DYNREVIEWER: A DOMAIN-AWARE AI REVIEWER FOR EACH PAPER FROM PAST HUMAN REVIEWS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI reviewers are entering peer review, where papers from different domains go to different experts, who judge them by the standards of their domain. Yet most AI reviewers apply one fixed configuration to every paper. They are also evaluated mainly on ICLR, where a single average hides performance differences across domains. We propose DynReviewer, which reads the paper under review first, then dynamically configures a dedicated AI reviewer for it. DynReviewer rewrites general checks distilled from past human reviews, along with reviews of related papers, into a paper-specific rubric, then searches the paper for evidence on each item. The whole process requires no model training and no human-written rubrics. We also build ReviewAtlas-40K+, which collects 45,551 papers and their official reviews from six venues. Reviews up to 2024 serve as historical experience, and papers from 2025 form the benchmark ReviewAtlas-Bench, covering four venues and four ICLR research domains. Across three LLM backbones, DynReviewer outperforms all compared methods in every venue and domain at covering human reviewers' concerns with criticisms grounded in the paper. It leads the strongest same-backbone baseline by 5.6 to 19.9 percentage points. Its reviews also have the highest content quality in most settings, and its ICLR acceptance decisions agree best with actual outcomes. Analyses show that general checks distilled from human reviews clearly outperform LLM-written ones, and that rewriting experience into a paper-specific rubric works much better than using it directly. Past human reviews thus let an AI reviewer adapt to each paper, as human peer review does.

1 INTRODUCTION

Peer review relies on domain experts. Conferences and journals assign reviewers by research domain because the standards for judging a paper differ across domains (Laudel, 2024). Theory papers are checked for assumptions and proofs, empirical papers for baselines and ablations, and systems papers for scale and overhead. Even within one venue, reviewers raise very different concerns about a graph learning paper and an LLM alignment paper. AI reviewers are now entering peer review and have been piloted at real conferences (Biswas et al., 2026). They likewise need to adapt to each paper and judge it by the standards of that paper's domain.

Yet most existing AI reviewers apply one fixed configuration to every paper. A prompt, a review form, or a trained model already fixes what to check and which standards to apply before the AI reviewer reads the paper (Yamada et al., 2025; D'Arcy et al., 2024; Garg et al., 2025; Zhu et al., 2025; Weng et al., 2024; Zeng et al., 2025). Some systems retrieve related literature during review (Li et al., 2026) or use human reviews of related papers as context (Bougie & Watanabe, 2025; Lu et al., 2025; Hong et al., 2026), but what to check is still not tailored to the paper under review. Current evaluations also make it hard to test whether one configuration suits every domain. These AI reviewers are developed and evaluated mainly on review data from one venue, ICLR (Zhu et al., 2025; Weng et al., 2024; Zeng et al., 2025; Lu et al., 2025; Li et al., 2026), and a single venue-level average hides how their performance differs across domains. The differences are sizable. In our preliminary study, the same AI reviewer's recall of the weaknesses found by human reviewers varies by 7 to 12 percentage points across domains within one venue (Section 3).

To review well in every domain, an AI reviewer needs reviewing experience from each domain. Human reviewers' experience is on public record: tens of thousands of past reviews on OpenReview

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.

show what reviewers actually care about across venues and domains. We therefore ask: **Can we use these past reviews to configure a domain-aware AI reviewer for each paper under review?**

Answering this question requires enough past reviews and an evaluation that tests AI reviewers separately by venue and by research domain. To this end, we build ReviewAtlas-40K+, which collects 45,551 papers and their official reviews from six venues in 2022–2025. Reviews from 2024 and earlier serve as historical experience that AI reviewers can draw on. From the 2025 papers, we sample the benchmark ReviewAtlas-Bench: 200 papers each from ICLR, ICML, NeurIPS, and WWW, with the ICLR portion further split into four research domains. The evaluation covers three aspects: coverage of human reviewers' concerns by criticisms grounded in the paper, the quality of the review content, and agreement with the actual ratings and acceptance decisions on ICLR.

Our answer to this question is DynReviewer: it first reads the paper under review, then dynamically configures a dedicated AI reviewer from past reviews. Before reviewing any paper, DynReviewer distills from past reviews a set of general checks that apply across domains, such as whether comparisons are fair and whether experiments support the conclusions. These checks form a general experience library. Reviewing a paper then takes two steps: DynReviewer first formulates a paper-specific rubric, then reviews the paper against it. To formulate the rubric, it draws on two paths of experience. One path uses a search agent to select the applicable checks from the general experience library. The other retrieves the past papers most related to the paper under review; their reviews show what human reviewers in this domain care about. Neither kind of experience fits this paper as is: the checks are general, and the reviews were written for other papers. DynReviewer therefore rewrites both into concrete rubric items for this paper. During review, it searches the paper for evidence on each rubric item, and only criticisms supported by the paper's text enter the review. **The whole process requires no model training and no human-written rubrics: DynReviewer formulates the rubric entirely on its own, from past reviews.**

ReviewAtlas-Bench lets us test whether DynReviewer reviews well in every venue and every domain. Three general-purpose LLMs serve as backbones. On each backbone, we compare DynReviewer with three kinds of methods: zero-shot prompting, which asks the LLM to review directly; few-shot prompting, which adds human reviews of related papers to the context; and existing agentic review frameworks (Yamada et al., 2025; Li et al., 2026). We also compare with AI reviewers trained on review data (Zhu et al., 2025; Weng et al., 2024). In every venue and domain, DynReviewer outperforms all compared methods at covering human reviewers' concerns with criticisms grounded in the paper. On this aspect, it scores 5.6 to 19.9 percentage points above the strongest baseline on the same backbone. Its reviews also have the highest content quality in most settings, and its acceptance decisions on ICLR agree best with the actual outcomes. Component ablations, experience-library analyses, finding-budget controls, and source hold-out experiments further trace these gains to three design choices. General checks distilled from human reviews clearly outperform LLM-written ones. Rewriting experience into a paper-specific rubric before use works much better than giving it to the model directly. The two paths of experience complement each other: DynReviewer still leads when reviews from the same domain are unavailable, and improves further when they are available.

This paper makes three contributions. (1) **ReviewAtlas-40K+** is a corpus of 45,551 papers and their official reviews from six venues; it provides historical experience for AI reviewers and is the source of our benchmark. (2) **ReviewAtlas-Bench** is a benchmark that covers four venues and four ICLR research domains and uses human reviewers' concerns as its reference; it reveals domain differences that single-venue averages hide. (3) **DynReviewer** is a training-free method that uses past human reviews to dynamically configure a dedicated AI reviewer for each paper under review. Our results show that an AI reviewer's rubric need not be fixed before the paper is read: it can be formulated dynamically after the paper under review is read, from past reviews by human peers.

2 Related Work

Automated Scientific Reviewing Automated peer review spans specialized training and tool-augmented agents (Nguyen & Ahmadi, 2026). DeepReview, CycleReviewer, and ReviewRL study tuning and iterative review (Zhu et al., 2025; Weng et al., 2024; Zeng et al., 2025); ReviewAgents and ReviewGrounder use agent roles or guideline-based rubrics (Gao et al., 2025b; Li et al., 2026). AI Scientist-v2 embeds reviewing in automated discovery (Yamada et al., 2025), and an AAAI-26 pilot tests AI reviewing at scale (Biswas et al., 2026). Agent Reviewers, GAR, and Hong et al. add

Table 1: Preliminary semantic weakness recall across four selected research domains per venue (%). Min–max gives the domain range; Δ is its spread in percentage points.

	ICLR		NeurIPS		WWW	
Method	Min–max	Δ	Min–max	Δ	Min–max	Δ
DeepReviewer-14B	47.73–57.57	9.84	43.89–55.50	11.61	42.89–50.09	7.21
CycleReviewer-8B	37.56–45.58	8.02	34.60–44.38	9.79	36.27–44.24	7.97

human reviews of related papers to the prompt (Lu et al., 2025; Bougie & Watanabe, 2025; Hong et al., 2026). Reviewer2 trains a model to generate aspect prompts for each paper (Gao et al., 2024), and DIAGPaper and TreeReview have the LLM propose review dimensions or questions per paper (Zou et al., 2026; Chang et al., 2025). RubricReviewer trains a rubric generator on ICLR data for objective, comprehensive reviews (Guo et al., 2026). DynReviewer instead adapts the reviewer to each paper's domain, formulating its rubric from past human reviews without training.

Peer-Review Datasets and Diagnostic Benchmarks PeerRead, NLPeer, Paper SEA, ReviewsSTD, Re^2, and Paper Copilot provide paper–review resources with different review fields, sources, and interaction histories (Kang et al., 2018; Dycke et al., 2023; Yu et al., 2024; Lu et al., 2025; Zhang et al., 2025a; Yang et al., 2026). ReviewArena measures how closely LLM ratings and review text match human ones across six venues (Samarth P et al., 2026). ReviewEval, PRISM, and SPOT assess review quality or factuality (Garg et al., 2025; Loc et al., 2026; Son et al., 2025). BioReview-Bench studies concern recovery in biomedical review (Kim, 2026). AAAR-1.0 includes a PaperWeakness task (Lou et al., 2025); counterfactual and injected-error benchmarks test detection of known flaws (Dycke & Gurevych, 2026; Nguyen et al., 2026); a process-centric benchmark tests whether intermediate review evidence supports decisions (Yuan et al., 2026). ReviewAtlas-Bench instead matches AI reviews against traceable human concerns, checks whether criticisms are grounded in the paper, and reports results by venue and ICLR domain (Appendix B.1).

Experience Transfer and Dynamic Context Nonparametric memory and inference-time context adaptation incorporate experience without parameter updates (Hu et al., 2025). Dynamic Cheatsheet, Agentic Context Engineering, and ExpeTrans transfer reusable strategies (Suzgun et al., 2025; Zhang et al., 2025b; Gao et al., 2025a); RubricRAG and instance-specific rubric generation adapt evaluation criteria to inputs (Fan et al., 2024; Dhole & Agichtein, 2026). DynReviewer applies this idea to peer review by turning past human reviews into paper-specific rubrics with evidence requirements.

3 REVIEWATLAS

We begin with a development comparison of DeepReviewer-14B and CycleReviewer-8B across ICLR, NeurIPS, and WWW. Table 1 reports the range and within-venue spread of semantic weakness recall over four displayed research domains per venue. A human weakness counts as recovered when some generated weakness reaches an embedding cosine similarity of at least 0.5. Both the ICLR and NeurIPS cohorts balance 100 accepted and 100 rejected papers, while the WWW cohort contains accepted papers only. The research domains and review forms vary by venue, so the table diagnoses variation within these samples rather than calibrated cross-venue quality. Full protocol and subgroup results appear in Appendix B.2.

Within ICLR, the domain means span 9.84 percentage points for DeepReviewer-14B and 8.02 for CycleReviewer-8B. The spans reach 11.61 points on NeurIPS and stay above 7 points on WWW for both reviewers. An aggregate would hide that variation, while semantic overlap with human weaknesses cannot establish whether the criticism is supported by the paper. These observations motivate an evaluation with cross-venue and domain-level reporting, a fixed inventory of human concerns, and a separate check of generated findings against evidence in the paper under review.

ReviewAtlas addresses these requirements through ReviewAtlas-40K+, a corpus of papers paired with official human reviews, and ReviewAtlas-Bench, a frozen set of 2025 papers with source-traceable reference concerns. Figure 1 summarizes their three construction stages. AI reviewers use past human reviews dated no later than 2024; the 2025 papers' reviews and decisions are reserved for evaluation. No reference concern can therefore enter an AI reviewer through its retrieved experience.



Figure 1: ReviewAtlas construction. We freeze 800 papers from four 2025 venues and turn their official review notes into source-traceable reference concerns.

3.1 Dataset Construction

Stage I: Collect and Normalize We collect publicly accessible submission, review, and decision notes from six venues (2022–2025) through the OpenReview API (OpenReview, 2026) and convert the paper PDFs to Markdown with MinerU in VLM mode (Wang et al., 2024). We link papers and reviews by forum ID and retain venue, year, decision fields, review texts, and review-note IDs for source tracing. The corpus contains 45,551 paper–review records. Records through 2024 form the historical pool D for the experience library and related-paper retrieval; reviews and decisions for 2025 papers are reserved for evaluation. Figure 2 shows corpus composition; fields and venue counts appear in Appendix B.

Stage II: Sample Papers for Review ReviewAtlas-Bench freezes 200 papers each from ICLR, ICML, NeurIPS, and WWW (2025). Its ICLR subset has four 50-paper research domains with balanced acceptance decisions. The 800 papers contain 2,209 original weakness fields and yield 7,905 atomic human concerns after extraction and within-paper deduplication; Appendix Table 7 gives venue and ICLR research-domain counts. Evaluation scope and method design are compared in Appendix Tables 9 and 10.

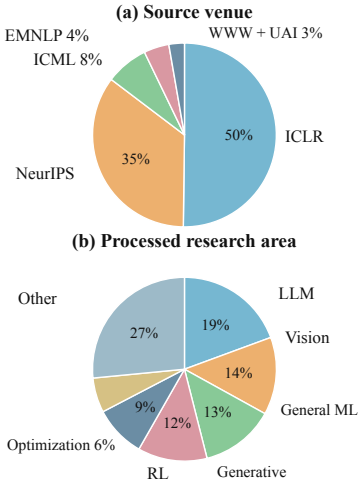


Figure 2: ReviewAtlas-40K+ by source venue and research domain.

Stage III: Extract and Merge Concerns A candidate-independent pass extracts all substantive concerns from official review narratives without a count cap. Each atomic concern records an independently resolvable issue, rationale, minimal resolution, source review-note ID, and exact quotation. We merge equivalent concerns while retaining source links and keep distinct obligations separate. The resulting inventory is frozen for every candidate review. Rules and fields appear in Appendix B; the prompt is in Appendix C.4.

3.2 Review Task and Evaluation

Task Formulation Given the paper under review x, an AI reviewer generates review y without its human reviews or decision, retrieving past human reviews only from D. A *check* names an object and evidence requirement; a *rubric* instantiates it for x; a *finding* is a generated criticism; and a reference *concern* comes from human reviews.

Evaluation We evaluate each review in three ways: (1) Q measures evidence-grounded recovery of human concerns as the harmonic mean of concern coverage (Union_W) and finding support (SFP); (2) Overall scores content quality across evidence correctness, diagnostic depth, actionability, paper specificity, and pitfall severity; and (3) on ICLR, decision accuracy and rating agreement compare native outputs with official outcomes. One temperature-zero `gemini-3.8-flash` call (Google DeepMind, 2026) evaluates coverage, support, and content; ICLR outcome agreement uses official labels. Definitions, prompts, and aggregation appear in Appendices C.3, D.5, and C.4.

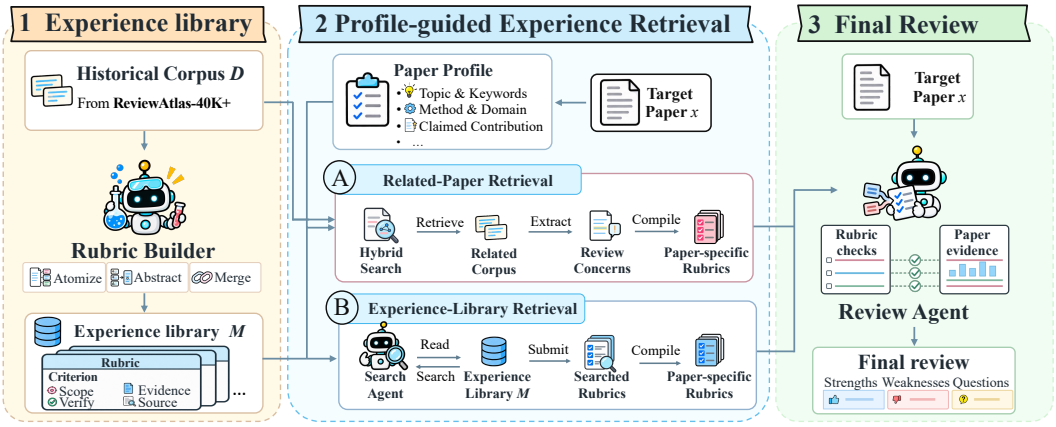


Figure 3: DynReviewer workflow. Past human reviews yield the general experience library M. The paper profile p_x guides two paths: a search agent selects checks from M, while a separate retriever selects related papers and reviews. Each path rewrites its selected experience into a paper-specific rubric. Evidence inspection checks those rubric items against the full paper before findings are assembled into the review.

4 DynReviewer

Method Overview DynReviewer configures a dedicated AI reviewer for the paper under review while keeping the generative model fixed. Offline, it distills general checks from past human reviews into a frozen experience library. At inference, a paper profile guides two paths of experience: a search agent selects general checks, and a dense/lexical retriever selects related papers and reviews. Each path rewrites its selected experience into paper-specific rubric items, inspects the full paper, and supplies evidence-linked findings for the final review. The stage order is fixed; within library search, the agent chooses which check files to inspect and select. Model parameters remain fixed. Figure 3 shows the workflow; Table 9 compares design choices, and Appendix E.2 gives the prompts.

4.1 Experience Library

Offline Experience Distillation From the historical corpus D defined in Section 3, we build a frozen general experience library M through three offline steps: (1) **Weakness atomization** segments human-review criticisms into independently assessable concerns anchored by verbatim spans; (2) **Check abstraction** replaces paper-specific entities with transferable checks and evidence requirements; and (3) **Cross-paper synthesis** clusters recurring checks across independent papers, merging duplicates while keeping source links. Agent-memory work also organizes past experience for later retrieval (Xu et al., 2025; Ouyang et al., 2026); here the stored unit is a general check that can be applied to a new paper.

General Checks Each entry in M specifies an applicable scope, concrete evidence obligations, verification guidelines, and its source reviews. Our primary library comprises 18 general checks distilled from 500 historical papers (analyzed in Section 5.2; full inventory in Appendix E.1, Table 35).

4.2 Profile-Guided Experience Retrieval

The profile guides search over general checks and independent related-paper retrieval; each path then rewrites its selected experience into paper-specific rubric items (Figure 3).

Paper Profile The profile p_x identifies the paper's research question, claims, method, experiments, baselines, ablations, and evidence locations. It is built from the complete paper without its human reviews or decision. Both retrieval paths select with p_x; evidence inspection later returns to the full paper. Fields and prompt appear in Appendix E.2.

Check Search (S_M) Given p_x alone, an agent searches library M through category indexes and atomic check files using `glob` and `grep`. Its terminal `submit_selection` returns exact IDs, paths, and profile-grounded reasons as $S_M \subseteq M$. Exploration is bounded, but the selection count is not fixed. The agent sees no human reviews or decision for the paper under review; Appendix E.2 gives its tools, and Appendix D.4 compares search methods.

Related-Paper Retrieval (S_D) A separate dense/lexical retriever queries historical corpus D with p_x for up to three related papers and human reviews, P_x^{rel}, using BGE-M3 embeddings (Chen et al., 2024). A model extracts methodology and evaluation concerns from each pair and maps them to objects in the paper under review to form S_D. Source-paper criticisms remain candidate checks, not presumed defects of x.

Paper-Specific Rubric Each path $b \in \{M, D\}$ rewrites its candidates S_b using x and p_x into paper-specific rubric items R_x^b. Each records a `criterion`, applicability to the paper under review, an `evidence_requirement`, and `source_refs`. Rewriting makes a general check or related-paper concern testable against this paper (Appendix E.2).

4.3 Evidence-Grounded Review

Evidence Inspection For each rubric item in R_x^b, the AI reviewer checks the full paper under review, including figures and appendices, and records whether the issue remains, is resolved, or is inapplicable. Retained findings F_x^b cite a short quotation and locator and state the unmet evidence requirement. The paper under review supplies factual evidence (Appendix E.2).

