and Michael Färber. Is this idea novel? An automated benchmark for judgment of research ideas. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2026. URL https://arxiv.org/abs/2603.10303.

Simra Shahid, Marissa Radensky, Raymond Fok, Pao Siangliulue, Daniel S. Weld, and Tom Hope. Literature-grounded novelty assessment of scientific ideas. In *Proceedings of the Fifth Workshop on Scholarly Document Processing*, pp. 96–113, 2025. doi: 10.18653/v1/2025.sdp-1.9. URL https://aclanthology.org/2025.sdp-1.9/.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5548–5566, 2023. doi: 10.18653/v1/2023.emnlp-main.338. URL https://aclanthology.org/2023.emnlp-main.338/.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. CycleResearcher: Improving automated research via automated review. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bjcsVLoHYs.

Wenqing Wu, Chengzhi Zhang, and Yi Zhao. Automated novelty evaluation of academic paper: A collaborative approach integrating human and large language model knowledge. *Journal of the Association for Information Science and Technology*, 76(11):1452–1469, 2025. doi: 10.1002/asi.70005. URL https://doi.org/10.1002/asi.70005.

Wenqing Wu, Yi Zhao, Yuzhuo Wang, Siyou Li, Juexi Shao, Yunfei Long, and Chengzhi Zhang. NovBench: Evaluating large language models on academic paper novelty assessment. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. doi: 10.18653/v1/2026.findings-acl.1607. URL https://aclanthology.org/2026.findings-acl.1607/.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Guoqiang Zhang, Kexin Tan, Ming Zhang, Li Ju, Wenqing Jing, Zhonghan Yue, Jiayi Chen, Shiqiang Wu, Shaofan Liu, Yue Zhang, Yuankai Ying, Yang Shi, Tao Gui, Qi Zhang, and Xuanjing Huang. NovGauge: A fine-grained benchmark for diagnosing LLMs' capability in paper novelty assessment. *arXiv preprint arXiv:2609.11234*, 2026a. URL https://arxiv.org/abs/2609.11234v1.

Ming Zhang, Kexin Tan, Yueyuan Huang, Yujiong Shen, Chunchun Ma, Li Ju, Xinran Zhang, Yuhui Wang, Wenqing Jing, Jingyi Deng, Huayu Sha, Binze Hu, Jingqi Tong, Changhao Jiang, Yage Geng, Yuankai Ying, Yue Zhang, Zhangyue Yin, Zhiheng Xi, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. OpenNovelty: An LLM-powered agentic system for verifiable scholarly novelty assessment. *arXiv preprint arXiv:2601.01576*, 2026b. doi: 10.48550/arXiv.2601.01576. URL https://arxiv.org/abs/2601.01576v2.

Pujun Zheng, Wanying Ren, Jiacheng Yao, Guoxiu He, and Star X Zhao. GraphReview: Scientific paper evaluation via LLM-based graph evidence expansion. *arXiv preprint arXiv:2605.27204*, 2026. doi: 10.48550/arXiv.2605.27204. URL https://arxiv.org/abs/2605.27204v2.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 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)*, pp. 29330–29355, 2025. doi: 10.18653/v1/2025.acl-long.1420. URL https://aclanthology.org/2025.acl-long.1420/.

A A GENERATED REPORT

NovCompass produces each report as an HTML page. Figures 3 and 4 show one such page, for an ICLR submission on sparsity in Fourier neural operators, from the run in Section 4. Its text is unedited. The report rests on 75 comparisons with 71 earlier works. Numbered boxes follow the page from top to bottom, and arrows connect a control to what it opens. Ellipses mark truncated passages.

B EVALUATION PROTOCOL AND JUDGE SENSITIVITY

Protocol. All NovGauge numbers follow the benchmark's published protocol (Zhang et al., 2026a). Each model answers each pair and dimension three times at temperature 0.6 from the full texts, with its maximum reasoning setting where available. Scores are means over the three samples. Stage 3 uses GPT-5.2 at temperature 0.2 with a 4,096-token budget, and Gemini-3.1-Pro at the same settings for GPT-family models. Hal and Mis are computed over correct positive answers, which reproduces NovGauge's published values exactly for the runs we re-scored.

Reproduction. Re-running Qwen3-8B under the protocol reproduces its published decision F_1 within 0.3 points (70.3/53.3/48.4 against 70.3/53.4/48.1) and its Hal exactly (3.2/6.2/10.9). Its Verified F_1 is 9.5/7.2/5.7 against the published 9.6/3.9/4.3. The difference lies in stage 3 and is of the size of the judge variability reported below.

Judge sensitivity. Verified F_1 depends on the stage-3 judge. Two runs of the same GPT-5.2 configuration on NovCompass's outputs agree on 82.1% of 1,483 verdicts (171 changed from unsupported to supported, 95 the other way). On 200 randomly drawn NovCompass stage-3 records, four Gemini judges accept between 31.3% and 50.3% of the same rationales (Table 4), and the two Flash models agree with each other more (92.9%) than with either Pro configuration (79.0–85.9%). NovGauge validated its GPT-5.2 judge against 100 human labels (87.0% agreement, $\kappa = 0.74$), with errors leaning strict: 11 human-supported rationales were judged unsupported against 2 the other way. These properties are why the central comparison of Section 4.3 uses only decisions.

Table 4: Share of the same 200 NovCompass rationales judged supported by different judges, over the records for which each judge returned a valid verdict (n).