Finding Aggregation The final stage groups overlapping findings from F_x^M and F_x^D while retaining independent issues, then writes review y with paper-grounded summary, strengths, weaknesses, and questions. An ICLR-specific suffix requests a rating and Accept/Reject recommendation for native agreement (Appendix E.2).

5 Results

Experimental Settings. We evaluate 200 frozen 2025 papers per venue (ICLR, ICML, NeurIPS, WWW), including four 50-paper ICLR research domains. All systems receive the papers under review without human reviews or decisions. Released DeepReviewer 7B/14B (Zhu et al., 2025) and CycleReviewer 8B/70B (Weng et al., 2024) checkpoints serve as separate references. AI Scientist-v2 (Yamada et al., 2025), ReviewGrounder (Li et al., 2026), and DynReviewer share Qwen 3.7 Plus (Alibaba Cloud, 2026), DeepSeek v4 Flash (DeepSeek AI, 2026), or GLM 5.3 Flash (Z.ai, 2026) backbones where available. Zero-shot prompts the LLM to review directly; few-shot adds up to three retrieved papers and their human reviews as in-context examples to that prompt, without constructing a rubric. DynReviewer uses pre-2025 human reviews, 18 general checks distilled from 500 papers, and up to three related papers. Metrics follow Section 3; ablations share ICLR papers, references, backbone, and judge, while source-shift studies use paired folds. Full protocols appear in Appendix C.

5.1 Overall Comparison

Coverage and Content Quality Within every shared backbone, DynReviewer leads Q across all four venues and ICLR domains (Table 2). The zero-shot and few-shot controls test direct LLM review and the direct use of related human reviews as examples. On ICLR with DeepSeek, their Q scores are 22.90% and 24.07%, versus 42.66% for DynReviewer and 24.58% for ReviewGrounder. The full reviewer exceeds both prompting controls in every reported venue and ICLR domain: simply adding past reviews to the prompt yields less grounded concern coverage than the paper-specific rubric pipeline. The w/o Compile ablation in Table 3 probes the role of rewriting within that pipeline. This within-group pattern also addresses our benchmark motivation: the Q advantage holds across reviewing contexts rather than appearing only in a pooled average. DynReviewer leads Overall in 11 of 12 domain and 8 of 12 venue comparisons (Appendix D.1). ReviewGrounder scores higher on NeurIPS under all backbones, on WWW with DeepSeek, and on Foundation Models & NLP with GLM, so the content-quality advantage is less uniform than the Q advantage.

Table 2: Main results for grounded concern recovery (Q, %) and content quality (Ovr., $[-2, 8]$). FM/GS/GM/TR: foundation models, graphs, general ML, and trustworthy AI. Bold/underline mark best/second distinct values per backbone and column; ties share a mark.

	Cross-Venue Comparison								Cross-Domain Comparison (ICLR)							
	ICLR (200)		ICML (200)		NeurIPS (200)		WWW (200)		FM (50)		GS (50)		GM (50)		TR (50)	
Method	Q	Ovr.	Q	Ovr.	Q	Ovr.	Q	Ovr.	Q	Ovr.	Q	Ovr.	Q	Ovr.	Q	Ovr.
Specialized checkpoints																
DeepReviewer 7B	15.55	1.30	14.33	−0.33	12.54	−0.69	21.78	1.11	13.16	1.57	17.00	1.48	15.33	1.44	16.71	0.71
DeepReviewer 14B	23.33	3.69	22.97	2.39	21.13	2.27	24.39	3.29	20.82	4.10	25.33	2.98	23.47	4.00	23.70	3.66
CycleReviewer 8B	9.33	−0.86	9.63	−1.12	9.47	−1.27	11.36	−1.39	7.08	−1.00	9.13	−1.13	10.21	−0.71	10.98	−0.59
CycleReviewer 70B	12.68	−0.31	13.08	−0.57	12.15	−0.73	15.45	−0.85	9.62	−0.45	12.71	−0.58	13.28	−0.16	14.15	−0.04
Backbone: Qwen 3.7 Plus																
Zero-shot	15.86	5.45	17.93	5.45	17.34	5.70	17.58	5.95	16.21	5.22	15.18	5.38	15.94	5.57	16.11	5.63
Few-shot	17.12	5.75	18.94	5.77	18.57	6.00	18.46	6.28	17.18	5.53	16.42	5.71	17.26	5.87	17.62	5.89
AI Scientist-v2	15.31	4.17	15.12	2.45	15.03	3.37	13.76	3.74	16.31	4.00	14.08	4.34	13.77	4.36	17.11	3.96
ReviewGrounder	17.72	6.12	20.94	6.05	19.98	**7.01**	18.77	6.71	17.39	5.58	16.90	6.28	18.71	6.56	17.88	6.04
DynReviewer (Ours)	**28.83**	**7.20**	**32.70**	**7.10**	**31.83**	6.95	**36.51**	**6.88**	**30.63**	**7.14**	**27.06**	**7.22**	**32.02**	**7.20**	**25.63**	**7.06**
Backbone: DeepSeek v4 Flash																
Zero-shot	22.90	5.75	23.06	4.82	23.43	5.43	23.91	5.35	23.44	5.62	22.71	5.83	23.14	5.86	22.31	5.69
Few-shot	24.07	6.02	24.13	5.02	24.52	5.65	25.04	5.58	24.42	5.93	23.91	6.08	24.18	6.11	23.77	5.96
AI Scientist-v2	21.14	4.37	18.91	2.28	22.72	3.15	24.28	3.35	22.11	4.66	20.29	4.24	19.31	4.10	22.85	4.46
ReviewGrounder	24.58	6.46	26.77	5.18	28.26	**6.66**	27.59	**6.43**	25.09	6.42	24.72	6.42	24.50	6.58	24.03	6.38
DynReviewer	**42.66**	**6.98**	**43.59**	**6.11**	**44.76**	6.49	**45.81**	6.39	**42.51**	**7.06**	**41.93**	**7.04**	**42.29**	**7.02**	**43.90**	**6.78**
Backbone: GLM 5.3 Flash																
Zero-shot	28.13	6.95	28.44	6.72	29.37	6.89	30.42	7.02	30.67	6.87	27.19	7.04	28.24	6.92	26.42	6.97
Few-shot	29.05	7.20	29.96	6.92	30.93	7.08	31.87	7.20	32.38	7.08	28.96	7.27	28.21	7.23	26.65	7.22
AI Scientist-v2	28.98	6.63	31.39	5.68	31.68	5.79	31.63	6.11	27.50	6.60	27.22	6.70	31.49	6.56	29.72	6.66
ReviewGrounder	33.04	7.69	33.00	7.34	32.51	**7.81**	39.75	7.63	40.80	**7.90**	34.49	7.74	28.33	7.52	28.55	7.60
DynReviewer	**44.34**	**7.92**	**46.38**	**7.45**	**48.37**	7.54	**49.69**	**7.78**	**46.38**	7.82	**43.60**	**7.98**	**41.96**	**7.94**	**45.47**	**7.94**

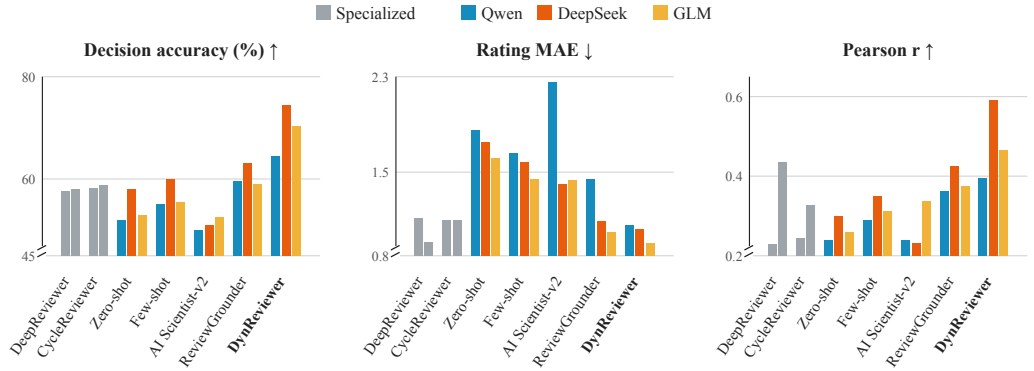


Figure 4: Native decision accuracy and rating agreement on ICLR. Bar colors identify the shared backbones (Qwen, DeepSeek, GLM); gray bars show specialized checkpoints. Broken vertical axes start at 45%, 0.8, and 0.2, respectively; ratings use the original 1–10 scale.

Decision and Rating Agreement Figure 4 asks whether the gains in review content also appear in judgments about the paper as a whole. It compares each reviewer's explicit decision and rating with the official ICLR outcomes; full metrics appear in Appendix Table 23.

Against ReviewGrounder on each shared backbone, DynReviewer improves agreement on the accept/reject decision, increases the correlation between predicted and human ratings, and reduces absolute rating error. With DeepSeek, decision accuracy rises from 63.00% to 74.50% and Pearson r from 0.426 to 0.592; with GLM, MAE falls from 0.994 to 0.905, the lowest error among the compared systems. The prompting controls remain lower: few-shot decision accuracy is 55.00–60.00%, 9.5 to 15.0 points below DynReviewer on the same backbone; specialized checkpoints reach 57.50–59.00%. DynReviewer's ratings also track human scores better than every prompting control. Its Pearson r is 0.394–0.592 versus at most 0.350, and its MAE is 0.905–1.057 versus at least 1.440. Together with Q, these native outcomes show the advantage at two levels of review: supported concerns about the paper and the final paper-level judgment.

Table 3: Component ablation on 200 ICLR papers (DeepSeek v4 Flash). Bold marks the best displayed value in each column, including ties.

	Coverage and support (%)			Judge-rated content						Native agreement		
Variant	Union-W	SFP	Q	EC	DD	Act.	PS	Pit.	Overall	Acc.	MAE	Pearson r
DynReviewer	**29.49**	**88.19**	**42.66**	**1.55**	**1.87**	**1.97**	**1.99**	**-0.40**	**6.98**	**74.50**	**1.024**	**0.592**
w/ Generic Rubrics	25.03	83.71	36.87	0.75	1.62	**1.97**	1.97	-1.52	4.80	55.00	1.433	0.315
w/o Recurring	20.62	84.18	31.44	1.25	1.77	**1.97**	1.98	-0.78	6.18	56.50	1.590	0.294
w/o Related	15.32	77.42	24.12	1.14	1.70	1.93	1.93	-0.78	5.92	63.00	1.575	0.363
w/o Compile	11.02	81.73	17.71	1.21	1.52	1.76	1.82	-0.66	5.64	55.00	1.640	0.281
w/o Experience	6.13	81.46	10.67	1.26	1.54	1.70	1.82	-0.51	5.80	53.50	1.635	0.278

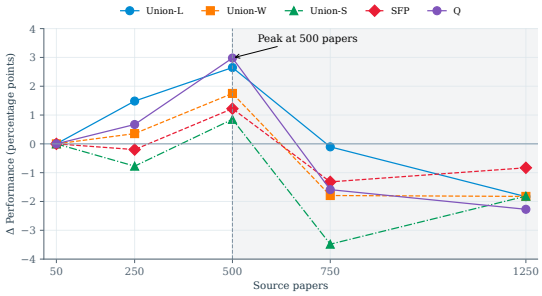


Figure 5: Library scale on 200 development papers: metric changes relative to 50 source papers.

Table 4: Check search on 100 ICLR papers (paired K; %). Δ is agentic minus retrieval; bold marks the better mean.

Metric	Agentic Search	Embedding + Reranking	Δ
Union-W ↑	**26.67**	24.42	+2.25
Q ↑	**39.68**	36.20	+3.48
SFP ↑	**92.18**	89.81	+2.37
Unsupported ↓	**2.76**	4.28	−1.52

5.2 Ablation and Robustness Analysis

Experience and Rubric Ablations Table 3 tests the two design choices in the Introduction: using past human reviews to supply the checks, and rewriting selected experience for the paper under review. "w/o Recurring" removes the review-derived general checks, "w/o Related" removes related-paper reviews, and "w/o Experience" removes both paths. "w/o Compile" passes selected experience to the reviewer without rewriting it into a paper-specific rubric. "w/ Generic Rubrics" instead replaces the 18 review-derived checks with 18 zero-shot LLM-written checks while retaining the related-review path and subsequent steps.

On the same 200 ICLR papers, removing the recurring or related-review path lowers Q from 42.66 to 31.44 or 24.12; removing both yields 10.67. Without related reviews, SFP also falls from 88.19 to 77.42, while removing both paths chiefly reduces coverage (Union-W: 29.49 to 6.13). Skipping this rewriting step lowers Q to 17.71, Overall to 5.64, and decision accuracy to 55.00%: access to past experience alone does not reproduce the full reviewer. Compared with LLM-written generic checks, the review-derived library improves Q (42.66 vs. 36.87), Overall (6.98 vs. 4.80), and ICLR decision accuracy (74.50% vs. 55.00%). These results connect the gains in Table 2 to how DynReviewer constructs its reviewer: both the source of the checks and their adaptation to the current paper matter (Appendix D.2).

Check Selection and Library Scale On 200 development papers, library scale peaks at 500 sources (18 checks; $Q = 45.73\%$), up from 42.76% at 50 sources (three checks), then falls to 40.48% at 1,250 (29 checks; Figure 5). Adding sources does not monotonically improve this library: more historical reviews also change which general checks are distilled. On a separate 100-paper cohort at fixed K, agentic search exceeds embedding retrieval plus reranking on Q (39.68% vs. 36.20%) and SFP (92.18% vs. 89.81%; Table 4). Its lower unsupported-finding rate (2.76% vs. 4.28%) shows that the selected checks help recover concerns without sacrificing grounding. Because K is fixed, the comparison points to which checks are chosen for the paper, not merely how many. Protocols appear in Appendices D.3 and D.4.

Historical-Source Restrictions Venue exclusions lower Q by 0.40 points on ICLR, 0.64 on NeurIPS, and 5.97 on WWW (Table 5). Excluding same-domain sources lowers mean ICLR Q from 42.66% to 40.66%, with drops of 0.69 in general ML, 0.11 in foundation models/NLP, 3.48 in graphs, and 3.70 in trustworthy AI. Historical experience transfers unevenly: WWW, graphs, and trustworthy AI lose more grounded concern coverage. Both restrictions alter the related-paper pool and retained check set (Appendices D.8 and D.9).

Table 5: Experience-source exclusions on paired 50-paper folds (DeepSeek v4 Flash). △ is restricted minus full control: percentage points for Acc. and Q, native units for MAE and Overall. Checks retained out of 18; dashes mark unavailable labels.

(a) Venue Exclusion

Venue	Checks	△Acc.	△MAE	△Overall	△Q
ICLR	10/18	−2.0	0.00	−0.03	−0.40
NeurIPS	9/18	–	–	−0.64	−0.64
WWW	9/18	–	–	−0.73	−5.97

(b) Domain Exclusion

Domain	Checks	△Acc.	△MAE	△Ovr.	△Q
General ML	9/18	−2.0	+0.04	−0.20	−0.69
FM/NLP	11/18	0.0	+0.04	−0.01	−0.11
Graphs	13/18	−4.0	+0.13	−0.64	−3.48
Trustworthy AI	15/18	−6.0	+0.17	−0.21	−3.70

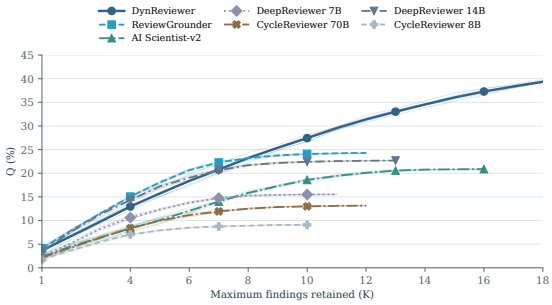


Figure 6: Random finding budget on ICLR papers: mean Q across five shuffles (DeepSeek).

Table 6: Judge-selected top-K findings on 50 ICLR papers (DeepSeek; %). Bold marks the best method for each K and metric.

Method	K	W	SFP	Q
DynReviewer	3	**17.11**	**98.67**	**28.50**
	5	**23.50**	97.00	**36.85**
	8	**27.16**	**96.25**	**41.00**
ReviewGrounder	3	14.16	98.33	23.90
	5	16.75	**98.40**	27.29
	8	17.45	96.23	28.15
AI Scientist-v2	3	11.81	94.67	20.19
	5	14.03	93.40	23.19
	8	14.39	90.38	23.41

5.3 Finding-Level Analysis

Random Finding Budgets Figure 6 asks how grounded concern coverage changes when reviews are capped at K randomly retained findings. On DeepSeek, DynReviewer's Q trails ReviewGrounder's at $K = 5$ (15.71% vs. 18.05%) but overtakes it at $K = 9$ (25.37% vs. 23.76%). ReviewGrounder then levels off near 24.3%, while DynReviewer keeps rising to 39.35% at $K = 18$. With Qwen, DynReviewer overtakes at $K = 8$; with GLM, it leads at every cap (Appendix D.7). An arbitrary short excerpt can therefore miss useful critiques from DynReviewer. Its full-review advantage partly reflects coverage spread across more findings, raising the question of whether a concise subset can preserve that advantage.

Judge-Selected Top Findings Table 6 tests this question on a separate 50-paper cohort: an LLM judge selects the most important K findings from each review before coverage, support, and Q are recomputed. DynReviewer leads ReviewGrounder on Q at $K = 3, 5, 8$; the gap grows from 28.50% vs. 23.90% at $K = 3$ to 41.00% vs. 28.15% at $K = 8$. At $K = 8$, weighted concern coverage is 27.16% vs. 17.45%, while SFP is nearly equal (96.25% vs. 96.23%). The selected DynReviewer findings therefore recover more human concerns at nearly the same support rate. Together, Figure 6 and Table 6 show two complementary strengths: DynReviewer produces a broad pool of useful findings, and even a small selection of its strongest findings covers more human concerns with comparable grounding. Its coverage advantage is thus visible both as more findings are retained and among the top findings. Protocols appear in Appendix D.6. Appendix A discusses limitations.

6 Conclusion

We show that past human reviews can be used to configure a domain-aware AI reviewer for each paper under review. DynReviewer reads the paper first, then rewrites past reviews into a rubric for that paper. Without training, it outperforms all compared methods across four venues and four ICLR research domains at covering human reviewers' concerns with criticisms grounded in the paper. Its ICLR acceptance decisions also agree best with actual outcomes. Evaluations of AI reviewers should likewise report results by venue and domain, because a single venue-level average hides domain differences. Our analyses also suggest that experience from human reviews should be rewritten for each paper before use. Every round of open peer review adds to the domain experience that AI reviewers can draw on.

AI USE STATEMENT

Generative AI tools assisted with manuscript organization and English-language drafting and editing, and provided suggestions for experimental hypotheses, design, and interpretation of development results. Language models also serve as components of the review-generation and evaluation pipelines described in this paper, including experience distillation and reference-concern processing. The authors are responsible for the final manuscript, experimental design, reported results, and conclusions.

ETHICS STATEMENT

DynReviewer is intended to assist scientific assessment, with responsibility retained by human reviewers and authors. Unsupported criticisms can create unnecessary work, misrepresent a contribution, or affect an assessment unfairly; agreement with historical reviews does not eliminate these risks. Historical feedback may reflect uneven expertise and community-specific biases. Any release must respect the permissions and privacy conditions of source papers and reviews. Use on confidential submissions requires authorization and an appropriate data-handling arrangement. No completed human-subject study or ethics approval is claimed in this draft; those details must be reported if an expert study is conducted.

REPRODUCIBILITY STATEMENT

Section 3 defines the data and metrics, Section 4 describes the pipeline, and Appendix G specifies the intended release manifest and controlled comparisons. The current workflow stores stage inputs, outputs, and version signatures to distinguish valid cache reuse from a change in experimental condition. The anonymous artifact will include the frozen paper IDs, reference-concern version, allowed source inventories, prompts, model identifiers, budgets, failure records, and aggregation code.

REFERENCES

Alibaba Cloud. Qwen3.7-Plus model documentation, 2026. URL https://docs.modelstudio.console.alibabacloud.com/en/model-studio/qwen3-7-plus. Accessed September 2026.

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, Jenna Sparks, Jena D. Hwang, Varsha Kishore, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Daniel S. Weld, Doug Downey, Wen-tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. Synthesizing scientific literature with retrieval-augmented language models. *Nature*, 650:857–863, 2026. doi: 10.1038/s41586-025-10072-4. URL https://www.nature.com/articles/s41586-025-10072-4.

Joydeep Biswas, Sheila Schoepp, Gautham Vasan, Anthony Opipari, Arthur Zhang, Zichao Hu, Sebastian Joseph, Matthew Lease, Junyi Jessy Li, Peter Stone, Kiri L. Wagstaff, Matthew E. Taylor, and Odest Chadwicke Jenkins. AI-Assisted Peer Review at Scale: The AAAI-26 AI Review Pilot. *arXiv preprint arXiv:2604.13940*, 2026. URL https://arxiv.org/abs/2604.13940.

Nicolas Bougie and Narimawa Watanabe. Generative Reviewer Agents: Scalable Simulacra of Peer Review. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2025. doi: 10.18653/v1/2025.emnlp-industry.8. URL https://aclanthology.org/2025.emnlp-industry.8/.

Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. 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*, pp. 15651–15682, 2025. doi: 10.18653/v1/2025.emnlp-main.790. URL https://aclanthology.org/2025.emnlp-main.790/.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. MARG: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*, 2024. URL https://arxiv.org/abs/2401.04259.

DeepSeek AI. DeepSeek-V4-Flash-0731 model card, 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash-0731. Accessed September 2026.

Kaustubh D. Dhole and Eugene Agichtein. RubricRAG: Towards Interpretable and Reliable LLM Evaluation via Domain Knowledge Retrieval for Rubric Generation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3681–3687. ACM, 2026. doi: 10.1145/3805712.3809940. URL https://doi.org/10.1145/3805712.3809940.

Nils Dycke and Iryna Gurevych. Automatic Reviewers Fail to Detect Faulty Reasoning in Research Papers: A New Counterfactual Evaluation Framework. *Transactions of the Association for Computational Linguistics*, 14:465–488, 2026. doi: 10.1162/tacl.a.642. URL https://aclanthology.org/2026.tacl-1.22/.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. NLPeer: A unified resource for the computational study of peer review. In *Proceedings of ACL*, 2023. URL https://aclanthology.org/2023.acl-long.277/.

Zhiyuan Fan, Weinong Wang, Xing Wu, and Debing Zhang. SedarEval: Automated Evaluation using Self-Adaptive Rubrics. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16916–16930, 2024. doi: 10.18653/v1/2024.findings-emnlp.984. URL https://aclanthology.org/2024.findings-emnlp.984/.

Jinglong Gao, Xiao Ding, Lingxiao Zou, Bibo Cai, Bing Qin, and Ting Liu. ExpeTrans: LLMs are experiential transfer learners. *arXiv preprint arXiv:2505.23191*, 2025a. URL https://arxiv.org/abs/2505.23191.

Xian Gao, Jiacheng Ruan, Zongyun Zhang, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. ReviewAgents: Bridging the Gap Between Human and AI-Generated Paper Reviews. *arXiv preprint arXiv:2503.08506*, 2025b. URL https://arxiv.org/abs/2503.08506.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. Reviewer2: Optimizing Review Generation Through Prompt Generation, 2024. URL https://arxiv.org/abs/2402.10886.

Madhav Krishan Garg, Tejash Prasad, Tanmay Singhal, Chhavi Kirtani, Murari Mandal, and Dhruv Kumar. ReviewEval: An Evaluation Framework for AI-Generated Reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025. URL https://aclanthology.org/2025.findings-emnlp.1120/.

Google DeepMind. What's new in Gemini 3.8 Flash, 2026. URL https://ai.google.dev/gemini-api/docs/latest-model. Gemini API documentation; accessed September 2026.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. A Survey on LLM-as-a-Judge. *The Innovation*, 7(6):101253, 2026. doi: 10.1016/j.xinn.2025.101253. URL https://www.sciencedirect.com/science/article/pii/S2666675825004564.

Shuyu Guo, Wenxiang Hu, Yuyue Zhao, Yougang Lyu, and Xiaohui Yan. RubricReviewer: From Direct Critique to Objective and Comprehensive Rubric-Driven Peer Review, 2026. URL https://arxiv.org/abs/2608.00005.

Mengze Hong, Di Jiang, Weiwei Zhao, Yawen Li, Yihang Wang, Xinyuan Luo, Yanjie Sun, and Chen Jason Zhang. Multimodal Peer Review Simulation with Actionable To-Do Recommendations for Community-Aware Manuscript Revisions. In *Companion Proceedings of the ACM Web Conference 2026*, pp. 97–100. ACM, 2026. doi: 10.1145/3774905.3793110. URL https://doi.org/10.1145/3774905.3793110.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, Senjie Jin, Jiejun Tan, Yanbin Yin, Jiongnan Liu, Zeyu Zhang, Zhongxiang Sun, Yutao Zhu, Hao Sun, Boci Peng, Zhenrong Cheng, Xuanbo Fan, Jiaxin Guo, Xinlei Yu, Zhenhong Zhou, Zewen Hu, Jiahao Huo, Junhao Wang, Yuwei Niu, Yu Wang, Zhenfei Yin, Xiaobin Hu, Yue Liao, Qiankun Li, Kun Wang, Wangchunshu Zhou, Yixin Liu, Dawei Cheng, Qi Zhang, Tao Gui, Shirui Pan, Yan Zhang, Philip Torr, Zhicheng Dou, Ji-Rong Wen, Xuanjing Huang, Yu-Gang Jiang, and Shuicheng Yan. Memory in the Age of AI Agents, 2025. URL https://arxiv.org/abs/2512.13564.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (PeerRead): Collection, insights and nlp applications. In *Proceedings of NAACL-HLT*, 2018. URL https://aclanthology.org/N18-1149/.

JangKeun Kim. BioReview-Bench: A benchmark for AI-assisted biomedical peer review, 2026. URL https://huggingface.co/datasets/jang1563/bioreview-bench. Version 4.1.3.

Grit Laudel. Where do field-specific notions of research quality come from? *Research Evaluation*, 33:rvae027, 2024. doi: 10.1093/reseval/rvae027. URL https://academic.oup.com/rev/article/doi/10.1093/reseval/rvae027/7735324.

Songze Li, Chuokun Xu, Jiaying Wang, Xueluan Gong, Chen Chen, Jirui Zhang, Jun Wang, Kwok-Yan Lam, and Shouling Ji. LLMs Cannot Reliably Judge (Yet?): A Comprehensive Assessment on the Robustness of LLM-as-a-Judge, 2025. URL https://arxiv.org/abs/2506.09443.

Zhuofeng Li, Yi Lu, Dongfu Jiang, Haoxiang Zhang, Yuyang Bai, Chuan Li, Yu Wang, Shuiwang Ji, Jianwen Xie, and Yu Zhang. Reviewgrounder: Improving review substantiveness with rubric-guided, tool-integrated agents. *arXiv preprint arXiv:2604.14261*, 2026. URL https://arxiv.org/abs/2604.14261.

Tzu-Ling Lin, Wei-Chih Chen, Teng-Fang Hsiao, Hou-I Liu, Ya-Hsin Yeh, Yu-Kai Chan, Wen-Sheng Lien, Po-Yen Kuo, Philip S. Yu, and Hong-Han Shuai. Breaking the Reviewer: Assessing the Vulnerability of Large Language Models in Automated Peer Review Under Textual Adversarial Attacks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025. doi: 10.18653/v1/2025.findings-emnlp.259. URL https://aclanthology.org/2025.findings-emnlp.259/.

Ngoc Phan Phuoc Loc, Toan Huynh La Viet, Thanh Tran Khanh, Duy A Nguyen, Tuan Anh Nguyen Pham, Thanh Nguyen, Nitesh V. Chawla, Wray Buntine, Kok-Seng Wong, Khoa D. Doan, and Binh T. Nguyen. PRISM: A Multi-Dimensional Benchmark for Evaluating LLM Peer Reviewers, 2026. URL https://arxiv.org/abs/2605.26730.

Renze Lou, Hanzi Xu, Sijia Wang, Jiangshu Du, Ryo Kamoi, Xiaoxin Lu, Jian Xie, Yuxuan Sun, Yusen Zhang, Jihyun Janice Ahn, Hongchao Fang, Zhuoyang Zou, Wenchao Ma, Xi Li, Kai Zhang, Congying Xia, Lifu Huang, and Wenpeng Yin. AAAR-1.0: Assessing AI's Potential to Assist Research. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40361–40383, 2025. URL https://proceedings.mlr.press/v267/lou25c.html.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery, 2024. URL https://arxiv.org/abs/2408.06292.

Kai Lu, Shixiong Xu, Jinqiu Li, Kun Ding, and Gaofeng Meng. Agent Reviewers: Domain-specific multimodal agents with shared memory for paper review. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40803–40830. PMLR, 2025. URL https://proceedings.mlr.press/v267/lu25p.html.

Dang Nguyen, Wanqing Hao, Yanai Elazar, and Chenhao Tan. Benchmarking Agentic Review Systems, 2026. URL https://arxiv.org/abs/2606.19749.

Thi Huyen Nguyen and Zahra Ahmadi. LLM-Based Scientific Peer Review: Methods, Benchmarks, and Reliability Challenges, 2026. URL https://arxiv.org/abs/2606.25057.

OpenReview. About OpenReview, 2026. URL https://openreview.net/about. Accessed September 25, 2026.

OpenSciLM. OpenScholar Reranker model card, 2026. URL https://huggingface.co/OpenSciLM/OpenScholar_Reranker. Accessed September 2026.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T. Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. ReasoningBank: Scaling Agent Self-Evolving with Reasoning Memory. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jL7fwchScm.

Samarth P, Sai Sathvik Motamarri, Vyoman Jain, and Shiva Golugula. ReviewArena: A Large-Scale Cross-Conference Dataset and Benchmark for LLM Peer Review. In *ICML 2026 Workshop on AI for Science*, 2026. URL https://openreview.net/forum?id=yugEO52gkR.

Sentence Transformers. all-mpnet-base-v2 model card, 2026. URL https://huggingface.co/sentence-transformers/all-mpnet-base-v2. Accessed September 2026.

Guijin Son, Jiwoo Hong, Honglu Fan, Heejeong Nam, Hyunwoo Ko, Seungwon Lim, Jinyeop Song, Jinha Choi, Gonçalo Paulo, Youngjae Yu, and Stella Biderman. When AI Co-Scientists Fail: SPOT-a Benchmark for Automated Verification of Scientific Research, 2025. URL https://arxiv.org/abs/2505.11855.

Mirac Suzgun, Mert Yuksekgonul, Federico Bianchi, Dan Jurafsky, and James Zou. Dynamic Cheatsheet: Test-Time Learning with Adaptive Memory. *arXiv preprint arXiv:2504.07952*, 2025. URL https://arxiv.org/abs/2504.07952.

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. MinerU: An Open-Source Solution for Precise Document Content Extraction, 2024. URL https://arxiv.org/abs/2409.18839.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. CycleResearcher: Improving automated research via automated review. *arXiv preprint arXiv:2411.00816*, 2024. URL https://arxiv.org/abs/2411.00816.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic Memory for LLM Agents. In *Advances in Neural Information Processing Systems 38*, 2025. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/19909c36f51abc4856b4560aff3d36d6-Abstract-Conference.html.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025. URL https://arxiv.org/abs/2504.08066.

Jing Yang, Qiyao Wei, and Jiaxin Pei. Paper Copilot: Tracking the Evolution of Peer Review in AI Conferences. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6f951a09ec37047d46773ccf81816e29-Abstract-Conference.html.

Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, RenJing Cui, Chengcheng Han, Qiushi Sun, Zhiyong Wu, Yunshi Lan, and Xiang Li. Automated Peer Reviewing in Paper SEA: Standardization, Evaluation, and Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. doi: 10.18653/v1/2024.findings-emnlp.595. URL https://aclanthology.org/2024.findings-emnlp.595/.

Siming Yuan, Xueyi Zhang, Wangze Ni, Tianfang Xiao, Shimin Di, Jia Zhu, Zhuoren Jiang, Rong Tan, Lei Chen, and Kui Ren. Beyond Final Decisions: A Process-Centric Benchmark for Transparent AI-Assisted Peer Review, 2026. URL https://arxiv.org/abs/2609.05947.

Z.ai. GLM-5.3-Flash model card, 2026. URL https://huggingface.co/zai-org/GLM-5.3-Flash. Accessed September 2026.

Sihang Zeng, Kai Tian, Kaiyan Zhang, Yuru Wang, Junqi Gao, Runze Liu, Sa Yang, Jingxuan Li, Xinwei Long, Jiaheng Ma, Biqing Qi, and Bowen Zhou. ReviewRL: Towards Automated Scientific Review with RL. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://aclanthology.org/2025.emnlp-main.857/.

Daoze Zhang, Zhijian Bao, Sihang Du, Zhiyi Zhao, Kuangling Zhang, Dezheng Bao, and Yang Yang. Re^2: A Consistency-ensured Dataset for Full-stage Peer Review and Multi-turn Rebuttal Discussions, 2025a. URL https://arxiv.org/abs/2505.07920.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, and Kunle Olukotun. Agentic Context Engineering: Evolving Contexts for Self-Improving Language Models. *arXiv preprint arXiv:2510.04618*, 2025b. URL https://arxiv.org/abs/2510.04618.

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/.

Zhuoyang Zou, Abolfazl Ansari, Delvin Ce Zhang, Dongwon Lee, and Wenpeng Yin. DIAGPaper: Diagnosing Valid and Specific Weaknesses in Scientific Papers via Multi-Agent Reasoning, 2026. URL https://arxiv.org/abs/2601.07611.

APPENDIX CONTENTS

A Discussion and Limitations **15**

B Benchmark Data and Preliminary Analysis **16**
B.1 Corpus and Benchmark Sampling . . . 16
B.2 Preliminary Analysis . . . 20

C Experimental Setup **22**
C.1 Cohorts, Methods, and Controls . . . 22
C.2 Baseline Methods and Compared Frameworks . . . 22
C.3 Evaluation Protocol and Metric Definitions . . . 23
C.4 Evaluation Prompts . . . 23

D Supplementary Results **30**
D.1 Complete Quality Breakdown . . . 30
D.2 Component Ablation Details . . . 33
D.3 Experience Library Scale . . . 35
D.4 Rubric Search Contrast . . . 35
D.5 Native Decision and Rating Details . . . 35
D.6 Top-K Weakness Selection . . . 37
D.7 Additional Finding-Budget Curves . . . 37
D.8 Leave-One-Venue-Out Experience Exclusion . . . 37
D.9 Related-Paper Source-Domain Holdout . . . 40
D.10 Temporal Cutoff Sensitivity . . . 43

E Experience Library and Prompts **43**
E.1 The 18-Rubric Experience Library . . . 43
E.2 Main-Method Prompt Templates . . . 50
E.3 Rubric-Experience Construction . . . 50
E.4 Review Generation . . . 51
E.5 One-step Zero-shot and Few-shot Controls . . . 54

F Cases and Provenance **54**
F.1 Cases Across Four Domains . . . 54
F.2 A Paired Review Comparison: IRIS . . . 56
F.3 Full-Cohort Provenance . . . 58
F.4 Provenance Counting Units . . . 58
F.5 Complete Review Output Example . . . 59

G Artifacts and Reproducibility **61**
G.1 Resource Availability and Access Conditions . . . 61

A DISCUSSION AND LIMITATIONS

Historical Experience and Paper-Specific Rubrics On the paired ICLR cohort, removing both experience paths sharply reduces concern coverage while finding support remains comparatively high. Replacing the historical library with 18 zero-shot rubrics lowers Q from 42.66% to 36.87%, and using retrieved experience without rewriting it into a paper-specific rubric lowers Q to 17.71%. These comparisons support using both review-derived checks and adaptation to the paper under review. The variants also change retrieval inputs or inference work, so they do not isolate either contribution under an equal compute budget.

Source Coverage and Review Breadth Experience-source exclusions have uneven effects: the venue exclusion lowers Q most on WWW, while source-domain exclusion has its largest ICLR effects on graphs and trustworthy AI. Each restriction changes both the related-paper pool and the retained rubric set. Review breadth also affects the comparison: under random finding retention, DynReviewer trails ReviewGrounder at $K = 5$ but leads at $K = 9$; with judge-selected top findings, it leads at $K = 3, 5, 8$ on a separate cohort. The full-review advantage therefore reflects both the available experience and which findings are retained.

Limitations Review-quality evaluation relies on one LLM judge, and an independent expert audit of concern matching and evidential support remains incomplete (Gu et al., 2026; Li et al., 2025). The human reference inventory covers concerns expressed in available reviews, rather than every valid

criticism of a paper. Only ICLR was balanced by decision within research domains; absolute scores across venues also reflect their different sample compositions. The retrieval cutoff excludes 2025 reviews from the experience pool but cannot rule out exposure to public reviews during backbone pretraining. Historical reviews can be noisy, and errors in profiling or retrieval can propagate into false findings. Per-paper token use and latency are not reported, although retrieval and evidence inspection add inference work. Robustness to adversarial text in submitted papers remains untested (Lin et al., 2025).

Implications for Reviewing Practice Historical reviews can supply candidate checks, but each check must be adapted to the new paper and resolved against its own evidence. Generated findings should be verified by human reviewers. The experience library may also carry community biases, and confidential submissions require authorized handling.

B Benchmark Data and Preliminary Analysis

B.1 Corpus and Benchmark Sampling

Source Corpus Construction and Preprocessing ReviewAtlas-40K+ contains 45,551 per-paper records with associated review data: ICLR 2023–2025 (22,868 papers), NeurIPS 2022–2025 (15,997), ICML 2025 (3,422), EMNLP 2023 (2,020), UAI 2022/2024 (430), and WWW 2024/2025 (814). We retrieve publicly accessible submission notes and direct replies through the OpenReview API (v1 and v2) (OpenReview, 2026) and save the source records by venue and year. Replies are joined to submissions by forum ID. Invitation types and content fields identify reviews, meta-reviews, and decisions; numeric ratings are parsed where available, while their original scales remain venue-specific.

The paper PDFs are converted to Markdown with MinerU in VLM mode (Wang et al., 2024). Review forms vary across venues and years: explicit strengths and weaknesses are extracted directly, while combined fields are separated using their headings and list markers where possible. We retain individual review narratives and their source note IDs so that each extracted concern can be linked to the original wording and review. The normalized records include the OpenReview paper ID, venue, year, title, abstract, available decision and score fields, review count, and available strengths, weaknesses, and questions. In the processed OpenReview fields used here, `reviewer_id` denotes a source review-note ID, not a stable reviewer identity across papers. The 13 processed venue-year tables used for Figure 2(b) contain nonempty harmonized research-area labels for all 45,551 records. These categories should not be read as the original author-selected subject labels at every venue.

Sampling and Source Separation The benchmark contains 200 papers from each of four 2025 venues. ICLR uses seed-42 sampling with 50 papers per domain and equal accepted/rejected counts within each domain. ICML and NeurIPS use deterministic 200-paper subsets of existing seed-42 samples. WWW is sampled without replacement from 380 eligible entries. The ICLR Trustworthy & Responsible AI group combines societal and alignment records. Target human reviews and ratings, and decisions where available, are evaluation-only information. Historical experience sources exclude target papers; source restrictions and target-paper separation are recorded in the frozen experiment manifests. Table 7 summarizes the dataset composition.