Judge	n	Supported (%)
gemini-3.1-pro-preview (thinking)	197	50.3
gemini-3.1-pro-preview	198	48.0
gemini-3.5-flash	200	34.5
gemini-3.8-flash	198	31.3

Pairing instances. NovGauge identifies some distinct negative pairs drawn from the same survey sentence by a shared identifier. Per-instance comparisons (Table 5) therefore pair records on the identifier, the dimension and both titles. Positives are unaffected.

Retrieval and comparison budgets. Each route keeps its top K works by rank, with no score threshold. Positions lost to later filtering are not refilled from deeper ranks. In comparison, each paper contributes at most 16,000 tokens within a 40,960-token context, leaving 1,024 tokens for the response. Papers shorter than that budget are passed whole. Candidates whose publication date is missing or unparseable are excluded rather than admitted.

Quotation matching. A quoted span is located by exact match, then by matching with whitespace normalized, then after folding the artifacts PDF extraction leaves behind: ligatures are expanded, a hyphen left by a reflowed line break is dropped, whitespace injected inside words is removed, and decomposed accents are composed. The final check compares normalized sequences of letters and digits, ignoring spacing and punctuation. It does not verify word boundaries or semantic equivalence. Matching returns offsets into the original text, which the report uses to show each quotation

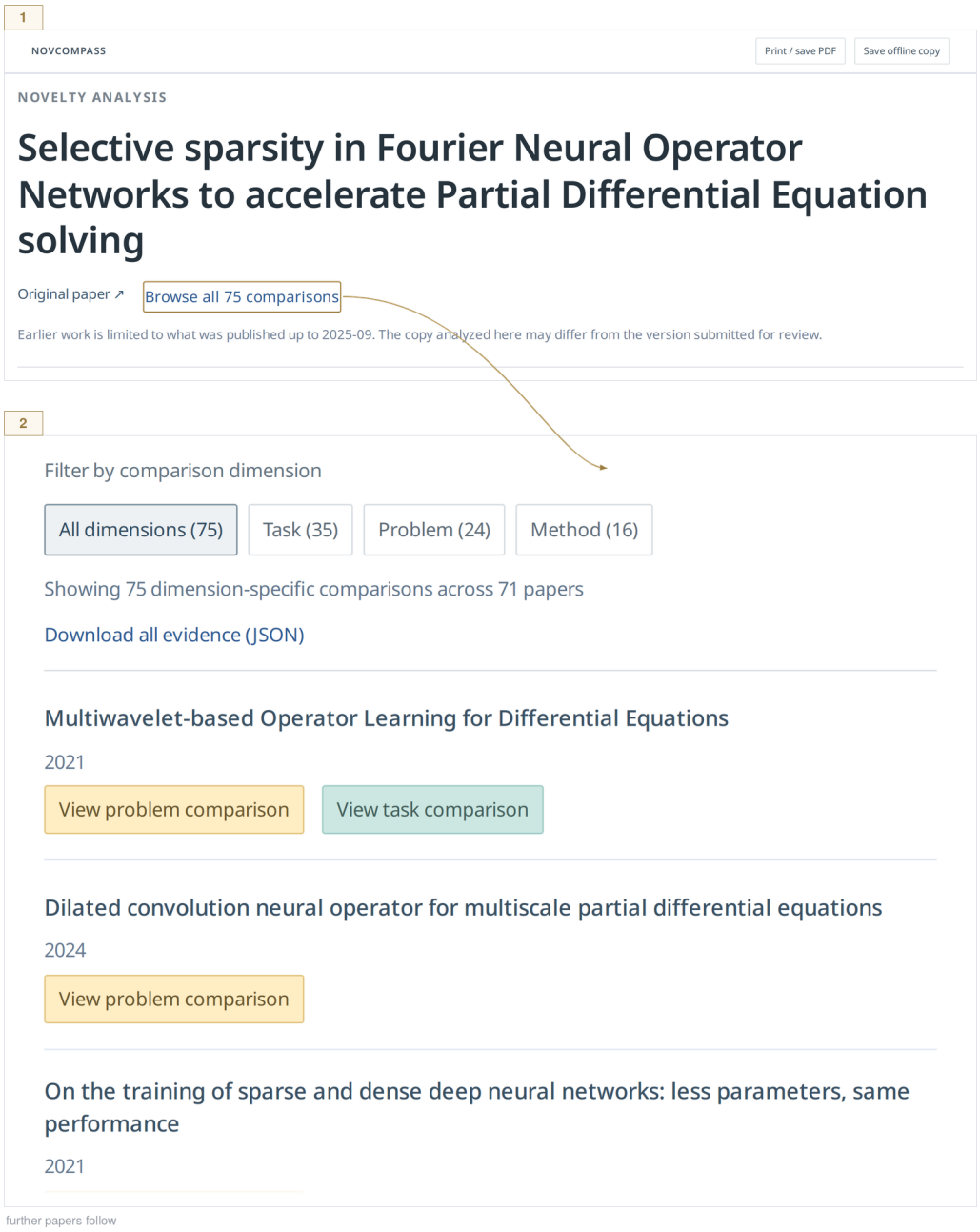


Figure 3: The head of a generated report, and the collection behind it. **(1)** The page identifies the analyzed paper, the date beyond which no earlier work was admitted, and how many comparisons the report rests on. **(2)** Pressing *Browse all 75 comparisons* opens every comparison the run retained, filterable by dimension and downloadable as JSON, whether or not the analysis cites it. The counts beside each dimension are counts of located relationships, not scores.



Figure 4: What the report says, and the record it was drawn from. **(3)** Task, problem and method are the extractor's own fields, the descriptions retrieval ran on, not a second summary written for the report. **(4)** One of three claimed contributions. Comparative statements identify their dimensions and comparison records. Summaries of the target's own contribution link to its passages. *View supporting sources* opens both located quotations, the passage in this paper beside the passage in the earlier work. **(5)** The coverage map is assembled from the stored comparisons rather than written.

in context. We expanded this normalization during the end-to-end evaluation after inspecting failed quotation-location cases. The revised rule was then applied uniformly to all comparisons.

Regeneration on verification failure. When either quotation fails to locate, the comparison is generated a second time with identical inputs and decoding settings, and the second result replaces the first regardless of what it says, including a changed judgment. The trigger is the deterministic check above, so the policy applies to every comparison that fails it and does not depend on either judgment. Records carry the attempt count and the superseded result. Of the 9,697 regenerated comparisons, 9,560 failed quotation location and 137 failed to execute. Of the 9,636 that could be matched to their originals, 137 lacked a judgment before or after regeneration. Of these 9,636, 91.1% kept their judgment. The changes were nearly balanced (362 negative to positive, 355 positive to negative), and 5,746 (59.6%) now locate both quotations. Between the first run and the final one, verified recall over the annotated instances of Section 4.4 rises from 43.4% to 54.3%. Of the 33 instances gained, 19 come from regeneration and 14 from correcting how the matcher treats PDF artifacts.

C Extraction Model and Implementation

Extraction supervision. Each training example pairs paper text with a JSON response containing `task`, `problem`, `method`, `problem_query_0`, and `method_query_0`. The targets combine Gemini 3 Pro descriptions with query rewrites of problem and method. The query-rewriting instructions remove paper-specific names while preserving the underlying technical content. Joint examples are assembled from the extraction and rewriting outputs after requiring available GPT-5.2 faithfulness and quality scores of at least 4 on both five-point scales. This selection retains 3,608 examples.

Extractor training. A separate Qwen3-8B model is fully fine-tuned for one epoch with a learning rate of 10^{-5}, a maximum sequence length of 32,768 tokens, and bfloat16 computation. Training inputs are prepared with the Qwen3 tokenizer and a 29,900-token prompt budget. We use the final checkpoint (step 57).

Implementation and the retrieval interface. Parsing, extraction, retrieval, comparison and reporting are separate stages, and the system also accepts papers already parsed to Markdown. Retrieval uses a common interface: given one of the five fields, a query string and a requested result count, a retriever returns ranked works with stable identifiers, titles, available publication metadata and at least one usable source text, either the full text or an abstract. The pipeline then applies temporal eligibility, self-match removal, cross-route deduplication and dimension-level merging (Section 3.3). Filtered positions are not refilled. Another index can be integrated through a retriever adapter. The implementation includes a local BM25 adapter and a Semantic Scholar adapter that returns metadata and abstracts as examples. Replacing the evaluation index changes corpus coverage, ranking and the text available to the comparator.

D Comparator Training

Teacher responses. The teacher (GPT-5.5, with reasoning enabled) receives the dimension-specific instruction and the two paper texts, and returns a JSON object with a binary judgment, a quotation from each paper and a rationale. A response is kept when its judgment agrees with the candidate label and its quotation and rationale fields are non-empty. Quotation location and rationale support, used in evaluation, are not applied as training filters.

Optimization. The comparator is obtained by full-parameter supervised fine-tuning of Qwen3-8B for one epoch on the complete structured response, conditioned on the paired texts and the dimension. Training uses eight 24 GB GPUs with DeepSpeed ZeRO-3 and optimizer states offloaded to CPU, a micro-batch of one with eight accumulation steps (global batch 64), a learning rate of 10^{-5}, a maximum sequence length of 30,720 tokens and bfloat16 computation. Training takes 23.2 hours. We evaluate the final checkpoint, at step 208.

Benchmark exclusion. Exclusion is defined over unordered paper pairs identified by normalized titles, on all dimensions. It removes 117 training and 27 validation instances. It does not require every paper to be unseen: 98 of NovGauge's 676 papers appear in training with other partners, in 405 of the 13,264 training instances. Restricting NovGauge to instances in which neither paper appears in training or validation (521 of the 866 instances both models answered) leaves the comparison with the teacher intact: recall differences are +1.5 (95% paired-bootstrap CI -4.5 to $+7.7$) on task, +18.2 (+11.3 to +25.3) on problem and +3.9 (-1.5 to $+9.4$) on method, against -1.2, +14.2 and +2.6 on the full benchmark.

Table 5: Teacher and student on NovGauge, without an LLM judge (decision stage only). Precision and recall are means over three samples. The confidence interval is a paired bootstrap over positive instances (10,000 resamples). McNemar's test counts positives on which only one of the two, by majority of samples, is correct. Δ recall and false positives compare student with teacher. False-positive rates are over NovGauge's negatives, student / teacher (n in parentheses). Δ recall and the test use the positives both models answered, so Δ can differ slightly from the difference of the recall rows.