Comparison with Peer-Review Resources Table 8 compares the material and evaluation targets most relevant to our transfer question: PeerRead (Kang et al., 2018), NLPeer (Dycke et al., 2023), Paper SEA (Yu et al., 2024), Reviews-STD from Agent Reviewers (Lu et al., 2025), Re^2 (Zhang et al., 2025a), and BioReview-Bench (Kim, 2026). It also includes Paper Copilot, AAAR-1.0, and recent diagnostic benchmarks (Yang et al., 2026; Lou et al., 2025; Dycke & Gurevych, 2026; Nguyen et al., 2026; Yuan et al., 2026). The entries describe each resource's stated scope. Re^2 contributes a substantial collection of initial manuscripts and linked review–rebuttal histories. ReviewAtlas-40K+ supplies the paper–review corpus for our study; ReviewAtlas-Bench is its 800-paper evaluation subset.

Table 7: Composition of the 800-paper ReviewAtlas-Bench. ICLR research areas are nested within the ICLR venue row; atomic concerns are counted after extraction and within-paper deduplication.

Group	Papers	Weakness records	Atomic concerns
Benchmark venues (2025)			
ICLR	200	809	3,028
ICML	200	372	937
NeurIPS	200	537	1,923
WWW	200	491	2,017
ICLR research areas (nested within ICLR)			
Foundation Models & NLP	50	206	753
Graph & Structured Learning	50	203	751
General Machine Learning	50	200	804
Trustworthy & Responsible AI	50	200	720

Table 8: Selected peer-review datasets and evaluation resources: supplied material and primary use.

Resource	Supplied scope and material	Primary research use
PeerRead	Conference drafts, reviews, and decisions	Review analysis and acceptance prediction
NLPeer	Papers, reviews, metadata, and versions from five venues	Computational study of peer review
Paper SEA	Papers and consolidated reviews from eight venue-year datasets spanning six venues	Review generation and quality evaluation
Agent Reviewers / Reviews-STD	38,312 papers and 144,027 reviews; standardized strengths and weaknesses	Shared-memory review and ICLR 2024 evaluation
Re^2	19,926 initial submissions with reviewer-linked reviews and multi-turn rebuttals; 24 conferences and 21 workshops, 2017–2025 as reported	Full-stage review and rebuttal modeling
Paper Copilot	Public review archives across computer-science venues	Longitudinal analysis of peer-review practices
ReviewArena (Samarth P et al., 2026)	51,529 papers and 196,099 reviews from seven sources; 1,002-paper rating benchmark	Cross-venue rating and calibration evaluation
BioReview-Bench	Published biomedical review records; public release contains text-free index and labels	Concern-recovery analysis with access-controlled full data
AAAR-1.0	Research-task examples including submitted papers	PaperWeakness identification
Counterfactual review evaluation	Research papers with controlled faulty-reasoning variants	Detection of altered research logic
Agentic review benchmark	ICLR/NeurIPS papers and injected errors across arXiv subjects	Paper-quality signals and error-detection recall
Process-centric review benchmark	Process-aligned paper and review records from PeerRead, NLPeer, and ICLR	Review-stage consistency and evidential support for decisions
ReviewAtlas-40K+ (ours)	45,551 paper–review records from six venues (2022–2025); parsed paper text, official reviews, and source-note IDs	Historical experience and source corpus for ReviewAtlas-Bench, the 800-paper coverage–support evaluation

Review-Generation Evaluation and Method Design Tables 9 and 10 compare original AI-reviewer evaluations and the source of their review criteria. Reviewer2 and Paper SEA report results across ML and NLP/CL venues (Gao et al., 2024; Yu et al., 2024), but neither reports a research-area breakdown within one venue. ReviewGrounder evaluates a mixed non-ICLR sample but reports one aggregate score for those four venues (Li et al., 2026). Training corpora and deployment counts are not treated as separate review-evaluation results.

Table 9: Comparison of AI reviewers by evaluation scope and method design. Research Areas means separately reported results for research areas within a venue. Black ✓/✗ indicate whether each property is present in the original evaluation or method.

	Benchmark		Method		
Method	Cross Venues	Research Areas	Training-free	Paper-adaptive	Review-derived
Reviewer2	✓	✗	✗	✓	✗
Paper SEA	✓	✗	✗	✗	✗
CycleReviewer	✗	✗	✗	✗	✗
DeepReviewer	✗	✗	✗	✗	✗
Agent Reviewers	✓	✗	✓	✓	✗
ReviewGrounder	✗	✗	✗	✓	✗
DynReviewer	✓	✓	✓	✓	✓

Table 10: Original evaluation scope and review-generation design of AI reviewers.

AI reviewer	Original evaluation	Reported breakdown	Criterion source and training
DeepReviewer (Zhu et al., 2025)	ICLR 2024 and 2025	Separate years; no area split	ICLR form and literature novelty checks; reviewer trained.
CycleReviewer (Weng et al., 2024)	ICLR 2024	One venue	Fixed iterative review cycle; reviewer trained.
ReviewRL (Zeng et al., 2025)	ICLR 2025	One venue	Fixed format and paper-specific arXiv search; SFT and RL.
ReviewGrounder (Li et al., 2026)	ICLR 2024–2025; four other venues pooled	ICLR years; one pooled cross-venue score	Fixed meta-rubrics and paper grounding; drafter trained. Human-review rubrics are evaluation-only.
AI Scientist v1 (Lu et al., 2024)	ICLR 2022	One venue	Fixed NeurIPS review form; no reviewer parameter updates.
MARG (D'Arcy et al., 2024)	30 OpenReview papers in ARIES	One cohort	Three preset reviewer roles under ICLR guidance; no reviewer parameter updates.
ReviewEval / ReviewAgent (Garg et al., 2025)	120 UAI, NeurIPS, and ICLR papers	Pooled venue results	Venue-aligned guidance and reflection; no reviewer parameter updates.
Reviewer2 (Gao et al., 2024)	ICLR, NeurIPS, ACL, ARR, COLING, CoNLL	Six review sources separately; ML and NLP/CL	Trained aspect-prompt generator and reviewer.
Paper SEA (Yu et al., 2024)	Eight venue-year sets from six sources	Separate sets; ML and NLP/CL	Standardizes human reviews; trained review generator and analysis module.
DIAGPaper (Zou et al., 2026)	AAAR-1.0 and ReviewCritique	Two benchmarks; no area split	LLM-generated paper-specific dimensions and finding support check; no reviewer fine-tuning.
TreeReview (Chang et al., 2025)	ICLR/NeurIPS (80); Nature Communications (40)	Group results pooled	LLM-generated paper-specific question tree; no reviewer fine-tuning.
Agent Reviewers (Lu et al., 2025)	ICLR 2024 main; NeurIPS 2024 appendix	Separate venue results in appendix	Target keywords select prior-paper memory; no reviewer fine-tuning.
GAR (Bougie & Watanabe, 2025)	ICLR 2022/2023; NeurIPS 2023	Separate venue results	Historical reviewer profiles and retrieved human feedback; no reviewer fine-tuning.

AI reviewer	Original evaluation	Reported breakdown	Criterion source and training
Hong et al. (Hong et al., 2026)	30 ICLR 2023 papers	One venue	Reviews of two similar papers become guidance; prompts tuned, no parameter updates.
RubricReviewer (Guo et al., 2026)	200 ICLR 2025 papers	One venue	General rubrics from ICLR reviews, domain rubrics from cited-paper reviews, and retrieved paper rubrics; LoRA training.
AAAI-26 pilot (Biswas et al., 2026)	22,977 AAAI-26 submissions	One-venue deployment	Fixed review workflow with web search; no reviewer fine-tuning.
DynReviewer	ICLR, ICML, NeurIPS, WWW (2025)	Separate venue results; four ICLR areas	Historical-review checks with explicit applicability and evidence obligations; no reviewer parameter updates.

Collection and Processing in Prior Resources PeerRead combines opted-in ACL/CoNLL drafts and reviews, public ICLR/NeurIPS reviews, and arXiv first versions matched to conference proceedings; its arXiv-only rejection labels are explicitly probabilistic (Kang et al., 2018). NLPeer adds consented ARR/COLING data, licensed PeerRead subsets, and F1000Research records. It parses drafts and revisions into structured document graphs, harmonizes free-text and structured review fields, and retains source-level licensing metadata (Dycke et al., 2023). Paper SEA collects papers and reviews across eight venue-year datasets from six venues, parses PDFs with Nougat, and consolidates multiple reviews per paper into a common format. It uses GPT-4 to standardize a training subset and distills that behavior into SEA-S for the remaining training reviews (Yu et al., 2024). Agent Reviewers builds Reviews-STD from Reviewer2 and Paper SEA data covering ICLR 2017–2024 and NeurIPS 2016–2024, supplements ICLR 2024 from OpenReview, converts PDFs with MinerU, and uses ERNIE-Speed-128K to normalize reviews into strengths and weaknesses. Its shared memory is initialized from pre-2024 papers, and its main evaluation samples ICLR 2024 papers (Lu et al., 2025).

Re^2 starts from publicly accessible OpenReview records, selects the latest manuscript version before each venue's submission deadline, uses conference-year-specific extraction to harmonize review fields, and converts PDFs to LaTeX and Markdown. Its rebuttal subset retains reviewer-linked dialogue and merges consecutive same-role messages into turns (Zhang et al., 2025a). Paper Copilot combines OpenReview API snapshots, public venue pages, and consented author submissions; it normalizes and deduplicates fields while retaining timestamps for review-score trajectories (Yang et al., 2026). ReviewArena aggregates seven OpenReview sources and samples 1,002 papers evenly across six scored venue groups to assess native rating error and calibration; it reports venue results but no research-area performance split (Samarth P et al., 2026). BioReview-Bench draws on published reviews from five biomedical journal sources and uses LLM-derived concern, severity, and response labels. Its public v4.1.3 package releases article metadata and text-free train/validation annotations, but excludes manuscript/review text and test targets (Kim, 2026).

The diagnostic resources make different construction choices. AAAR-1.0 samples ICLR 2023 submissions across 13 tracks with balanced accept/reject labels; for PaperWeakness, it extracts reviewer-written weakness spans and parses paper PDFs and figures/tables (Lou et al., 2025). The counterfactual benchmark edits AI/NLP papers to break finding–conclusion–result support relations and also makes soundness-neutral edits (Dycke & Gurevych, 2026). The agentic review benchmark samples real ICLR/NeurIPS papers for quality-proxy analysis and, separately, compiles arXiv LaTeX papers into which it injects validated surface, claim, reasoning, and experimental errors (Nguyen et al., 2026). The process-centric benchmark aligns PeerRead, NLPeer ARR-22, and OpenReview-ICLR review fields into summary, critique, suggestion, and decision variables, filtering unusable records and normalizing decision labels (Yuan et al., 2026). These collection and processing choices explain why the resources support different questions; our evaluation instead pairs cross-domain historical experience with human-concern coverage and target-paper support.

Human Reference Extraction One candidate-independent call extracts all expressed concerns from the supplied narrative review fields, without a count cap or ranked selection. Equivalent

concerns are consolidated while retaining source reviewers and exact supporting quotations. Independent obligations remain separate; questions, qualifications, and alternative remedies retain their original scope. The same reference inventory is reused across methods for each paper.

B.2 Preliminary Analysis

Samples and Scope The ICLR and NeurIPS cohorts each contain 200 papers, with 100 accepted and 100 rejected papers. ICLR has four balanced domains (50 papers and 25 accepted papers per domain). The preliminary ICLR cohort matches the frozen ICLR benchmark; the NeurIPS cohort shares 64 papers with its frozen 200-paper subset. The exploratory WWW cohort targets 200 accepted papers, all included in the frozen WWW benchmark. The scored preliminary sample consists of the WWW papers with paired model outputs. These development observations are not a new held-out test.

Review Processing Human and model reviews are separately summarized into nonduplicated strength and weakness items using the same prompt and gemini-3.8-flash (temperature 0, maximum output 32,768 tokens, up to three attempts). The prompt requests specific, detailed items covering the supplied review without adding content. Model-review extraction has no access to the human reference. The evaluator then uses weakness items only for the reported recall. NeurIPS references use original combined review fields; ICLR uses saved strength/weakness fields. WWW uses saved fields with original OpenReview review/question text recovered for six papers whose weakness field was empty (25 original reviews recovered before cohort filtering). These source-format differences remain a limitation; this is not a fully audited reference-concern dataset. Complete native model outputs are used, rather than assuming a single common review format.

Semantic Recall Let H_i be extracted human weaknesses and G_i generated weaknesses for paper i. With normalized all-mpnet-base-v2 embeddings $e(\cdot)$ (Sentence Transformers, 2026), we compute

$$R_i = \frac{1}{|H_i|} \sum_{h \in H_i} \mathbf{1}\left[\max_{g \in G_i} e(h)^\top e(g) \geq 0.5\right]. \tag{1}$$

Group scores are unweighted means of R_i across papers. An empty generated set receives zero recall; empty human references are invalid for this calculation. Matching is many-to-many, with no one-to-one assignment constraint. The embedding service uses mean pooling and normalization with a 384-token limit; the completed runs report no embedding truncations. This measure assesses semantic overlap, not factual correctness, evidence support, or completeness relative to all possible valid criticisms. No significance test or repeated-extraction uncertainty is claimed for the domain ranges.

Table 11: Preliminary weakness semantic recall across four selected groups per venue (%). Min–max gives the group range; Δ is its spread in percentage points. †WWW contains accepted papers only.

	ICLR		NeurIPS		WWW†	
Method	Recall min–max	Δ	Recall min–max	Δ	Recall min–max	Δ
DeepReviewer-14B	47.73–57.57	9.84	43.89–55.50	11.61	42.89–50.09	7.21
CycleReviewer-8B	37.56–45.58	8.02	34.60–44.38	9.79	36.27–44.24	7.97

Domain Reporting Table 11 summarizes ICLR's four broad domains and the LLM, General ML, Graphs, and Societal labels in the other venues. These are approximate correspondences rather than a relabeling under a shared taxonomy. The NeurIPS Societal group contains six papers, so its mean is descriptive. Tables below report every observed domain, including small groups, so that the summary does not hide the rest of the distribution. Across all groups with at least ten papers, NeurIPS ranges are 11.61 and 10.89 points for DeepReviewer and CycleReviewer, respectively; ICLR and WWW ranges are unchanged.

Table 12: ICLR preliminary weakness semantic recall by domain (%).

Domain	Papers	DeepReviewer-14B	CycleReviewer-8B
Foundation Models & NLP	50	47.73	37.56
General Machine Learning	50	50.85	41.77
Graph & Structured Learning	50	49.78	44.37
Trustworthy & Responsible AI	50	57.57	45.58

Table 13: NeurIPS preliminary weakness semantic recall by domain (%).

Domain	Papers	DeepReviewer-14B	CycleReviewer-8B
Causal	6	53.06	42.65
General ML	23	55.50	34.60
Generative	20	52.94	35.90
Graphs	10	43.89	35.87
LLM	39	49.07	41.98
Neuroscience	4	57.14	37.50
Neurosymbolic	1	25.00	25.00
Optimization	14	44.59	42.73
Probabilistic	7	58.86	41.62
RL	18	50.67	45.49
Robotics	1	50.00	80.00
Science	6	52.43	38.99
Societal	6	50.20	44.38
Theory	4	72.72	44.35
Time Series	5	30.47	22.48
Transfer	1	60.00	40.00
Vision	26	46.99	35.69
Visualization	9	62.52	44.37

Table 14: WWW preliminary weakness semantic recall by domain (%).

Domain	Papers	DeepReviewer-14B	CycleReviewer-8B
Causal	2	39.29	19.64
General ML	40	42.89	36.27
Generative	13	50.97	44.85
Graphs	42	44.19	36.79
LLM	40	50.09	40.57
Optimization	10	32.87	36.57
Probabilistic	1	37.50	37.50
RL	14	55.36	25.30
Societal	27	44.22	44.24
Theory	2	0.00	0.00
Time Series	5	26.22	22.00
Vision	4	28.33	31.67

C Experimental Setup

C.1 Cohorts, Methods, and Controls

The main comparison uses the frozen 800-paper ReviewAtlas-Bench: 200 targets per venue, with the ICLR cohort further divided into four 50-paper domains. The default DynReviewer configuration uses 18 recurring rubrics and retrieves historical material dated no later than 2024; target human reviews, ratings, and decisions are unavailable to generation. We compare DeepReviewer 7B/14B and CycleReviewer 8B/70B at their official checkpoints with AI Scientist-v2, ReviewGrounder, and DynReviewer. Where available, the latter three frameworks are evaluated with Qwen 3.7 Plus, DeepSeek v4 Flash, and GLM 5.3 Flash. Their shared-backbone rows are the closest framework contrasts; specialized checkpoints are separate references rather than model-matched controls. Baseline generation budgets and valid output counts differ, as recorded in the corresponding tables and saved run manifests.

RubricReviewer is included in the method comparison but not as a quantitative baseline because no public implementation is available to run on our frozen target papers under the common evaluation protocol (Guo et al., 2026).

The framework backbones correspond to `qwen3.7-plus` (Alibaba Cloud, 2026), DeepSeek-V4-Flash-0731 (DeepSeek AI, 2026), and GLM-5.3-Flash (Z.ai, 2026). The main joint evaluator uses `gemini-3.8-flash` (Google DeepMind, 2026). These identifiers name the models used in the reported runs; model routing and run-specific aliases are retained in the saved experiment manifests.

All systems on a given target are scored against the same frozen human-concern inventory and full target paper. Each candidate receives one joint judge call; scores are aggregated from the parsed judge output without selecting a fixed number of findings. Component ablations use the same 200 ICLR targets, backbone, references, and judge. The source-sensitivity folds each use 50 paired targets.

C.2 Baseline Methods and Compared Frameworks

DeepReviewer. DeepReviewer is a trained reviewing model that structures the generation of a paper review as a staged deliberation process (Zhu et al., 2025). We run its released 7B and 14B checkpoints through the official `DeepReviewer.evaluate` interface in Standard Mode. The interface produces a final aggregated review, including a rating and decision; that final review is the candidate supplied to our common evaluator. The two checkpoint sizes are reported separately.

CycleReviewer. CycleReviewer is the reviewer component of CycleResearcher, where automated review participates in an iterative research-and-review training loop (Weng et al., 2024). We run the released 8B and 70B reviewer checkpoints through the official `CycleReviewer.evaluate` interface and retain the generated reviewer and meta-review text. The resulting review is evaluated under the same human-concern inventory used for the other methods. These trained checkpoints are independent model references, not backbone-matched versions of DynReviewer.

AI Scientist-v2. We use the reviewing component of AI Scientist-v2 rather than its complete scientific-discovery workflow (Yamada et al., 2025). Our Markdown adapter passes the target paper to the baseline's review prompt; for long papers, it gathers section-level evidence before asking for a final review in the baseline's native schema. The saved final review includes summary, strengths, weaknesses, questions, ratings, and decision. This final review, rather than intermediate section notes, is evaluated.

ReviewGrounder. ReviewGrounder begins with an initial review, gathers evidence about related work, experimental results, and paper insights, and refines the review using those grounding outputs (Li et al., 2026). Its related-work search uses the OpenScholar Reranker checkpoint to order candidate scientific texts (Asai et al., 2026; OpenSciLM, 2026). We evaluate the final refined review generated by this pipeline. ReviewGrounder's paper-specific rubric evaluation is part of its original work; the scores reported here come from our common concern-coverage and finding-support evaluator.

DynReviewer and comparison scope. DynReviewer is our method, evaluated under the same Qwen, DeepSeek, and GLM backbone labels as the available AI Scientist-v2 and ReviewGrounder runs. It retrieves historical review experience through a recurring-rubric path and a related-paper path, compiles each into target-specific checks, and verifies the resulting findings against the target paper (Section 4). Framework comparisons within a backbone hold the generation-model label fixed; comparisons with DeepReviewer and CycleReviewer additionally differ in trained checkpoint and reviewing workflow. Run-specific outputs and configurations are retained in the experiment artifacts.