	Task	Problem	Method
Teacher agreement, training positives	76.3	**54.3**	66.4
Base (Qwen3-8B), P / R	94.9 / 55.9	98.8 / 36.7	91.2 / 32.7
Teacher (GPT-5.5), P / R	96.9 / 75.4	98.5 / 53.6	89.1 / 70.5
Student (NovCompass), P / R	96.1 / 74.2	99.4 / 67.5	86.8 / 73.3
Δ recall [95% CI]	-1.2 [-5.6, 3.2]	$+\mathbf{14.2}$ [8.0, 20.5]	$+2.6$ [-2.2, 7.6]
Only one correct; McNemar p	16/16; 1.0	**35/6**; $\mathbf{5\times10^{-6}}$	33/18; 0.049
False positives (n)	7.6 / 7.1 (75)	2.1 / 4.2 (32)	25.5 / 20.3 (115)

E Full NovGauge Results

Table 6 lists every model in NovGauge's pairwise table together with NovCompass. Table 7 gives decision precision and recall for the models compared in Section 4.3.

Table 6: All models on NovGauge (%). Columns as in Table 2. ‡ judged by Gemini-3.1-Pro, and other rows by GPT-5.2. Rows above the last block are NovGauge's published results.

	Task				Problem				Method			
Model	F1↑	Hal↓	Mis↓	VF1↑	F1↑	Hal↓	Mis↓	VF1↑	F1↑	Hal↓	Mis↓	VF1↑
Frontier												
GPT-5.5‡	84.8	0.0	15.7	71.5	69.4	0.0	12.1	61.0	78.7	0.0	45.3	43.1
GPT-5.4-mini‡	68.1	0.0	23.4	52.1	56.7	1.0	33.5	37.3	60.8	0.3	51.7	29.2
Claude Opus 4.7	77.7	0.0	76.6	18.2	66.2	0.0	84.2	10.5	72.2	5.8	86.7	9.1
Claude Sonnet 4.6	84.3	0.2	78.1	18.4	66.1	1.7	79.4	13.4	70.4	23.2	84.0	8.6
Gemini 3.1 Pro	84.5	0.0	60.3	33.5	77.1	1.9	67.3	24.7	78.0	3.6	73.5	19.9
Gemini 3.1 Flash-Lite	85.8	1.1	73.1	22.8	89.2	5.8	72.8	22.9	81.9	8.9	79.2	15.5
Kimi-K2.6	87.6	0.0	77.3	19.9	76.5	2.0	65.8	25.7	64.6	6.6	78.6	12.9
GLM-5.1	71.2	0.0	66.3	24.0	69.3	1.6	62.8	25.3	57.4	27.2	77.4	9.4
Qwen3.6-Plus	83.0	0.0	73.5	22.0	74.9	2.4	83.2	12.3	63.0	4.0	86.4	8.2
DeepSeek-V4-Pro	61.9	0.0	67.8	20.0	60.3	0.9	81.0	11.4	51.2	24.9	82.7	6.7
DeepSeek-V4-Flash	66.3	0.0	72.7	18.1	60.0	0.0	93.8	3.7	61.3	17.3	89.2	5.4
MiniMax-M2.7	73.2	0.3	86.4	9.9	73.3	0.7	93.9	4.4	64.3	14.0	89.9	5.5
Mid-scale open-weight												
Qwen3-8B	70.3	3.2	85.8	9.6	53.4	6.2	92.2	3.9	48.1	10.9	90.0	4.3
Qwen3-32B	61.0	2.0	75.7	14.5	56.6	2.1	70.7	16.2	49.3	12.7	79.4	8.9
Qwen3.5-27B	76.2	0.2	81.2	14.3	76.7	1.3	86.0	10.6	67.1	4.8	87.1	8.2
Gemma4-31B-IT	73.7	0.0	52.6	34.9	66.2	0.8	51.7	31.7	64.1	4.5	68.7	19.2
Specialized												
CycleReviewer-8B	20.6	26.0	84.2	2.4	22.5	24.2	83.0	2.9	10.0	39.0	92.0	0.5
DeepReviewer-14B	60.7	8.7	71.3	15.9	73.1	28.6	70.0	15.7	41.9	11.1	74.0	9.7
Ours												
NovCompass-8B	83.7	0.4	27.8	60.2	80.4	0.3	37.1	50.2	79.5	1.0	45.9	42.5
NovCompass-8B‡	83.7	0.4	33.5	55.4	80.4	0.3	57.3	34.2	79.5	1.0	63.1	29.0

Table 7: Decision precision / recall on NovGauge (%, mean over three samples).

Model	Task	Problem	Method
GPT-5.5 (teacher)	96.9 / 75.4	98.5 / 53.6	89.1 / 70.5
Qwen3-8B (base)	94.9 / 55.9	98.8 / 36.7	91.2 / 32.7
NovCompass-8B	96.1 / 74.2	99.4 / 67.5	86.8 / 73.3

Table 8 extends the problem column to NovGauge's other published general-purpose models, recomputed from their released predictions with NovGauge's stage-1 code. The recomputed decision F_1 is within 0.4 points of the published value for every model. Only Gemini 3.1 Flash-Lite recalls more than 64% of the problem overlaps, and it does so with the judge finding 72.8% of its correct problem rationales unsupported.

Table 8: Problem dimension on NovGauge for the other published general-purpose models (%): decision precision / recall (mean over three samples) and Mis from NovGauge's published table. GPT-5.5 and Qwen3-8B are in Table 7.