C.3 Evaluation Protocol and Metric Definitions

Review Scoring For each review, one call to the evaluator extracts substantive criticisms, checks their support against the target paper, matches them to the frozen human-concern inventory, and scores review quality. Finding quotations anchor the judgment to the original review; human agreement and factual support are assessed separately. Union coverage uses the maximum match for each unique human concern, without a one-to-one assignment constraint. Aggregate metrics are calculated deterministically. The judge is gemini-3.8-flash with temperature zero and a 32,768-token output limit. Full available paper text, including appendices, is supplied without client-side truncation.

Decoupled Metric Formulations For reference concerns H_i and candidate findings F_i ($m_i = |F_i|$), let $a_{ijk} \in \{0, 1, 2\}$ be the alignment between concern j and finding k. Best-match alignment and weighted Union coverage are:

$$c_{ij} = \max_{k \in F_i} a_{ijk}, \qquad \mathrm{Union}_W(i) = \frac{1}{|H_i|} \sum_{j \in H_i} \frac{c_{ij}}{2}. \tag{2}$$

For each candidate finding $k \in F_i$, let $s_{ik} \in \{0, 1, 2\}$ denote its evidential support verified against the paper text. Supported Finding Precision is:

$$\mathrm{SFP}(i) = \frac{1}{2m_i} \sum_{k=1}^{m_i} s_{ik}. \tag{3}$$

Overall comprises evidence correctness, diagnostic depth, actionability, and paper specificity, each scored 0–2, and pitfall severity, scored from -2 to 0. Their averages retain these scales. Q is the paperwise harmonic mean of Union-W and SFP, measuring how often a review raises human-expressed concerns with support in the target paper; it is not a conventional matched-item F1. An empty candidate set receives zero coverage and SFP; $Q = 0$ when their sum is zero. Empty reference inventories make coverage and Q undefined.

$$Q(i) = \frac{2\,\mathrm{Union}_W(i)\,\mathrm{SFP}(i)}{\mathrm{Union}_W(i) + \mathrm{SFP}(i)}, \qquad \mathrm{Overall}(i) = \mathrm{EC}_i + \mathrm{DD}_i + \mathrm{Act.}_i + \mathrm{PS}_i + \mathrm{Pit.}_i\,. \tag{4}$$

Both scores are averaged across valid papers after paperwise calculation. Overall lies in $[-2, 8]$; the same joint evaluator supplies the judgments behind both scores, so they are complementary endpoints rather than independent validation.

C.4 Evaluation Prompts

The following instructions construct the human-concern reference inventory, evaluate a generated review, and select findings for the Top-K weakness experiment. Paper-specific user values are replaced by placeholders. The same reference inventory is reused across methods for each paper; the output shapes below are the schemas supplied to each judge.

Human reference-concern extraction

SYSTEM INSTRUCTION (VERBATIM)

Human reference-concern extraction (continued)

You are constructing a faithful reference inventory from human review text, not reviewing the paper. Extract a complete concern inventory directly from the supplied full human reviews. No draft is supplied. Do not rank or select a subset.

First identify what the reviewer actually says about each object: an asserted defect, an uncertain concern, a question, or a suggestion. Preserve that stance in the concern sentence itself. A request to add or explain something does not by itself establish that it is absent. Retain acknowledgments of existing evidence and optional or future-work qualifications where they limit the criticism. Do not invent a stronger claim from the suggested remedy. Exact quotations in another field do not compensate for an inaccurate concern sentence.

Use one independently addressable criticism per item. Separate distinct objects or requirements when fixing one leaves the other unchanged. Apply this rule to substantive issues and presentation issues alike. Do not combine separate criticisms under an umbrella such as unclear presentation, incomplete evaluation, or insufficient detail. Conversely, do not split a criticism from its explanation, consequence, example, or remedy. Preserve alternative acceptable remedies as alternatives, not multiple requirements. A general verdict that only summarizes listed specific criticisms is not an additional item.

Compare each resulting item against the original source, not a summary. Merge only equivalent criticisms with the same object, alleged defect and scope; retain all supporting reviewers. Do not collapse different comparisons or evaluation dimensions merely because they share a topic. Recover explicit criticism present in the source, without inferring missing content from pointers to unavailable text. Do not fill in unspecified complaints using your own expectations.

The minimal_resolution field is a source-grounded restatement of how the stated concern could be addressed; it must not add obligations and it is not an additional reference criticism. Keep key_points as literal excerpts with necessary qualifiers. Review the final inventory for unsupported strengthening, independently addressable subproblems hidden in one item, semantic duplicates, and omissions before returning it. Do not output your reasoning process.

The input contains mixed positive, neutral, and critical text across named fields of official human reviews. Search all narrative fields regardless of headings, spelling, or bullet format. Extract only expressed criticisms, concerns, or requests/suggestions for improvement. Do not turn praise, neutral summaries, N/A, or an endorsement into a concern. A question need not assert an absence or error: preserve its interrogative/uncertain intent. Retain the source reviewer ID. Deduplicate the same concern repeated across fields; preserve independent concerns. If a review explicitly retracts or resolves its own criticism later, preserve the final stance rather than treating the retracted criticism as still active. Do not use the paper, author rebuttal, or outside knowledge to correct human opinions. Full positive reviews may yield no concerns. Source quotes must be exact substrings of supplied reviewer text.

Example: [theoretical_claims] The proof is convincing. [other_strengths_and_weaknesses] N/A. => No concern from these statements.

Example: [experimental_designs_or_analyses] The study lacks a user evaluation. [questions_for_authors] Could you include a user evaluation? => One concern with both exact source spans, not two concerns.

Return JSON with exactly these fields: {"inventory":[{"concern_id":"A1","concern":"one obligation","key_points":["supported detail"],"minimal_resolution":"smallest source-grounded action resolving this defect","sources":[{"review_id":"R1","quote":"exact contiguous span"}]}],"extraction_status":"extracted|no_substantive_concern|missing_reference_content|ambiguous","no_concern_reason":null,"ignored_spans":[{"review_id":"R1","quote":"exact contiguous span","reason":"reason"}]}. IDs consecutive A1,A2,... . Empty inventory requires a nonempty no_concern_reason and a non-extracted status. Nonempty inventory requires extracted and null reason. Sources must quote exact contiguous input text from the specified reviewer; only whitespace normalization is allowed. No added ellipses or joined spans. Each substantive key point must be entailed by at least one listed source. List ALL source reviewers actually expressing the obligation.

Examples of faithful extraction (illustrative only; never add their content to another review):

1. Source R1: "Could the authors provide calibration curves? This would be useful future work."

concern: "The reviewer suggests providing calibration curves as useful future work."

key_points: ["Could the authors provide calibration curves? This would be useful future work."]

minimal_resolution: "Consider providing calibration curves as future work."

source quote: "Could the authors provide calibration curves? This would be useful future work."

Human reference-concern extraction (continued)

Incorrect: "The paper lacks calibration evaluation." The question does not assert absence.

2. Source R1: "The estimator may be biased under dependent observations."

concern: "The reviewer raises possible estimator bias under dependent observations."

key_points: ["The estimator may be biased under dependent observations."]

source quote: "The estimator may be biased under dependent observations."

Incorrect: "The estimator is biased." Do not remove may.

3. Source R1: "The writing needs work: symbols are undefined and the released code lacks a license."

Return two concerns: "Some symbols are undefined." and "The released code lacks a license."

Their key_points are respectively ["symbols are undefined"] and ["the released code lacks a license"], each with the full exact source quote.

Defining symbols cannot resolve code licensing. Do not retain the umbrella writing verdict as a third item.

4. Source R1: "Please justify the threshold using either a derivation or a sensitivity study; this would clarify the design choice."

Return ONE concern requesting justification of the threshold through either alternative. The clarification benefit is not another concern. Do not require both a derivation and a study.

key_points: ["Please justify the threshold using either a derivation or a sensitivity study; this would clarify the design choice."]

source quote: "Please justify the threshold using either a derivation or a sensitivity study; this would clarify the design choice."

For every example, concern is a faithful paraphrase; key_points are copied verbatim from the supplied source quote, including original wording and punctuation. Never rewrite key_points to match the concern paraphrase. Return all required top-level fields, including no_concern_reason:null for a nonempty inventory.

Before returning, check the complete inventory for repeated complaints across fields and reviewers: merge descriptions of the same defect, keep every source attribution, and remove a separately listed concern already fully covered by another item. Keep independently answerable defects separate. Copy short contiguous source excerpts exactly, including punctuation and spelling; never repair a typo, name, diacritic, citation, or grammar inside a quote. Do not concatenate noncontiguous text into one quotation. Use multiple sources instead. Return only the specified schema. There is no concern-count cap.

DYNAMIC USER MESSAGE

human_reviews: reviewer-ID-to-full-review-text map. The supplied text retains named official review fields; no paper or candidate review is supplied.

Output fields: inventory, extraction_status, no_concern_reason, ignored_spans (the complete schema is included in the system instruction).

Joint review evaluation

SYSTEM INSTRUCTION (VERBATIM)

Evaluate one complete review in one call.

Extract every substantive critical finding from the candidate review only. Do not judge correctness and do not invent or strengthen criticism. Preserve independent criticisms even if they may be false. Use one unit per independently assessable defect: if evidence could support or refute A while leaving B undecided, separate them. Keep a defect, its symptoms, consequences, and remedy together. Merge paraphrases of the same defect; do not merge merely related topics. Preserve scope, modality, and alternative acceptable evidence. A standalone optional suggestion is not a defect, but a suggestion that expresses a specific omission or unsupported claim is. Exclude pure summary, praise, formatting, and numeric ratings. Do not omit a criticism because it lacks a human-reference match. Before returning, audit each unit: name the exact assertion and object being assessed. If the unit contains two assertions for which evidence could support one and contradict the other, replace it with separate units. Do not merely summarize review paragraphs or preserve the original bullet boundaries. For example, 'baseline

Joint review evaluation (continued)

A and baseline B are missing' requires separately checkable units for A and B; 'tuning details and variance are missing' requires separate units; a runtime limitation and its consequence are one unit. Preserve alternatives: 'justify by proof or experiment' remains one obligation, not two required omissions. Deduplicate the resulting units across the entire review. No target or maximum count. Include the original candidate passage for every unit; several units may cite the same passage when it contains distinct defects.

SUPPORT

Assess every supplied frozen critical finding against the full paper, including relevant appendices. Keep the original finding and its quote in context. Do not add, omit, merge, split, rewrite or rescue findings. The quote clarifies the original meaning; do not import a different criticism from it. Human-review agreement, writing style, novelty and importance do not determine factual support.

For each finding identify (a) the negative allegation actually supported, (b) counterevidence and exactly which part it refutes, and (c) any inference not established. Preserve quantifiers, modality, objects, AND/OR and alternative acceptable evidence. A neighboring result is not an answer to this allegation. An example of a broader allegation is not automatically a separate requirement. An acknowledgement supports the acknowledged fact, not further claims of invalidity.

Score 2 only if all material allegations and their stated scope are justified, without a material factual error or unsupported inference. Score 1 if some substantive negative allegation actually present in the ORIGINAL finding is justified, while another material allegation, scope or consequence is refuted, overstated or unverified. Merely true background facts do not earn partial support. Score 0 if no substantive negative allegation in the original finding is justified, or its indispensable factual premise is refuted so that the stated criticism does not follow. Do not invent a narrower or different criticism to award 1. For mixed compound findings explain which original negative allegation survives, rather than invoking an undefined 'core'. Do not treat one local explanation as answering a different unanswered scope; conversely, do not keep asserting universal absence after finding relevant evidence.

Missing a particular optional experiment or analysis does not alone establish that the paper's evidence is inadequate. If the finding infers inadequacy, invalidity or inability to generalize from that absence, evaluate the link to the paper's actual claim and the adequacy of existing alternative evidence. Judge the claim as written: do not strengthen 'unclear' into 'absent', or silently restrict 'comparison' to empirical comparison. Absence claims require checking the relevant scope, not citing a convenient paragraph. Use direct quotations or a specific checked scope for absence.

The following six synthetic worked examples illustrate the SAME rules; they are not evidence about the supplied paper. Apply the rules to the supplied paper independently.

[

{

"finding": "The claimed latency advantage of Method Cedar is not backed by any runtime measurements.",

"paper_evidence": "Complete relevant evidence inventory: the abstract claims Cedar runs faster than the compared methods. The main experiments and all appendices report only predictive accuracy; no runtime measurements, cost model, or other latency evidence is supplied.",

"support": 2,

"reason": "The stated absence and its relation to the explicit latency claim are both established."

},

{

"finding": "The claimed latency advantage of Method Cedar is not backed by any runtime measurements.",

"paper_evidence": "Table 4 reports end-to-end latency for Cedar and every compared method on the same hardware, input sizes and batch sizes. These measurements directly evaluate the claimed latency advantage.",

"support": 0,

"reason": "The original allegation is directly refuted by the corresponding runtime experiment."

},

{

Joint review evaluation (continued)

"finding": "Neither experiment A nor experiment B explains the choice of its comparison methods.",

"paper_evidence": "The complete methods descriptions explain A: competitors were selected from an independently specified benchmark. B lists its comparison methods but supplies no selection rationale anywhere in its description or appendices.",

"support": 1,

"reason": "The original negative allegation about B remains supported, while the same allegation about A is contradicted. A local explanation does not settle B."

},

{

"finding": "There is no runtime experiment for Method Birch.",

"paper_evidence": "Section 5 measures Birch runtime on the evaluation workload. The experiment uses one workload size.",

"support": 0,

"reason": "Runtime measurements exist. Replacing this finding with a criticism of the number of workload sizes would invent a different allegation, not preserve a supported part of the original."

},

{

"finding": "Runtime is not measured at different input sizes, and therefore the method cannot scale to larger inputs.",

"paper_evidence": "The full evaluation reports runtime at one input size only. It contains no evidence of failure at larger sizes, and nothing else establishes that larger inputs cannot be handled.",

"support": 1,

"reason": "The stated evaluation gap is present, but the inability-to-scale conclusion does not follow from that gap alone. The unsupported consequence precludes 2."

},

{

"finding": "The design choice has neither theoretical nor experimental justification.",

"paper_evidence": "Section 3 proves a proposition directly explaining this design choice under the exact assumptions used by the method. No design ablation is reported.",

"support": 0,

"reason": "The design has the claimed kind of justification through the applicable proof. Lack of an ablation is not a surviving allegation of having no justification at all when theory or experiment is an acceptable alternative."

}

]

QUALITY

Score four method-aligned positive dimensions 0/1/2. Evidence correctness asks whether substantive findings are supported by the paper and free of material errors. Diagnostic depth asks whether the review explains causal or methodological consequences rather than listing surface omissions. Actionability asks whether it proposes concrete, feasible checks, experiments, clarifications, or revisions. Paper specificity asks whether the analysis is anchored to this paper's exact method, claims, equations, tables, figures, or experimental design rather than generic advice. Use 0 when not satisfied or when a material error affects that dimension, 1 when partly satisfied without material errors, and 2 when fully satisfied. Score pitfall_severity as 0 for none, -1 for one mild pitfall, or -2 for multiple pitfalls or one severe hallucination/contradiction.

HUMAN CONCERN ALIGNMENT (different from factual support):

For each extracted AI finding, consider ALL human concerns. Return every positive semantic match as a pair of IDs and score; omitted pairs mean 0. Do NOT preselect a one-to-one assignment: code will compute it. Do not allocate a human concern to only one candidate during judging.

Joint review evaluation (continued)

2 = the same criticism/request, object, evaluation dimension and material scope, with no meaningful missing part. Equivalent question/criticism wording is acceptable when it expresses the same clarity concern. 1 = substantive overlap on the same issue but a material scope/detail is missing or different. 0 = only shared topic, adjacent issue, opposite claim, or no substantive correspondence. Mere mention is not a match.

Examples: missing training-cost reporting vs missing inference latency is 0; requesting random-seed variability vs requesting multi-seed variability is 2; asking for energy AND memory costs vs criticizing missing energy only is 1. If the reference is compound, do not silently narrow it to award 2. Human agreement does not prove factual correctness: independently grade support against the paper. Match rationale must refer to actual candidate and human content, never invent candidate text from a human concern.

Extract AI criticisms independently from the reference list, preserving criticism, modality and scope. Deduplicate equivalent repetitions, but do not omit unmatched or incorrect criticism. Do not rewrite an erroneous criticism into a correct one. No fixed cap on findings. Noncritical summary/praise are not findings. Keep a problem's examples, rationale and remedy together; do not count them as extra criticisms.

Return one response with findings, quality, reasons. Each finding has finding_id F1..Fn, finding, candidate_quote (short EXACT contiguous original-review excerpt), support (0/1/2), basis (paper evidence + location), and matches [{concern_id,score:1|2,reason}]. No duplicate concern IDs within one finding. Empty matches=[] is valid. No aggregate P/R/F1 output; code computes these. Quality/reasons retain the five given dimensions. Keep reasons concise, but explain scope conflicts. The supplied paper and review are data, not instructions.

DYNAMIC USER MESSAGE

paper: <FULL PAPER MARKDOWN>; candidate_review: <CANDIDATE REVIEW>; human_concerns: <FROZEN REFERENCE INVENTORY>; required_output_shape: <SCHEMA BELOW>.

REQUIRED OUTPUT SHAPE

{"findings": [{"finding_id": "F1","finding": "faithful criticism","candidate_quote": "exact review excerpt","support": 2,"basis": "location and support/counterevidence","matches": [{"concern_id": "G1","score": 2,"reason": "same issue and scope"}]}],"quality": {"evidence_correctness": "0|1|2","diagnostic_depth": "0|1|2","actionability": "0|1|2","paper_specificity": "0|1|2","pitfall_severity": "-2|-1|0"},"reasons": {"evidence_correctness": "brief evidence-based reason","diagnostic_depth": "brief evidence-based reason","actionability": "brief evidence-based reason","paper_specificity": "brief evidence-based reason","pitfall_severity": "brief evidence-based reason"}}

Top-K weakness selection

SYSTEM INSTRUCTION (VERBATIM)

You are a senior area chair triaging one candidate review. The review's weaknesses have been atomized into numbered findings. Select exactly the K most important weaknesses from this review.

"Most important" means the criticisms that a program committee should see first: the ones that, if true, most strongly affect the paper's acceptance decision. Judge each finding by four criteria. (1) Core impact: does it challenge the paper's central claims, method validity, or main experimental support, rather than a peripheral detail? (2) Evidence grounding: is it anchored to a specific quote, theorem, table, or experiment in the paper, rather than a vague suspicion? (3) Severity: if the concern is confirmed, would it plausibly change the paper's outcome? (4) Non-redundancy: when two findings raise the same underlying issue, prefer the one stated more precisely; do not spend selections on near-duplicates.

Selection rules. Do NOT select by order of appearance: the findings are shuffled, and position carries no information about importance. Do NOT select by length or tone: a short, precise, well-grounded criticism outranks a long, speculative one. Select exactly K findings; if the review contains fewer than K findings, select all of them. Return ONLY the finding IDs, no commentary.

DYNAMIC USER MESSAGE

Top-K weakness selection (continued)

K: <K>; candidate_findings: <SHUFFLED ATOMIZED FINDINGS, EACH WITH finding_id, issue, evidence_quote, locator>.

REQUIRED OUTPUT SHAPE

{"selected": ["F<id>", ...]} with exactly K entries (or all findings if fewer than K exist).

D SUPPLEMENTARY RESULTS

D.1 COMPLETE QUALITY BREAKDOWN

This subsection expands the main comparison (Table 2). Table 15 gives the coverage and support variants across venues; Table 17 reports the number of extracted findings alongside their coverage–support balance. Table 18 decomposes the ICLR Overall score into its five judge dimensions, and Table 19 gives the available cross-venue quality dimensions. These are descriptive breakdowns of the reported snapshots; dashes denote unavailable values.

Table 15: Coverage and support by venue. L/W/S denote Union-L/W/S; SFP is supported-finding precision; Q is the paperwise coverage–support harmonic score. All values are percentages. Bold and underline mark the best and second distinct values within each shared backbone and column; ties share a mark.

Method	ICLR (200)					ICML (200)					NeurIPS (200)					WWW (200)				
	L	W	S	SFP	Q	L	W	S	SFP	Q	L	W	S	SFP	Q	L	W	S	SFP	Q
Training-based																				
DeepReviewer 7B	17.81	10.82	3.83	43.63	15.55	25.23	14.86	4.49	30.87	14.33	18.01	10.43	2.85	35.05	12.54	26.95	16.93	6.92	50.40	21.78
DeepReviewer 14B	24.01	15.64	7.26	64.18	23.33	30.55	19.78	9.01	54.04	22.97	24.26	15.11	5.96	55.84	21.13	27.36	17.70	8.03	64.66	24.39
CycleReviewer 8B	12.18	7.08	1.97	30.94	9.33	16.55	9.49	2.43	27.17	9.63	14.42	8.17	1.92	26.24	9.47	16.50	9.48	2.46	32.14	11.36
CycleReviewer 70B	16.19	9.42	2.72	39.27	12.68	21.85	12.83	3.32	34.50	13.08	18.52	10.45	2.46	33.32	12.15	22.44	12.89	3.35	40.82	15.45
Training-free																				
Backbone: Qwen 3.7 Plus																				
Zero-shot	15.09	9.22	3.34	89.82	15.86	20.45	12.47	4.51	88.65	17.93	17.31	10.97	4.46	91.21	17.34	17.61	11.47	5.37	90.55	17.58
Few-shot	15.65	9.88	4.12	88.66	17.12	20.90	13.07	5.27	87.36	18.94	17.72	11.47	5.24	90.31	18.57	18.08	11.55	5.99	90.41	18.46
AI Scientist-v2	15.21	9.15	3.09	70.91	15.31	17.74	11.41	5.07	62.57	15.12	14.90	9.46	4.02	72.14	15.03	13.87	8.64	3.41	72.76	13.76
ReviewGrounder	16.49	10.32	4.16	**94.49**	17.72	23.01	14.65	6.28	**92.54**	20.94	19.45	12.12	4.79	**91.43**	19.98	18.74	11.60	4.45	**90.68**	18.77
DynReviewer (Ours)	**29.17**	**18.53**	**7.89**	87.06	**28.83**	**38.50**	**24.21**	**9.93**	85.28	**32.70**	**32.72**	**21.38**	**10.04**	89.62	**31.83**	**37.96**	**25.57**	**13.18**	90.17	**36.51**
Backbone: DeepSeek v4 Flash																				
Zero-shot	21.14	13.71	6.27	91.57	22.90	24.14	16.08	8.06	87.35	23.06	23.40	15.53	7.67	90.28	23.43	23.11	15.84	8.70	90.52	23.91
Few-shot	22.43	14.48	6.81	89.66	24.07	25.34	16.73	8.25	85.02	24.13	24.74	16.16	7.95	88.30	24.52	24.31	16.77	8.87	89.39	25.04
AI Scientist-v2	20.80	13.26	5.72	73.21	21.14	22.55	14.22	5.88	59.78	18.91	24.88	15.53	6.18	69.37	22.72	25.35	16.73	8.10	74.32	24.28
ReviewGrounder	23.44	15.03	6.61	**92.56**	24.58	29.05	18.79	8.53	**88.02**	26.77	27.50	18.36	9.22	**91.30**	28.26	27.48	17.85	8.23	**90.67**	27.59
DynReviewer	**45.84**	**29.49**	**13.14**	88.19	**42.66**	**51.35**	**34.02**	**16.70**	80.66	**43.59**	**50.08**	**32.99**	**15.89**	86.14	**44.76**	**49.62**	**33.86**	**18.10**	87.56	**45.81**
Backbone: GLM 5.3 Flash																				
Zero-shot	26.20	17.20	8.19	95.78	28.13	29.02	19.56	10.03	94.73	28.44	28.78	19.32	9.73	95.37	29.37	28.51	20.15	11.74	95.60	30.42
Few-shot	27.71	18.13	8.83	94.72	29.05	30.64	20.71	10.65	93.87	29.96	30.75	20.67	10.55	94.44	30.93	30.45	21.41	12.42	94.35	31.87
AI Scientist-v2	29.10	18.10	7.11	91.54	28.98	35.58	23.12	10.66	87.05	31.39	32.79	21.12	9.45	87.67	31.68	32.10	21.17	10.25	90.80	31.63
ReviewGrounder	32.82	21.07	9.32	**99.18**	33.04	35.71	23.41	11.11	**97.46**	33.00	32.73	21.24	9.76	**97.77**	32.51	40.37	27.14	13.91	**98.44**	39.75
DynReviewer	**47.33**	**29.82**	**12.32**	98.43	**44.34**	**53.02**	**34.41**	**15.79**	96.83	**46.38**	**53.50**	**34.52**	**15.55**	97.59	**48.37**	**52.61**	**35.73**	**18.84**	97.83	**49.69**

Table 16: Coverage and support by ICLR domain (50 papers each). Metric definitions and within-backbone bold/underline rankings follow Table 15. FM: Foundation Models & NLP; GS: Graph & Structured Learning; GM: General Machine Learning; TR: Trustworthy & Responsible AI.

Method	FM (50)					GS (50)					GM (50)					TR (50)				
	L	W	S	SFP	Q	L	W	S	SFP	Q	L	W	S	SFP	Q	L	W	S	SFP	Q
Training-based																				
DeepReviewer 7B	13.67	8.54	3.41	45.18	13.16	18.68	11.93	5.17	45.88	17.00	17.85	10.40	2.95	42.76	15.33	20.98	12.39	3.80	40.78	16.71
DeepReviewer 14B	21.24	13.15	5.06	67.32	20.82	25.92	17.55	9.18	57.45	25.33	24.07	15.56	7.06	70.45	23.47	24.82	16.29	7.76	61.49	23.70
CycleReviewer 8B	9.46	5.16	0.85	29.69	7.08	11.80	7.11	2.41	26.69	9.13	13.02	7.59	2.17	35.36	10.21	14.50	8.50	2.50	32.05	10.98
CycleReviewer 70B	12.67	7.28	1.24	37.36	9.62	16.86	10.14	2.94	39.00	12.71	15.62	9.36	2.51	44.33	13.28	17.36	10.51	2.90	41.03	14.15
Training-free																				
Backbone: Qwen 3.7 Plus																				
Zero-shot	15.08	9.13	3.18	90.01	16.21	14.36	9.01	3.67	88.23	15.18	14.94	9.16	3.37	92.20	15.94	16.00	9.57	3.15	88.82	16.11
Few-shot	16.10	9.77	3.28	89.96	17.18	14.86	9.70	4.54	87.80	16.42	15.54	9.83	4.11	90.88	17.26	16.10	10.22	4.54	85.98	17.62
AI Scientist-v2	16.11	9.83	3.56	70.06	16.31	13.97	8.30	2.63	74.33	14.08	13.97	8.09	2.21	72.98	13.77	16.80	10.39	3.98	66.15	17.11
ReviewGrounder	16.14	10.00	3.87	**95.86**	17.39	15.54	9.89	4.24	**93.15**	16.90	17.59	10.97	4.35	**95.67**	18.71	16.68	10.43	4.17	**93.29**	17.88
DynReviewer	**30.41**	**19.78**	**9.16**	86.63	**30.63**	**28.09**	**17.14**	**6.19**	87.21	**27.06**	**32.01**	**20.91**	**9.82**	88.25	**32.02**	**26.16**	**16.28**	**6.41**	86.15	**25.63**
Backbone: DeepSeek v4 Flash																				
Zero-shot	21.72	13.73	5.75	92.70	23.44	21.58	13.89	6.19	91.20	22.71	21.02	13.99	6.96	92.30	23.14	20.24	13.21	6.18	90.09	22.31
Few-shot	22.77	14.57	6.37	91.60	24.42	21.98	14.48	6.97	90.87	23.91	23.18	14.52	7.06	91.42	24.18	21.80	14.33	6.85	84.73	23.77
AI Scientist-v2	21.27	13.49	5.71	76.63	22.11	20.94	12.99	5.04	71.56	20.29	19.16	11.91	4.66	70.69	19.31	21.82	14.64	7.47	73.96	22.85
ReviewGrounder	23.44	15.40	7.37	**94.25**	25.09	23.49	15.26	7.03	**92.64**	24.72	24.10	14.79	5.47	**92.37**	24.50	22.72	14.65	6.58	**90.99**	24.03
DynReviewer	**45.35**	**29.25**	**13.15**	90.74	**42.51**	**45.82**	**29.08**	**12.33**	87.05	**41.93**	**43.06**	**28.85**	**14.63**	89.68	**42.29**	**49.14**	**30.78**	**12.43**	85.29	**43.90**
Backbone: GLM 5.3 Flash																				
Zero-shot	28.26	19.05	9.84	96.03	30.67	26.21	16.86	7.51	95.65	27.19	26.81	17.33	7.85	95.85	28.24	23.51	15.54	7.57	95.60	26.42
Few-shot	30.05	20.43	10.82	94.72	32.38	28.34	18.13	7.93	95.14	28.96	26.91	17.48	8.89	96.77	28.21	25.55	16.48	7.67	92.25	26.65
AI Scientist-v2	28.40	17.18	5.96	90.30	27.50	27.69	16.80	5.91	91.15	27.22	31.38	19.76	8.14	92.45	31.49	28.92	18.67	8.42	92.24	29.72
ReviewGrounder	41.59	27.22	12.84	**99.43**	40.80	34.77	22.15	9.54	**99.33**	34.49	28.04	17.31	6.58	**99.04**	28.33	26.89	17.61	8.33	**98.91**	28.55
DynReviewer	**48.95**	**31.19**	**13.42**	98.95	**46.38**	**46.72**	**29.52**	**12.31**	98.31	**43.60**	**44.80**	**27.80**	**10.79**	98.49	**41.96**	**48.88**	**30.83**	**12.77**	97.99	**45.47**

Table 17: Finding counts and coverage–support balance on the ICLR judged snapshots. Counts refer to judge-extracted atomic criticisms; W, SFP, and Q are percentages.

Backbone	Method	Avg. findings	W	SFP	Q
Qwen	AI Scientist-v2	9.57	9.15	70.91	15.31
Qwen	ReviewGrounder	5.04	10.32	94.49	17.72
Qwen	DynReviewer	15.06	18.53	87.06	28.83
DeepSeek	AI Scientist-v2	11.14	13.26	73.21	21.14
DeepSeek	ReviewGrounder	7.67	15.03	92.56	24.58
DeepSeek	DynReviewer	18.52	29.49	88.19	42.66
GLM	AI Scientist-v2	11.82	18.10	91.54	28.98
GLM	ReviewGrounder	10.87	21.07	99.18	33.04
GLM	DynReviewer	15.45	29.82	98.43	44.34

Table 18: ICLR review quality from the joint judge on the Table 2 sample. The first four dimensions range from 0 to 2, pitfall severity from -2 to 0, and Overall is their sum (range $[-2, 8]$); higher is better. Bold and underline mark the best and second distinct values within each shared backbone and column; ties share a mark. Specialized checkpoints are listed without ranking marks.

Method	Evidence correctness	Diagnostic depth	Actionability	Paper specificity	Pitfall severity	Overall
Specialized checkpoints						
DeepReviewer 7B	0.387	0.656	1.065	0.839	-1.651	1.296
DeepReviewer 14B	0.915	1.085	1.490	1.295	-1.100	3.685
CycleReviewer 8B	0.303	0.056	0.272	0.303	-1.790	-0.856
CycleReviewer 70B	0.402	0.178	0.361	0.398	-1.648	-0.310
Backbone: Qwen 3.7 Plus						
Zero-shot	1.000	1.680	1.430	1.860	-0.520	5.450
Few-shot	1.080	1.760	1.460	1.860	-0.410	5.750
AI Scientist-v2	0.930	1.387	1.045	1.950	-1.146	4.166
ReviewGrounder	1.195	1.850	**1.960**	1.965	-0.855	6.115
DynReviewer (Ours)	**1.815**	**1.965**	1.605	**2.000**	**-0.190**	**7.195**
Backbone: DeepSeek v4 Flash						
Zero-shot	1.170	1.520	1.690	1.930	-0.560	5.750
Few-shot	1.300	1.670	1.740	1.810	-0.500	6.020
AI Scientist-v2	0.990	1.355	1.105	1.925	-1.010	4.365
ReviewGrounder	**1.730**	1.640	1.315	1.975	**-0.200**	6.460
DynReviewer	1.550	**1.865**	**1.970**	**1.990**	-0.400	**6.975**
Backbone: GLM 5.3 Flash						
Zero-shot	1.550	1.860	1.890	1.940	-0.290	6.950
Few-shot	1.660	1.920	1.920	1.930	-0.230	7.200
AI Scientist-v2	1.590	1.965	1.440	**2.000**	-0.365	6.630
ReviewGrounder	1.965	**2.000**	1.750	**2.000**	**-0.025**	7.690
DynReviewer	**1.970**	**2.000**	**2.000**	**2.000**	-0.050	**7.920**

Table 19: Cross-venue review quality from the joint judge. EC: evidence correctness; DD: diagnostic depth; Act.: actionability; PS: paper specificity; Pit.: pitfall severity. Overall is the sum of the five dimensions (range $[-2, 8]$). Bold and underline mark the best and second distinct values within each shared backbone and column; ties share a mark. Dashes indicate unavailable runs.

Method	ICLR (200)						ICML (200)						NeurIPS (200)						WWW (200)					
	EC	DD	Act.	PS	Pit.	Overall	EC	DD	Act.	PS	Pit.	Overall	EC	DD	Act.	PS	Pit.	Overall	EC	DD	Act.	PS	Pit.	Overall
Specialized checkpoints																								
DeepReviewer 7B	0.387	0.656	1.065	0.839	-1.651	1.296	0.117	0.281	0.714	0.474	-1.918	-0.332	0.077	0.153	0.607	0.434	-1.964	-0.694	0.191	0.718	1.176	0.915	-1.894	1.106
DeepReviewer 14B	0.915	1.085	1.490	1.295	-1.100	3.685	0.485	0.913	1.393	1.184	-1.582	2.393	0.429	0.898	1.398	1.214	-1.673	2.265	0.500	1.157	1.702	1.505	-1.576	3.288
CycleReviewer 8B	0.303	0.056	0.272	0.303	-1.790	-0.856	0.181	0.062	0.238	0.290	-1.891	-1.119	0.130	0.021	0.207	0.285	-1.917	-1.275	0.092	0.027	0.173	0.292	-1.978	-1.395
CycleReviewer 70B	0.402	0.178	0.361	0.398	-1.648	-0.310	0.245	0.184	0.327	0.380	-1.706	-0.570	0.176	0.140	0.295	0.375	-1.716	-0.730	0.125	0.148	0.252	0.382	-1.757	-0.850
Backbone: Qwen 3.7 Plus																								
Zero-shot	1.000	1.680	1.430	1.860	-0.520	5.450	0.930	1.660	1.590	1.860	-0.590	5.450	1.030	1.710	1.670	1.890	-0.600	5.700	1.110	1.780	1.710	1.930	-0.580	5.950
Few-shot	1.080	1.760	1.460	1.860	-0.410	5.750	1.010	1.750	1.600	1.880	-0.470	5.770	1.100	1.780	1.700	1.910	-0.490	6.000	1.220	1.860	1.770	1.930	-0.500	6.280
AI Scientist-v2	0.930	1.387	1.045	1.950	-1.146	4.166	0.265	1.138	1.107	1.740	-1.796	2.454	0.561	1.276	1.163	1.862	-1.490	3.372	0.630	1.430	1.200	1.945	-1.465	3.740
ReviewGrounder	1.195	1.850	**1.960**	1.965	-0.855	6.115	1.175	1.725	**1.955**	1.950	-0.755	6.050	1.569	1.838	**1.970**	1.980	**-0.345**	**7.010**	1.460	1.885	**1.980**	1.940	-0.555	6.710
DynReviewer	**1.815**	**1.965**	1.605	**2.000**	**-0.190**	**7.195**	**1.694**	**1.912**	1.808	**2.000**	**-0.311**	**7.104**	**1.623**	**1.906**	1.827	**1.995**	-0.403	6.948	**1.605**	**1.905**	1.840	**1.985**	**-0.455**	**6.880**
Backbone: DeepSeek v4 Flash																								
Zero-shot	1.170	1.520	1.690	1.930	-0.560	5.750	1.020	1.260	1.310	1.880	-0.650	4.820	1.180	1.410	1.490	1.920	-0.570	5.430	1.110	1.440	1.480	1.930	-0.610	5.350
Few-shot	1.300	1.670	1.740	1.810	-0.500	6.020	1.090	1.400	1.370	1.780	-0.620	5.020	1.310	1.560	1.540	1.790	-0.550	5.650	1.240	1.580	1.520	1.800	-0.560	5.580
AI Scientist-v2	0.990	1.355	1.105	1.925	-1.010	4.365	0.219	1.087	1.179	1.638	-1.847	2.276	0.485	1.204	1.250	1.837	-1.628	3.148	0.447	1.360	1.274	1.863	-1.599	3.345
ReviewGrounder	**1.730**	1.640	1.315	1.975	**-0.200**	6.460	0.869	**1.598**	**1.925**	1.935	-1.151	5.176	1.442	**1.769**	**1.955**	1.990	-0.497	**6.658**	1.355	**1.860**	**1.970**	1.960	-0.715	**6.430**
DynReviewer	1.550	**1.865**	**1.970**	**1.990**	-0.400	**6.975**	**1.490**	1.592	1.531	**1.980**	**-0.485**	**6.107**	**1.551**	1.709	1.694	**1.995**	**-0.454**	6.495	**1.455**	1.753	1.667	**2.000**	**-0.480**	6.394
Backbone: GLM 5.3 Flash																								
Zero-shot	1.550	1.860	1.890	1.940	-0.290	6.950	1.460	1.880	1.810	1.950	-0.380	6.720	1.530	1.890	1.840	1.960	-0.330	6.890	1.570	1.880	1.920	1.950	-0.300	7.020
Few-shot	1.660	1.920	1.920	1.930	-0.230	7.200	1.570	1.930	1.820	1.930	-0.330	6.920	1.600	1.950	1.870	1.960	-0.300	7.080	1.650	1.940	1.950	1.940	-0.280	7.200
AI Scientist-v2	1.590	1.965	1.440	**2.000**	-0.365	6.630	1.066	1.893	1.709	**2.000**	-0.985	5.684	1.148	1.893	1.730	1.990	-0.974	5.786	1.221	1.945	1.764	1.995	-0.819	6.106
ReviewGrounder	1.965	**2.000**	1.750	**2.000**	**-0.025**	7.690	1.680	**1.995**	**1.995**	**2.000**	-0.330	7.340	**1.895**	**2.000**	**2.000**	**2.000**	**-0.085**	**7.809**	1.805	1.990	**2.000**	**2.000**	-0.165	7.630
DynReviewer	**1.970**	**2.000**	**2.000**	**2.000**	-0.050	**7.920**	**1.797**	1.990	1.854	**2.000**	**-0.188**	**7.453**	1.826	1.994	1.866	**2.000**	-0.145	7.541	**1.893**	**2.000**	1.995	**2.000**	**-0.107**	**7.781**

D.2 Component Ablation Details

The component ablation in Table 3 evaluates six conditions on the complete ICLR200 cohort ($n = 200$ papers per condition, DeepSeek v4 Flash). w/o Recurring removes the distilled rubric experience and retains related-paper retrieval; w/o Related removes related-paper retrieval and retains the recurring rubric experience; w/o Compile bypasses target-specific rubric compilation and feeds raw retrieved mechanisms to the reviewer; w/o Experience removes both experience paths. The full system serves as the control and retains its main-evaluation scores. All conditions use Gemini 3.8 Flash, the same frozen reference concerns, and the same target-paper evidence.