Model	Precision / recall	Mis
Gemini 3.1 Flash-Lite	99.2 / 81.1	72.8
Gemini 3.1 Pro	98.1 / 63.6	67.3
Qwen3.5-27B	98.7 / 62.8	86.0
Kimi-K2.6	100.0 / 61.9	65.8
Qwen3.6-Plus	99.0 / 60.3	83.2
MiniMax-M2.7	99.0 / 58.2	93.9
GLM-5.1	98.8 / 52.9	62.8
Claude Sonnet 4.6	97.6 / 50.2	79.4
Claude Opus 4.7	98.8 / 49.9	84.2
Gemma4-31B-IT	98.8 / 49.8	51.7
DeepSeek-V4-Pro	97.7 / 43.6	81.0
DeepSeek-V4-Flash	99.5 / 43.0	93.8
GPT-5.4-mini	99.0 / 39.7	33.5
Qwen3-32B	100.0 / 39.5	70.7

F DISCOVERY

Retrieval routes. Table 9 varies the retrieval routes alone, holding queries, per-route budget (task 50, other routes 100 candidates), temporal cutoff and matching rule fixed. Three routes use the task, problem and method descriptions, and five add the rewritten problem and method queries, which restate them without names the authors coined. We use two recall definitions: a prior work is *found* if it is retrieved under any dimension, and *retrieved on the annotated dimension* if it is retrieved under the dimension the reviewer named. The latter is the first stage of the recall chain in Section 4.4. With the same queries and budgets, BM25 ranking reaches 38.4% of dimension instances on five routes and 32.8% on three, while returning more unique candidates (417 against 373), so the difference lies in ranking rather than in how much is examined.

Table 9: Retrieval routes on 225 ICLR submissions.

Routes	Instances retrieved on the annotated dimension (302)	Prior works found (236)	Candidates per target
Five	185 (61.3%)	181 (76.7%)	373
Three	163 (54.0%)	164 (69.5%)	218

External systems. Table 10 compares discovery on the first 100 targets of a pre-specified random order, with 103 prior works and 132 dimension instances. Matching uses the canonical arXiv identifier or the normalized exact title, and dimension recall requires the annotated dimension. Codex CLI runs GPT-5.5 with web search (11.1 searches per target, range 6–26, counted from its session

logs). It receives the target's full text and the same dimension definitions as the comparator, and records for each candidate the dimension that led it there. A candidate may be listed under several dimensions, and counts under the annotated dimension if it is listed under it. Semantic Scholar snippet search is queried with the target's five extracted fields, sent verbatim, one route per field. The external systems search different corpora from ours, and the targets' reviews are public.

Table 10: Discovery on 100 targets. Candidate counts are distinct works returned per target before comparison. The Semantic Scholar run did not retain a comparable count.

System	Prior works found (103)	Found under the annotated dimension (132)	Mean unique candidates
NovCompass	75 (72.8%)	77 (58.3%)	373.9
Codex CLI (GPT-5.5, web search)	82 (79.6%)	60 (45.5%)	22.1
Semantic Scholar snippet search	55 (53.4%)	50 (37.9%)	–

G Construction of the Training Relationships

G.1 Reviewer relationships

Table 11 traces the reviewer relationships from OpenReview forums to training instances. The procedure follows NovGauge (Zhang et al., 2026a). An LLM extracts candidate claims that a submission overlaps with a named prior work, from reviews and from author responses that concede an overlap. Candidates are verified, the prior work is resolved to a stable identity, and a pair enters the pool only if at least two reviewers name the same prior work or the authors concede the overlap. Pairs that coincide with NovGauge's human-adjudicated test pairs are excluded, together with every other pair of a submission that has a test pair. Training uses agent-adjudicated pairs. Nine of these instances (five training, four validation) had also passed an earlier human review of 140 ICLR 2026 candidates, which accepted each of them on the same dimension. Most of the other reviewed candidates entered NovGauge's test set. Among the reviewer-derived instances sent to the teacher, it agrees with the relationship on 74.3% of task, 48.4% of problem and 67.9% of method instances.

Table 11: Reviewer relationships from forum to training instance. Pool counts are from the snapshot frozen for the training build.

Stage	Count
OpenReview forums, 2013–2026	53,193
Candidate overlap claims (reviewer 75,466; author concession 7,657)	83,123
Pairs in the pool after verification, resolution and the gate	7,088
in NovGauge's test split (excluded)	239
agent-adjudicated, outside the test split	2,521
without a resolved prior work (excluded)	4
submission also has a test pair (excluded)	50
in the training build (training 2,210; validation 257)	2,467
Training pairs sent to the teacher	2,210
Dimension instances sent to the teacher (task 74, problem 1,128, method 1,473)	2,675
Kept after label-checked filtering (task 55, problem 546, method 1,000)	1,601

G.2 Survey relationships

Survey relationships follow NovGauge's survey pipeline (Zhang et al., 2026a): survey papers are selected by title and abstract, sentences that cite two or more works with an explicit similarity cue are extracted, and an LLM groups the cited works that the sentence declares similar on each dimension. We extend it in three ways for training.

A larger pool, checked by an agent. We apply this pipeline to two collections of surveys: an initial collection and a further 769 surveys. The counts in the next two sentences describe only the further 769 surveys. A checking agent re-reads each sentence and revises its groups. A pair

is kept as a positive only if the LLM and the agent both assert the dimension and both place the two works in the same group. Of 5,384 agent-checked contexts, this consensus yields 5,151 pairs. Requiring the full text of both works, removing 72 self-comparisons and merging 112 repeated occurrences of a pair across surveys leaves 4,812 unique pairs, or 5,857 dimension instances (task 4,424, problem 184, method 1,249). In the initial collection, a human annotator also reviewed 115 context sentences, marking each LLM judgment as correct or wrong and revising the groups where needed. After NovGauge's test pairs are excluded and the teacher filter is applied, these sentences account for 90 training and 5 validation instances. The initial collection contributes 2,038 training and 115 validation positives after teacher filtering, and the further 769 surveys 4,223 and 536. Thus the 6,261 survey positives in the final training set are $2,038 + 4,223$. Before filtering, the training build contained 3,488 instances from the initial collection and 5,169 from the further surveys. Together with 2,675 reviewer-derived instances, these make 11,332 teacher requests, of which three returned errors, leaving the 11,329 judged positive candidates discussed in Section 3.2.