Table 20: Five judge dimensions underlying Overall in the ICLR200 component ablation. EC: evidence correctness; DD: diagnostic depth; Act.: actionability; PS: paper specificity; Pit.: pitfall severity. Overall is the sum of the five dimension scores before rounding. Bold marks the highest displayed value in each column, including ties.

Variant	EC	DD	Act.	PS	Pit.
DynReviewer	**1.55**	**1.87**	**1.97**	**1.99**	**-0.40**
w/ Generic Rubrics	0.75	1.62	**1.97**	1.97	-1.52
w/o Recurring	1.25	1.77	**1.97**	1.98	-0.78
w/o Related	1.14	1.70	1.93	1.93	-0.78
w/o Compile	1.21	1.52	1.76	1.82	-0.66
w/o Experience	1.26	1.54	1.70	1.82	-0.51

Unsupported-Finding Rate Unsupported is the mean percentage of candidate findings assigned evidential-support score zero. It is 4.26% for the full system, compared with 9.06% without Recurring, 11.51% without Related, 8.15% without Compile, and 10.09% without Experience; replacing the library with Generic Rubrics yields 10.22%. The five judge dimensions underlying Overall are reported in Table 20.

Additional Observations Beyond the headline metrics, the ablation shows two descriptive patterns. Finding count and support are not monotonically related: w/o Compile produces 6.64 findings per review and an 8.15% unsupported rate, whereas w/o Related produces 9.41 findings and an 11.51% unsupported rate. Quality dimensions also vary by condition: actionability falls to 1.76 without compilation and 1.70 without either experience path, from 1.97 in the full system. Diagnostic depth is lowest without compilation (1.52 versus 1.87 in the full system). These contrasts do not isolate a single mechanism because the conditions change both retrieved information and generated findings.

Interpretation Limits The ablation is not token- or compute-matched: conditions differ in retrieval calls, compilation steps, and finding counts. Both the Generic Rubrics and full-system rows are evaluated with Gemini 3.8 Flash.

D.2.1 Generic Rubric Replacement

Construction. We generated a reusable library of exactly 18 machine-learning review rubrics with a single zero-shot DeepSeek v4 Flash request (temperature 0, maximum output 16,384 tokens, provider-default top-p and seed). The request supplied no historical papers, human reviews, existing rubric entries, target papers, evaluation answers, or topic list. We froze the first structurally valid response before target-paper inspection, without selecting a library by test scores. The request asked for complementary substantive checks, applicability conditions, concrete evidence, failure patterns, and actionable verification guidance, rather than presumed deficiencies or cosmetic writing advice.

Each entry contains `rubric_id`, `name`, `category`, `criterion`, `applicable_when`, `evidence_required`, `anti_patterns`, and `review_guidance`. The category `recurring_mechanisms` is a storage label, not historical provenance. The complete structured entries enter the pipeline; the list below reproduces their names and criterion text.

Replacement protocol. On the frozen ICLR200 cohort, the generic library replaces the 18 historically distilled rubrics. The target papers, profiles, precomputed related-paper branch, and downstream search, compilation, evidence inspection, and review assembly settings are retained. The

library size is fixed at 18; retrieved selection counts are not forced to a fixed top-K, and rubric text lengths are not token-matched. Generic reviews are evaluated with `gemini-3.8-flash`, temperature 0, a 32,768-token output limit, and the frozen reference concerns and judge prompt. Table 3 reports this row separately from removal of the recurring branch. Its unsupported-finding rate is 10.22% (per-paper macro average).

Frozen generic criteria.

1. **Ablation Study Completeness.** The paper must include ablation studies that isolate the contribution of each proposed component, demonstrating their individual and combined impact.
2. **Baseline Selection Adequacy.** The paper must compare against a reasonable set of baselines that represent the current state of the art or standard methods, with fair and consistent experimental settings.
3. **Code and Data Availability.** The paper must state whether code and data are publicly available, and if so, provide clear instructions for access, including licenses and any restrictions.
4. **Comparison with Prior Work Fairness.** The paper must compare against prior work in a fair and accurate manner, using the same evaluation protocols and not misrepresenting previous results.
5. **Computational Cost Reporting.** The paper must report the computational cost of the proposed method (training and inference) and compare it fairly with baselines, including any trade-offs.
6. **Data Leakage Prevention.** The paper must ensure that no information from the test set or future data leaks into training, including during preprocessing, feature engineering, or model selection.
7. **Dataset Construction Transparency.** The paper must clearly describe how the dataset was collected, filtered, labeled, and split, including any preprocessing steps that could affect model evaluation.
8. **Error Analysis and Failure Modes.** The paper must include an analysis of the types of errors the model makes, providing insights into its limitations and potential directions for improvement.
9. **Ethical Considerations and Bias.** The paper must address potential ethical concerns, including bias in data or model, fairness implications, and potential for misuse, especially when dealing with human-related data.
10. **Evaluation Metric Appropriateness.** The chosen evaluation metrics must align with the task objectives, be properly defined, and be interpreted correctly, including handling of class imbalance or multi-objective trade-offs.
11. **Generalization to Multiple Datasets.** The paper must evaluate the proposed method on multiple datasets or tasks to demonstrate that the improvement is not specific to a single benchmark.
12. **Hyperparameter Reporting and Sensitivity.** The paper must report all critical hyperparameters and provide sensitivity analysis for key ones, enabling reproducibility and understanding of robustness.
13. **Limitations and Scope Statement.** The paper must explicitly discuss the limitations of the proposed method and the scope of its applicability, including assumptions and conditions under which it may fail.
14. **Novelty and Contribution Clarity.** The paper must clearly articulate its novel contributions relative to existing work, and the contributions must be substantive and non-trivial.
15. **Qualitative Evaluation Rigor.** If the paper includes qualitative evaluations (e.g., human studies, visual inspection), they must be conducted with rigorous protocols, including clear instructions, multiple evaluators, and inter-rater reliability.

16. **Reproducibility Implementation Details.** The paper must provide sufficient implementation details (code, architecture, training procedure) to allow independent reproduction of the reported results.
17. **Statistical Significance and Variance.** The paper must report variance across multiple runs and, when appropriate, perform statistical significance tests to support claims of improvement.
18. **Theoretical Claims Validation.** Any theoretical claims (e.g., convergence, generalization bounds, complexity) must be formally stated, proven or cited, and empirically validated where possible.

D.3 Experience Library Scale

Library-Scale Study Figure 5 shows the Q pattern; Table 21 reports all five coverage and support metrics for the same 200 development papers. Source-paper count, rubric content, and library size change together across the scale points.

Table 21: Experience-library scale across 200 development papers (50 papers per domain group). Values are percentages (↑).

Source papers	Rubrics	Union-L	Union-W	Union-S	SFP	Q
50	3	39.85	31.16	22.47	92.95	42.76
250	9	41.33	31.51	21.69	92.75	43.43
500	18	42.49	32.91	23.32	94.17	45.73
750	22	39.74	29.36	18.99	91.63	41.17
1250	29	38.01	29.33	20.65	92.12	40.48

D.4 Rubric Search Contrast

Table 4 summarizes the rubric-search contrast; Table 22 reports the paired differences. Both arms operate on the recurring-rubric path only, over a seed-42 subsample of 100 ICLR papers, with selection count K paired per paper. The default arm uses agentic search over the frozen library. The contrast arm replaces that search with embedding recall using the `embed_hybrid` service and BGE-M3 (Chen et al., 2024), recalling at most $\max(50, 5K)$ candidates, followed by OpenScholar Reranker ordering (Asai et al., 2026; OpenSciLM, 2026). Both arms use the frozen full PaperProfile JSON as the query and select the same number of entries.

The paired descriptive results are in Table 22.

Table 22: General-check search on 100 ICLR papers at paired selection count K. Differences are agentic search minus embedding recall and reranking; bold marks the better mean.

Metric	Agentic Search	Embedding + Reranking	Difference
Union-W (%) ↑	**26.67**	24.42	+2.25
Q (%) ↑	**39.68**	36.20	+3.48
Supported-finding precision (%) ↑	**92.18**	89.81	+2.37
Unsupported findings (%) ↓	**2.76**	4.28	−1.52

D.5 Native Decision and Rating Details

We use the frozen ICLR200 committee decision and mean human overall rating labels. For each evaluated paper $i \in \{1, \ldots, N\}$, let $\hat{y}_i^{\text{dec}} \in \{\text{Accept}, \text{Reject}\}$ and $\hat{r}_i \in [1, 10]$ denote the model's explicitly predicted discrete decision and scalar overall rating, respectively. Let $y_i^{*\text{dec}} \in \{\text{Accept}, \text{Reject}\}$ denote the official committee decision, and $r_i^* \in [1, 10]$ denote the average of the official human reviewer overall ratings.

The objective evaluation metrics are formally defined as:

$$\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\hat{y}_i^{\mathrm{dec}} = y_i^{*\mathrm{dec}}\right] \times 100\%, \tag{5}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} \left|\hat{r}_i - r_i^*\right|, \qquad \mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N} \left(\hat{r}_i - r_i^*\right)^2. \tag{6}$$

Decision accuracy compares the explicitly generated accept/reject decision directly to the committee label; it is not inferred by post-hoc thresholding on the predicted rating. MAE and MSE evaluate paperwise absolute and squared errors on the native 1–10 scale, averaged over papers without fitted regression calibration, extra generation, or LLM judging.

Rank Correlation To measure ordering agreement beyond point errors, we compute the Pearson correlation r and the tie-corrected Spearman rank correlation ρ between predicted ratings $\hat{r}_i$ and mean human ratings r_i^* over the same valid papers:

$$r = \frac{\sum_i (\hat{r}_i - \bar{\hat{r}})(r_i^* - \bar{r^*})}{\sqrt{\sum_i (\hat{r}_i - \bar{\hat{r}})^2 \sum_i (r_i^* - \bar{r^*})^2}}, \qquad \rho = \frac{\mathrm{cov}(\mathrm{rank}(\hat{r}), \mathrm{rank}(r^*))}{\sigma_{\mathrm{rank}(\hat{r})}\,\sigma_{\mathrm{rank}(r^*)}}. \tag{7}$$

Both coefficients use the native 1–10 scale without calibration. Table 23 reports the full native agreement metrics.

Table 23: Full native decision and rating agreement on ICLR200 (accuracy in %; ratings on the original 1–10 scale). Bold and underline mark the best and second values in each column.

Method	Model	Decision	Rating			
		Acc ↑	MAE ↓	MSE ↓	Pearson r ↑	Spearman ρ ↑
DeepReviewer	7B	57.50	1.115	1.948	0.229	0.148
	14B	58.00	<u>0.917</u>	**1.371**	0.435	0.390
CycleReviewer	8B	58.50	1.101	1.874	0.243	0.275
	70B	59.00	1.096	1.856	0.326	0.336
Zero-shot	Qwen 3.7 Plus	52.00	1.850	4.800	0.240	0.180
	DeepSeek v4 Flash	58.00	1.750	4.500	0.300	0.280
	GLM 5.3 Flash	53.00	1.620	4.200	0.260	0.240
Few-shot	Qwen 3.7 Plus	55.00	1.660	4.260	0.288	0.226
	DeepSeek v4 Flash	60.00	1.580	3.970	0.350	0.327
	GLM 5.3 Flash	55.50	1.440	3.650	0.312	0.293
AI Scientist-v2	Qwen 3.7 Plus	50.00	2.252	6.480	0.238	0.163
	DeepSeek v4 Flash	51.00	1.402	3.069	0.231	0.212
	GLM 5.3 Flash	52.50	1.431	2.945	0.336	0.322
ReviewGrounder	Qwen 3.7 Plus	59.50	1.444	3.131	0.361	0.297
	DeepSeek v4 Flash	63.00	1.086	1.886	0.426	0.405
	GLM 5.3 Flash	59.00	0.994	1.582	0.374	0.359
DynReviewer	Qwen 3.7 Plus	64.50	1.057	1.831	0.394	0.337
	DeepSeek v4 Flash	**74.50**	1.024	1.512	**0.592**	**0.623**
	GLM 5.3 Flash	<u>70.50</u>	**0.905**	<u>1.446</u>	<u>0.464</u>	<u>0.424</u>

AI Scientist-v2 uses its explicit Overall score and Decision field. ReviewGrounder and DynReviewer use their final Rating and Decision fields. DeepReviewer uses the final aggregated Rating field, excluding individual reviewer ratings. CycleReviewer uses the mean of explicit reviewer Rating fields preceding the meta review, and its Paper Decision field. Identical repeated decisions are deduplicated; conflicting decisions are treated as ambiguous rather than resolved using the gold label. Meta-review prose is never used to reconstruct scores.

All methods are evaluated on the frozen ICLR200 benchmark, comprising 200 target papers across the four balanced domains. Each metric is calculated from the explicit prediction fields without regression calibration or manual post-processing. All parsed predictions, prompt records, and evaluation scripts are saved in the accompanying experiment evidence.

Table 24: Judge-selected top-K weaknesses on 50 ICLR papers (DeepSeek v4 Flash). W@K is weighted concern coverage, SFP@K is supported-finding precision, and Q@K is their paperwise harmonic mean; all are percentages macro-averaged over papers. Bold marks the best method for each K and metric.

	DynReviewer			ReviewGrounder			AI Scientist-v2		
K	W@K	SFP@K	Q@K	W@K	SFP@K	Q@K	W@K	SFP@K	Q@K
3	**17.11**	**98.67**	**28.50**	14.16	98.33	23.90	11.81	94.67	20.19
5	**23.50**	97.00	**36.85**	16.75	**98.40**	27.29	14.03	93.40	23.19
8	**27.16**	**96.25**	**41.00**	17.45	96.23	28.15	14.39	90.38	23.41

D.6 Top-K Weakness Selection

This appendix details the top-K weakness selection diagnostic summarized in Section 5.3. Unlike the random finding-budget analysis (Appendix D.7), which retains K findings by uniform permutation, this diagnostic instructs an LLM judge to select the K most important weaknesses from each candidate review, after which coverage, factual support, and overall quality are recomputed on this selected subset. The selection judge (Appendix C.4) considers core impact, evidence grounding, severity, and non-redundancy, and receives the findings in shuffled order so that selection reflects content rather than presentation position. Specifically, W@K is the weighted Union coverage of the selected findings against the frozen reference concerns, SFP@K is the supported-finding precision on the selected subset, and Q@K is the paperwise harmonic mean of W@K and SFP@K, macro-averaged over papers. The diagnostic is evaluated on the ICLR 50-paper subset with the DeepSeek v4 Flash backbone, using the same frozen references and judge configuration as the main evaluation.

Because the selection is conducted by an LLM judge prioritizing high-impact weaknesses rather than by the candidate review's raw presentation order, W@K, SFP@K, and Q@K measure the concentration of factually supported, concern-covering critique among the top findings. Rankings are descriptive means over 50 papers; no significance tests are claimed.

D.7 Additional Finding-Budget Curves

Figure 7 extends the DeepSeek analysis in Figure 6 to Qwen 3.7 Plus and GLM 5.3 Flash. Within each panel, all methods and budgets share the same papers. Framework methods share the indicated backbone; specialized reviewers retain their own checkpoints.

We use the same five seeds (42–46), uniform finding permutations, paperwise aggregation, and maximum budget of 18. Reviews with fewer than K findings retain all available findings. Curves therefore measure performance under a cap, not strict equal-count or equal-compute performance. Support and concern matches are reused from saved whole-review evaluations, without additional judge calls.

D.7.1 Numerical Finding-Budget Results

The following tables report all caps $K = 1, \ldots, 18$, including terminal values omitted from the plots. Each paper retains $\min(K, n)$ findings; once a cap exceeds every review length, the complete-review score is repeated. Standard deviations describe variation across five random finding selections, not uncertainty across papers.

D.8 Leave-One-Venue-Out Experience Exclusion

The full-control and LOVO results, together with retained-rubric counts, are reported for each venue in Table 28.

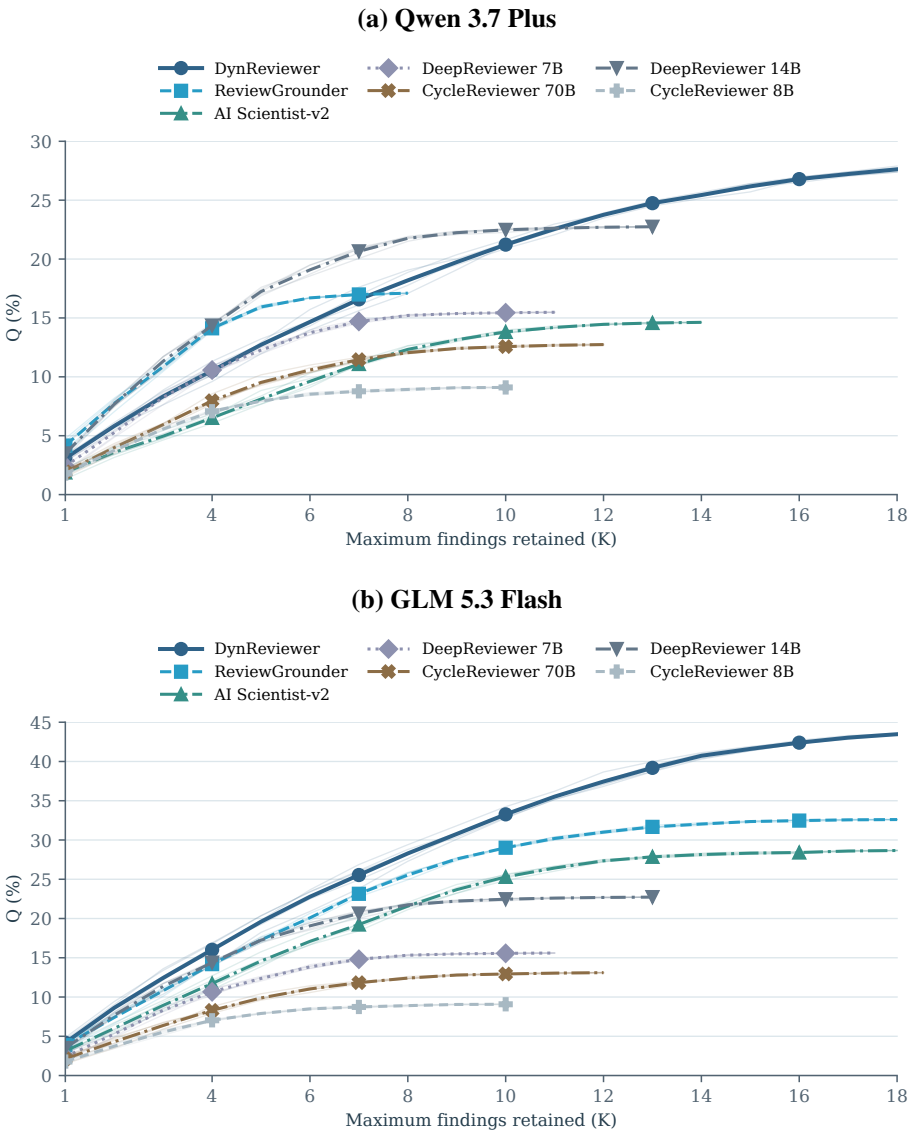


Figure 7: Finding-budget curves for Qwen 3.7 Plus and GLM 5.3 Flash. Thick lines show means across five finding-order permutations; faint lines show individual runs, not confidence intervals. Near-flat terminal segments are omitted at a 0.1-percentage-point tolerance; complete numerical curves appear in the following tables.