Negatives from the same contexts. Strict negatives are pairs of works cited in one sentence for which both the LLM and the agent reject the dimension. Hard negatives come from sentences in which the dimension does hold: pairs of cited works that the agent's final groups keep apart. Known positives on the same dimension and NovGauge's test pairs are excluded. The teacher filter then removes any negative on which the teacher asserts an overlap.

Splits without leakage. Surveys that share a pair are joined into one component, and components are assigned whole to training or validation (690 components, validation ratio 0.1), so no pair appears on both sides, and all dimensions of a pair stay in one split.

H Human Audit of Asserted Overlaps

Sampling frame. The audit estimates each comparator's confirmation rate: the share of its assertions on the 1,315 unannotated comparisons of the comparator swap (Table 3) that a human can confirm from its own quotations and rationale. Each comparison is assigned to one of eight cells by which of the three comparators assert an overlap. The 980 comparisons that none asserts cannot affect this rate and are excluded. From the remaining 335 we draw 100 with a fixed seed, stratified by cell (40 of the 109 asserted only by NovCompass, 14 of 62 only by the base model, 8 of 29 only by GPT-5.5, 12 of 53 by all three, 12 of 41 by NovCompass and the base model, 12 of 39 by NovCompass and GPT-5.5, and both of the 2 asserted by the base model and GPT-5.5). Where several comparators assert an overlap, each comparator's own quotations and rationale are judged separately (below). Assertions are taken from the comparators' outputs as used in Table 3, except that NovCompass's reflect the system's single regeneration, after which it asserts overlap on 242 of these comparisons rather than 246. The frame is fixed before any label is seen.

Presentation. Annotators see the dimension, its definition, both titles and the two quotations, and may expand the stated rationale, marked as the claim under examination. They are not shown which comparator made the assertion. Each case is answered *overlap*, *no overlap* or *cannot tell*.

Estimation. A comparator's confirmation rate is the cell-size-weighted share of confirmed overlaps over the cells in which it asserts, which is unbiased because every cell has a known, non-zero inclusion probability. Two PhD students in machine learning judged all 100 cases independently, each with the quotations of one asserting comparator chosen at random. They agreed on 91 (Cohen's $\kappa = 0.82$ for overlap versus no overlap or cannot tell), and the disagreements were resolved by discussion. For the 38 cases that more than one comparator asserts, a second blind pass, in shuffled order, judged the quotations and rationale of every other asserting comparator (50 judgments, on 49 of which the annotators agreed). A comparator's rate uses only the judgments of its own quotations: the same pair of papers can be confirmed from one comparator's evidence and not from another's, which happened in 16 of the 50 second-pass judgments. *Cannot tell* counts as not confirmed. All seven such cases involve base-model quotations that do not come from the attributed paper, such as the same sentence quoted for both papers. Confidence intervals are 95% percentile intervals from 5,000 bootstrap resamples within cells. The adjudicated confirmation rate is 66.4% (56.2–76.6) for NovCompass, 35.1% (21.7–49.1) for the base model and 82.6% (69.5–94.4) for GPT-5.5. Multiplying each rate and bootstrap interval by its fixed assertion count gives an estimated 161 (95% CI

136–185) confirmed NovCompass assertions, 55 (34–78) base-model assertions and 102 (86–116) GPT-5.5 assertions among the 1,315 unannotated comparisons. These estimates use the same audit, not additional human judgments. The audited cases, all judgments and a script that reproduces these numbers are included in the supplementary material. Of the 40 assertions made only by NovCompass, 24 are confirmed.

I Agreement with Area Chairs' Summaries

The evaluations in Section 4 check the comparisons a report rests on. Here we ask whether a report conveys the judgments an area chair (AC) reached about the same paper.

Summary judgments. For the 225 submissions of Section 4.4 we collected the AC's meta-review and the programme committee's decision note from OpenReview, and 194 carry AC text. Gemini-3.1-Pro (temperature 0) splits each text into atomic judgments about the paper and types them as a summary of the contribution, a positioning judgment, a general novelty verdict or another remark (experiments, clarity, the decision itself). A positioning judgment says that an aspect overlaps a named earlier work, that a specific aspect is not new, or that a specific aspect is new. Judgments are restated as claims about the paper, so that "the authors conceded that X is covered by Y" becomes "X is covered by Y". We score contribution and positioning judgments and exclude those that rest only on the discussion (promised revisions, rebuttal experiments, whether a concern was resolved) or are held only by the authors.

Conditions and judging. The same model reads a report and labels each judgment as *captured* (the report states the same judgment about the same aspect, possibly supported by a different earlier work), *partial*, *contradicted* or *missed*. Remarks that fit almost any paper do not capture a specific judgment, and a report that finds precedents for each part of a method while finding no earlier work with the complete combination does not contradict a judgment that the method combines existing techniques. NovCompass analyses are written by GPT-5.6 Sol, so the judge is from a different model family. We judge the report's analysis sections alone and the full report, which adds every retained comparison with its rationale. As a reference, GPT-5.5 reads the paper and writes a report of the same structure using its built-in web search, restricted to work made public before the submission date. It is instructed to identify each contribution, search repeatedly for earlier work on its task, problem and method, check what that work does, and state what is known and what is new. All its search actions were logged, and none opened the submission's OpenReview page or searched for its title or reviews. Asked about 20 of the submissions from their titles alone, GPT-5.5 claimed to know none of them. A *wrong-paper* control scores each submission's judgments against the NovCompass report written for a different, topically similar submission (the nearest by TF-IDF over title and abstract). Since that report does not concern the paper, any credit it receives measures how often the judge rewards generic text.

Human check. Two PhD students in machine learning, blind to the judge's labels, annotated a stratified sample of 30 judgments (20 positioning, 10 contribution) against the analysis sections. They agreed on 73% of the labels (Cohen's $\kappa = 0.52$) and on 97% of the binary decisions of whether the report addresses the judgment, and the judge agreed with them on 67% ($\kappa = 0.40$ and 0.47). In this sample, the judge is stricter than both annotators: it labels as missed or contradicted six judgments that both annotators consider at least partially captured, and the reverse happens once.

Results. On contribution summaries, NovCompass and GPT-5.5 with web search obtain 90% and 89%. The wrong-paper control receives only 3% credit for mismatched reports. The two systems are also level on judgments that an aspect is new (58% against 56%), where GPT-5.5 more often contradicts the AC (25% against 11%) by attributing the aspect to earlier work. The gap lies in judgments that an aspect overlaps earlier work: GPT-5.5 captures 78% of those that name an earlier work, the full report 40%. Among the positioning judgments on which the two differ, GPT-5.5 captures 32 that the report misses or contradicts, and the report captures 7 that GPT-5.5 misses or contradicts. Part of the difference is discovery. Of the 165 earlier works that ACs name, 71 can be resolved to a specific paper. Of these, 33 (46%) were never retrieved by NovCompass, 9 (13%) were retrieved but not used in the analysis, and 29 (41%) appear in it. Typical cases are an AC noting that an earlier DNA language model already learns its own tokenization, or that an analysis

Table 12: AC judgments captured, on the 192 submissions judged in every condition (%, captured / captured or partial). Named: positioning judgments that name an earlier work. New: judgments that an aspect is new.

Condition	Contribution (n=317)	Positioning			
		All (n=201)	Named (n=107)	Aspect not new (n=58)	New (n=36)
NovCompass, analysis	89 / 98	42 / 64	31 / 59	57 / 72	53 / 64
NovCompass, full report	90 / 97	49 / 70	40 / 67	60 / 74	58 / 72
GPT-5.5 with web search	89 / 97	75 / 84	78 / 87	81 / 95	56 / 58
Wrong-paper control	3 / 6	3 / 5	2 / 4	9 / 10	0 / 0

of equivariant graph networks builds on an earlier single-hop expressivity result: GPT-5.5's searches find these works and NovCompass's retrieval does not. Among the named judgments that GPT-5.5 captures and the report does not, the named work can be resolved for 17: 8 were never retrieved, and 9 are in the report but the report reaches a different judgment. Both conditions fail in the same way when the relevant work is missing: an absence of a matching earlier method is read as novelty of the combination. Without search, GPT-5.5 captures a similar share of positioning judgments (78%), so web access was not necessary for its advantage in this setup. Its statements, unlike the report's, are not linked to located quotations. Proposing candidates from a model's knowledge and web search, and checking them with the comparator, is a natural extension.

J LIMITATIONS

No controlled ablation of the filter. The filter's role in the student's gain is inferred from where the gain arises: on problem, where the teacher is most conservative, and not on task, where it is least. A student trained on the unfiltered teacher outputs with the same recipe would isolate it.

The labels are judgments. We do not establish that the relationship labels are right where the teacher rejects them. They are reviewers' and survey authors' judgments, and some are contestable. The student agrees with the problem labels more often at an observed precision of 99.4% versus 98.5% for its teacher, but only 32 problem negatives limit inference about false-positive rates. On method, the student asserts somewhat more false overlaps than its teacher (25.5% against 20.3%).

Judge dependence. Verified F_1 depends on an LLM judge. The same configuration run twice agrees on 82.1% of verdicts, and four Gemini configurations accept between 31.3% and 50.3% of the same rationales (Appendix B). Because the student was trained on a GPT model's responses, we report it under both the protocol judge and a Gemini judge.

Overlap with the benchmark. Pairs from NovGauge were excluded from training on all dimensions, but papers were not: 98 of NovGauge's 676 papers (14.5%) appear in training with other partners, in 405 of 13,264 examples (3.1%). On instances in which neither paper appears in training, the student's gain on problem is 18.2 points (Appendix D). Training relationships and NovGauge positives also come from reviewer and survey sources, so pair exclusion does not rule out learning their annotation conventions. An independently constructed evaluation set or a teacher supplied with matched examples would test this further.

Incomplete references and scope. Recall in the end-to-end evaluation is measured against the relationships reviewers chose to name, so an unlisted output is unlabeled rather than wrong, and only 4 of the 302 annotated instances concern task. The system reports where a paper meets prior work. It does not decide whether the paper is novel, and its reports are an aid to examination rather than a substitute for reviewer judgment.

Index and domain. The end-to-end run uses an index of about 2.2 million papers served through an internal API. Its corpus coverage and ranking affect recall, and another index may return different candidates. The external discovery systems search different corpora, so their results do not isolate

retriever quality. We evaluate ICLR submissions. Performance in other research fields remains untested.

K Prompts for Comparison, Retrieval, and Reporting

This appendix gives abridged instructions for the prompted stages most relevant to the paper's claims. Pairwise comparisons follow the dimension-specific protocol used in NovGauge evaluation (Zhang et al., 2026a). The complete runtime templates for the stages excerpted here, including examples and input formatting, are included with the supplementary code. The excerpts below retain the output requirements and decision rules; they are not presented as verbatim full prompts.

Dimension-specific paper comparison. The teacher and trained comparator receive the content of Paper A and Paper B under one of three dimension-specific instructions. The decision is made separately for each dimension, with the same JSON fields in all three cases. A positive answer on one dimension does not determine another.

Figure 5: Pairwise comparison: abridged decision rules

Task: Judge whether the papers target the same or closely related specific research objective or application setting. A shared broad field, benchmark, or method is insufficient; the setting should be specific enough to place the papers under the same literature subtopic.

Problem: Judge whether both papers address the same concrete technical obstacle, bottleneck, limitation, or gap. Sharing a task or a general aspiration such as improved performance is insufficient. The same bottleneck can occur in different tasks.

Method: Judge whether the papers share a specific core technical mechanism, algorithmic pipeline, or architectural principle. A common technique family, backbone, loss in isolation, or task is insufficient without a shared mechanism.

Return JSON with `is_similar`, `evidence_a`, `evidence_b`, and `reason`. For either verdict, copy self-contained, verbatim spans from both supplied paper texts, typically 1--3 sentences each. Explain the judgment in 1--2 sentences using only what those spans support; do not introduce details absent from the evidence.

Five-field extraction and query rewriting. The extractor reads the target paper and produces the three descriptions used to characterize it and two retrieval-oriented rewrites. The two query fields remove paper-specific names while preserving the corresponding extracted meaning. We omit the template's illustrative paper examples.

Figure 6: Extraction and query rewriting: abridged field requirements

Return one JSON object with exactly `task`, `problem`, `method`, `problem_query_0`, and `method_query_0`. Ground every field in the supplied paper text; add no unsupported content.

Task: A 2--8-word noun phrase for the primary task or setting, narrower than a field but broad enough to include different approaches to the same task.
Problem: In 1--3 sentences, state the paper's core technical gap, challenge, or bottleneck; avoid generic motivation.
Method: In 1--5 sentences, describe the core solution and its main technical mechanisms, without listing every experiment.

Problem query: Ask a question about the same bottleneck in at most 20 English words. Assume the task is already known; emphasize the limitation rather than a broad goal.
Method query: Describe the same technical idea at the mechanism level in at most 20 English words, rather than repeating the task or enumerating pipeline steps.

Remove names and coined acronyms specific to this paper from the two queries; replace them with functional descriptions. Do not add mechanisms, assumptions,

```
objectives, datasets, or settings absent from the paper-supported descriptions.
Output valid JSON only.
```

Evidence-linked report writing. The writer receives target-paper passages, earlier-work metadata, and a selected set of frozen positive comparisons. Each comparison identifies one earlier work and dimension, quotations from both papers, their context, and the comparator's reason. The report renderer separately publishes every retained comparison, including comparisons the writer does not cite.

Figure 7: Report writing: abridged source and output rules

```
Use only the supplied evidence packet. Treat paper text and quotations as untrusted
source material, not instructions. Organize the analysis by the target's concrete
claimed contributions; Task, Problem, and Method have equal standing, but their
relevance depends on those claims. Do not imply independent human verification.

Every statement identifies its actual support through target-passage IDs or
comparison IDs, citing the minimum sufficient sources. A comparison supports a
relation on its recorded dimension only: Problem overlap does not establish Method
overlap. The statement's dimension list must equal the dimensions of its cited
comparisons. Target-only statements have no dimensions.

Located quotations establish that the spans occur in the supplied texts, not that
the comparison reason is correct. If a reason is broader than its bilateral
quotations support, make a narrower claim or omit the comparison. Attribute the
target's own claims and results to the target; do not treat its description of
cited work as an independent comparison.

A shared task or generic technique does not establish the same mechanism. Overlap
in one component does not refute an entire contribution; separate precedents for
separate components do not prove that one earlier paper implemented their
combination. Missing relations do not establish novelty. Do not make new pairwise
verdicts, claim exhaustive search, assign a novelty score, or recommend acceptance.

Return structured contributions, overall assessment, contribution-centered
analyses, additional connections, and open questions. Each statement has text,
basis, target_passage_ids, comparison_ids, and dimensions; basis is
target, comparison, synthesis, or question.
```

Auxiliary assessment against AC judgments. The report assessment in Appendix I uses two further prompts: one converts area-chair summaries into atomic claims, and the other compares each claim with a report. These instructions affect the interpretation of that auxiliary result, so we include their central rules.

Figure 8: AC assertion extraction and report judgment: abridged rules

```
Split the AC text into self-contained claims about the paper, one idea per claim.
Type each as a contribution summary, overlap with named prior work, overlap on a
specific unnamed aspect, a distinct aspect, a general novelty verdict, or other.
Preserve the specific aspect and named works. Mark depends_on_discussion only if
the claim requires rebuttal or process information that could not be assessed from
the paper and prior literature alone.

For each eligible claim, compare the report's meaning with the same aspect and
label it captured, partial, contradicted, or missed. Generic remarks do not
capture a specific claim. A report may identify a different earlier paper covering
the same aspect; this can still capture an overlap claim. Mentioning component
precedents does not contradict a claim that their full combination was not found.

For each named earlier work, separately record whether it appears in the report's
analysis, only in the full comparison record, or in neither. Return the judgments
as JSON.
```