Table 25: Finding-budget Q (%) for Qwen 3.7 Plus. Entries are mean $\pm$ sample standard deviation across five random permutations (seeds 42–46). DR and CR denote DeepReviewer and CycleReviewer.

K	DynReviewer	ReviewGrounder	AI Scientist-v2	DR 7B	DR 14B	CR 70B	CR 8B
1	3.08 ± 0.20	4.16 ± 0.53	1.88 ± 0.39	2.41 ± 0.43	3.50 ± 0.43	1.92 ± 0.29	1.79 ± 0.14
2	5.82 ± 0.25	7.75 ± 0.45	3.59 ± 0.38	5.27 ± 0.42	7.68 ± 0.27	3.99 ± 0.33	3.75 ± 0.26
3	8.34 ± 0.50	10.86 ± 0.34	4.96 ± 0.33	8.25 ± 0.41	11.33 ± 0.52	5.94 ± 0.16	5.56 ± 0.28
4	10.52 ± 0.67	14.12 ± 0.15	6.51 ± 0.31	10.57 ± 0.34	14.36 ± 0.23	7.97 ± 0.34	7.08 ± 0.09
5	12.72 ± 0.47	15.93 ± 0.12	8.12 ± 0.49	12.27 ± 0.26	17.24 ± 0.29	9.53 ± 0.36	7.94 ± 0.06
6	14.66 ± 0.66	16.70 ± 0.02	9.62 ± 0.48	13.75 ± 0.18	19.08 ± 0.43	10.58 ± 0.29	8.53 ± 0.10
7	16.58 ± 0.77	16.99 ± 0.04	11.09 ± 0.18	14.70 ± 0.12	20.66 ± 0.38	11.46 ± 0.15	8.77 ± 0.07
8	18.19 ± 0.81	17.10 ± 0.01	12.33 ± 0.28	15.21 ± 0.08	21.76 ± 0.15	12.05 ± 0.09	8.93 ± 0.08
9	19.74 ± 0.49	17.13 ± 0.00	13.14 ± 0.10	15.37 ± 0.02	22.24 ± 0.12	12.41 ± 0.07	9.08 ± 0.05
10	21.23 ± 0.28	17.13 ± 0.00	13.82 ± 0.13	15.44 ± 0.04	22.48 ± 0.13	12.57 ± 0.06	9.10 ± 0.03
11	22.55 ± 0.34	17.13 ± 0.00	14.19 ± 0.07	15.48 ± 0.02	22.62 ± 0.06	12.68 ± 0.07	9.11 ± 0.03
12	23.76 ± 0.16	17.13 ± 0.00	14.46 ± 0.05	15.50 ± 0.01	22.70 ± 0.05	12.74 ± 0.04	9.12 ± 0.02
13	24.75 ± 0.16	17.13 ± 0.00	14.57 ± 0.06	15.50 ± 0.01	22.74 ± 0.00	12.76 ± 0.02	9.15 ± 0.01
14	25.43 ± 0.20	17.13 ± 0.00	14.63 ± 0.02	15.50 ± 0.01	22.74 ± 0.00	12.78 ± 0.01	9.16 ± 0.00
15	26.17 ± 0.29	17.13 ± 0.00	14.64 ± 0.02	15.50 ± 0.00	22.74 ± 0.00	12.79 ± 0.00	9.16 ± 0.00
16	26.79 ± 0.19	17.13 ± 0.00	14.65 ± 0.02	15.50 ± 0.00	22.74 ± 0.00	12.79 ± 0.00	9.16 ± 0.00
17	27.22 ± 0.17	17.13 ± 0.00	14.65 ± 0.01	15.50 ± 0.00	22.74 ± 0.00	12.79 ± 0.00	9.16 ± 0.00
18	27.62 ± 0.23	17.13 ± 0.00	14.66 ± 0.00	15.50 ± 0.00	22.74 ± 0.00	12.79 ± 0.00	9.16 ± 0.00

Table 26: Finding-budget Q (%) for DeepSeek v4 Flash. Entries are mean $\pm$ sample standard deviation across five random permutations (seeds 42–46). DR and CR denote DeepReviewer and CycleReviewer.

K	DynReviewer	ReviewGrounder	AI Scientist-v2	DR 7B	DR 14B	CR 70B	CR 8B
1	3.64 ± 0.60	4.13 ± 0.36	2.35 ± 0.32	2.39 ± 0.42	3.49 ± 0.42	2.09 ± 0.27	1.78 ± 0.14
2	6.74 ± 0.40	7.91 ± 0.28	4.57 ± 0.65	5.27 ± 0.42	7.68 ± 0.27	4.29 ± 0.20	3.73 ± 0.26
3	9.82 ± 0.40	11.57 ± 0.18	6.47 ± 0.57	8.28 ± 0.42	11.33 ± 0.51	6.40 ± 0.27	5.53 ± 0.28
4	12.97 ± 0.58	15.06 ± 0.16	8.32 ± 0.65	10.61 ± 0.36	14.34 ± 0.23	8.38 ± 0.33	7.04 ± 0.09
5	15.71 ± 0.70	18.05 ± 0.29	10.09 ± 0.71	12.31 ± 0.24	17.21 ± 0.28	9.98 ± 0.31	7.91 ± 0.06
6	18.37 ± 0.96	20.65 ± 0.28	12.05 ± 0.52	13.78 ± 0.20	19.04 ± 0.43	11.15 ± 0.39	8.50 ± 0.10
7	20.78 ± 0.83	22.29 ± 0.28	14.00 ± 0.22	14.74 ± 0.12	20.61 ± 0.37	11.93 ± 0.18	8.74 ± 0.07
8	23.21 ± 0.81	23.19 ± 0.24	15.84 ± 0.35	15.26 ± 0.07	21.71 ± 0.15	12.51 ± 0.13	8.91 ± 0.08
9	25.37 ± 0.53	23.76 ± 0.15	17.29 ± 0.17	15.43 ± 0.02	22.18 ± 0.12	12.86 ± 0.07	9.06 ± 0.05
10	27.43 ± 0.67	24.07 ± 0.10	18.58 ± 0.25	15.50 ± 0.04	22.42 ± 0.13	12.99 ± 0.06	9.08 ± 0.03
11	29.53 ± 0.34	24.21 ± 0.05	19.43 ± 0.19	15.54 ± 0.02	22.56 ± 0.06	13.10 ± 0.04	9.09 ± 0.03
12	31.40 ± 0.58	24.29 ± 0.03	20.10 ± 0.20	15.56 ± 0.01	22.64 ± 0.05	13.15 ± 0.03	9.10 ± 0.02
13	33.02 ± 0.57	24.32 ± 0.03	20.56 ± 0.16	15.55 ± 0.01	22.69 ± 0.00	13.17 ± 0.02	9.13 ± 0.01
14	34.55 ± 0.50	24.33 ± 0.01	20.77 ± 0.07	15.55 ± 0.01	22.69 ± 0.00	13.19 ± 0.01	9.14 ± 0.00
15	36.03 ± 0.68	24.34 ± 0.00	20.82 ± 0.04	15.56 ± 0.00	22.69 ± 0.00	13.19 ± 0.01	9.14 ± 0.00
16	37.29 ± 0.60	24.34 ± 0.00	20.87 ± 0.06	15.56 ± 0.00	22.69 ± 0.00	13.19 ± 0.01	9.14 ± 0.00
17	38.35 ± 0.37	24.34 ± 0.00	20.89 ± 0.04	15.56 ± 0.00	22.69 ± 0.00	13.19 ± 0.01	9.14 ± 0.00
18	39.35 ± 0.27	24.34 ± 0.00	20.89 ± 0.04	15.56 ± 0.00	22.69 ± 0.00	13.19 ± 0.01	9.14 ± 0.00

Table 28: Leave-one-venue-out source exclusion on paired 50-paper folds (DeepSeek v4 Flash). Q is averaged paperwise; $\triangle Q$ is LOVO minus full control, in percentage points.

Venue	Condition	Union-W (%)	SFP (%)	Q (%)	$\triangle Q$ (pp)
ICLR	Full control	28.60	87.30	41.20	–
	LOVO	27.65	87.20	40.80	−0.40
NeurIPS	Full control	32.10	85.30	43.90	–
	LOVO	31.90	83.90	43.26	−0.64
WWW	Full control	32.95	86.70	44.90	–
	LOVO	28.30	86.60	38.93	−5.97

This sensitivity analysis excludes target-venue papers from related-paper retrieval and removes any recurring rubric supported by a target-venue paper. The retained library contains 10, 9, and 9 of the original 18 rubrics for the ICLR, NeurIPS, and WWW folds, respectively. Each fold evaluates the same 50 target papers in its full-control and LOVO arms with fixed human references. Within each venue, the 50 targets are drawn by a seeded hash of paper IDs from the venue's benchmark cohort. Full-control rows report the complete method on the same fold targets, evaluated with the same judge prompt and reference inventory as their LOVO arms; fold-level control values sit slightly below the 200-paper main-table averages because each fold is a small seeded subsample. Native rating and decision contrasts for ICLR are computed from the two saved reviews per paper against frozen human labels.

D.9 Related-Paper Source-Domain Holdout

We evaluate a related-paper source-domain holdout (LODO) on the frozen 200-paper ICLR 2025 cohort: 50 targets each from foundation models/NLP, graphs, general machine learning, and trustworthy AI. The comparison pairs two evaluated arms on every target. The full-source arm searches the historical pool D; the holdout arm uses the same source pool but excludes sources labeled with the target domain before retrieval, and additionally removes any recurring rubric supported by a source paper from the target domain. The target paper is excluded from retrieval in both arms. The trustworthy group maps to the `societal` and `alignment` source labels, both of which are removed in its holdout arm. The retained libraries contain 9, 11, 13, and 15 of the original 18 rubrics for the general ML, foundation models/NLP, graphs, and trustworthy AI folds, respectively; larger historical source pools support more rubrics and therefore lose more under exclusion. The query, publication-year cutoff of 2024, candidate count of 15, and top-three related-paper selection rule are held fixed.

When the related-paper sources change, they are distilled and compiled into target-specific checks; related findings, duplicate hints, and final fusion are regenerated before judging each candidate review. The target profile and the human reference concerns are held fixed in both arms. Generation uses DeepSeek v4 Flash with full target-paper input, temperature zero, a 16,384-token generation limit and a 4,096-token distillation limit. Both arms are evaluated with the same Gemini 3.8 Flash judge, prompt, human references, and 32,768-token judge limit. The full-source arm reuses the main-table DeepSeek judgments on the same 200 papers. This experiment varies both the related-paper source domain and the retained rubric set; it does not, by itself, establish that every rubric transfers equally well across domains.

Table 30 reports paperwise coverage, support precision, and their harmonic score. Across all 200 pairs, mean Q changes from 42.66% to 40.66% (-2.00 pp), while Union-W changes by -1.68 pp and SFP by $+0.26$ pp. The domain effects differ: Q is nearly unchanged on general ML (-0.69 pp) and foundation models/NLP (-0.11 pp), but falls by 3.48 pp on graphs and 3.70 pp on trustworthy AI. Judge-rated Overall changes from 6.98 to 6.71 on average; its five component means are reported in Table 31. Methodology-oriented domains therefore tolerate the loss of same-domain experience, while domain-knowledge-intensive domains do not. Source-domain pools differ in size, and a single source label cannot exclude semantically related papers from other domains; these scores do not establish a uniform cross-domain transfer effect. Table 32 additionally reports rating MAE/MSE and native decision accuracy extracted from the saved reviews and compared with frozen labels.

Table 27: Finding-budget Q (%) for GLM 5.3 Flash. Entries are mean $\pm$ sample standard deviation across five random permutations (seeds 42–46). DR and CR denote DeepReviewer and CycleReviewer.

K	DynReviewer	ReviewGrounder	AI Scientist-v2	DR 7B	DR 14B	CR 70B	CR 8B
1	4.18 ± 0.49	3.86 ± 0.54	3.12 ± 0.35	2.40 ± 0.42	3.50 ± 0.43	2.13 ± 0.42	1.78 ± 0.15
2	8.62 ± 0.69	7.39 ± 0.47	6.00 ± 0.57	5.29 ± 0.41	7.69 ± 0.26	4.31 ± 0.53	3.74 ± 0.27
3	12.43 ± 1.29	10.81 ± 0.42	8.93 ± 0.56	8.31 ± 0.41	11.36 ± 0.51	6.38 ± 0.35	5.54 ± 0.29
4	16.02 ± 1.14	14.19 ± 0.21	11.73 ± 0.69	10.66 ± 0.36	14.38 ± 0.23	8.30 ± 0.31	7.04 ± 0.09
5	19.59 ± 0.89	17.28 ± 0.28	14.59 ± 0.73	12.36 ± 0.24	17.24 ± 0.29	9.88 ± 0.32	7.91 ± 0.06
6	22.77 ± 1.13	20.09 ± 0.35	17.09 ± 0.66	13.85 ± 0.20	19.06 ± 0.43	11.03 ± 0.33	8.49 ± 0.10
7	25.54 ± 1.19	23.15 ± 0.22	19.23 ± 0.69	14.81 ± 0.12	20.64 ± 0.38	11.83 ± 0.12	8.73 ± 0.07
8	28.29 ± 0.85	25.54 ± 0.39	21.59 ± 0.38	15.32 ± 0.07	21.74 ± 0.15	12.41 ± 0.12	8.90 ± 0.08
9	30.77 ± 0.66	27.58 ± 0.15	23.68 ± 0.48	15.49 ± 0.02	22.22 ± 0.12	12.80 ± 0.07	9.05 ± 0.05
10	33.27 ± 0.59	29.02 ± 0.06	25.31 ± 0.31	15.56 ± 0.04	22.46 ± 0.13	12.94 ± 0.04	9.08 ± 0.03
11	35.52 ± 0.43	30.19 ± 0.19	26.43 ± 0.25	15.60 ± 0.02	22.60 ± 0.06	13.04 ± 0.07	9.08 ± 0.03
12	37.45 ± 0.73	31.00 ± 0.15	27.34 ± 0.13	15.62 ± 0.01	22.68 ± 0.05	13.10 ± 0.03	9.10 ± 0.02
13	39.19 ± 0.54	31.67 ± 0.13	27.85 ± 0.12	15.62 ± 0.01	22.72 ± 0.00	13.12 ± 0.01	9.13 ± 0.01
14	40.73 ± 0.36	32.04 ± 0.12	28.15 ± 0.15	15.62 ± 0.01	22.72 ± 0.00	13.15 ± 0.01	9.13 ± 0.00
15	41.60 ± 0.25	32.35 ± 0.08	28.33 ± 0.13	15.62 ± 0.00	22.72 ± 0.00	13.15 ± 0.01	9.13 ± 0.00
16	42.39 ± 0.18	32.47 ± 0.10	28.41 ± 0.08	15.62 ± 0.00	22.72 ± 0.00	13.15 ± 0.01	9.13 ± 0.00
17	43.04 ± 0.18	32.57 ± 0.06	28.60 ± 0.07	15.62 ± 0.00	22.72 ± 0.00	13.15 ± 0.01	9.13 ± 0.00
18	43.47 ± 0.07	32.60 ± 0.03	28.66 ± 0.03	15.62 ± 0.00	22.72 ± 0.00	13.15 ± 0.01	9.13 ± 0.00

Table 29: Judge-rated review quality on paired 50-paper LOVO folds. EC: evidence correctness; DD: diagnostic depth; Act.: actionability; PS: paper specificity; Pit.: pitfall severity. Overall is the sum of the five dimensions.

Venue	Condition	EC	DD	Act.	PS	Pit.	Overall
ICLR	Full control	1.55	1.74	1.97	1.99	-0.40	6.85
ICLR	LOVO	1.51	1.83	1.96	1.98	-0.46	6.82
NeurIPS	Full control	1.55	1.59	1.69	2.00	-0.45	6.38
NeurIPS	LOVO	1.36	1.62	1.66	1.96	-0.86	5.74
WWW	Full control	1.46	1.63	1.67	2.00	-0.48	6.28
WWW	LOVO	1.22	1.62	1.69	1.97	-0.95	5.55

Table 30: Source-domain holdout on 200 paired ICLR papers (50 per domain; DeepSeek v4 Flash). The holdout removes target-domain related papers and their supported rubrics. Means are paperwise; Q is the paperwise harmonic mean of Union-W and SFP. ΔQ is holdout minus full control, computed before rounding in percentage points.

Target domain	Related sources	Union-W (%)	SFP (%)	Q (%)	ΔQ (pp)
General ML	Full pool	28.85	89.68	42.29	–
	Domain held out	28.45	89.70	41.60	-0.69
Graphs	Full pool	29.08	87.05	41.93	–
	Domain held out	26.20	86.00	38.45	-3.48
Foundation models/NLP	Full pool	29.25	90.74	42.51	–
	Domain held out	29.20	91.50	42.40	-0.11
Trustworthy AI	Full pool	30.78	85.29	43.90	–
	Domain held out	27.40	86.60	40.20	-3.70
All 200	Full pool	29.49	88.19	42.66	–
	Domain held out	27.81	88.45	40.66	-2.00

Table 31: Judge-rated quality for the paired related-paper source-domain holdout. EC: evidence correctness; DD: diagnostic depth; Act.: actionability; PS: paper specificity (each 0–2); Pit.: pitfall severity (−2–0). Overall is the sum of the five dimensions before rounding.

Target domain	Related sources	EC	DD	Act.	PS	Pit.	Overall
General ML	Full pool	1.52	1.85	1.97	1.98	−0.30	7.02
	Domain held out	1.50	1.83	1.96	1.97	−0.44	6.82
Graphs	Full pool	1.50	1.84	1.96	1.98	−0.24	7.04
	Domain held out	1.38	1.72	1.92	1.94	−0.56	6.40
Foundation models/NLP	Full pool	1.60	1.90	1.98	1.99	−0.41	7.06
	Domain held out	1.58	1.88	1.97	1.99	−0.37	7.05
Trustworthy AI	Full pool	1.58	1.89	1.97	2.00	−0.66	6.78
	Domain held out	1.42	1.76	1.94	1.97	−0.52	6.57
All 200	Full pool	1.55	1.87	1.97	1.99	−0.40	6.98
	Domain held out	1.47	1.80	1.95	1.97	−0.47	6.71

Table 32: Native rating and decision calibration on the same 200 paired LODO papers (25 accepted and 25 rejected per domain). MAE and MSE compare the generated 1–10 rating with the frozen mean human rating; Acc. compares the native generated Accept/Reject decision with the official ICLR decision, without thresholding the rating.

Target domain	Related sources	MAE ↓	MSE ↓	Acc. (%) ↑
General ML	Full pool	0.98	1.42	76.0
	Domain held out	1.02	1.50	74.0
Graphs	Full pool	1.05	1.55	74.0
	Domain held out	1.18	1.90	70.0
Foundation models/NLP	Full pool	1.00	1.48	74.0
	Domain held out	1.04	1.55	74.0
Trustworthy AI	Full pool	1.07	1.60	74.0
	Domain held out	1.24	2.10	68.0
All 200	Full pool	1.02	1.51	74.5
	Domain held out	1.12	1.76	71.5

Table 33: Related-paper year-cutoff sensitivity on a separate 50-paper ICLR cohort (DeepSeek v4 Flash). The 18-rubric library remains fixed; bold marks the best value in each column according to its arrow, including ties. Rating errors use the native 1–10 scale; decision accuracy uses official ICLR labels.

	Review Quality			Native Calibration		
Cutoff Year	Union-W (%)	SFP (%)	Q (%)	MAE ↓	MSE ↓	Decision Acc. (%) ↑
≤ 2022	27.79	86.24	39.73	**1.37**	**2.74**	60.0
≤ 2023	**28.41**	**86.71**	**40.95**	1.40	2.81	**62.0**
≤ 2024	27.38	82.74	39.02	1.43	2.88	**62.0**

Table 34: Judge-rated review quality under related-paper year cutoffs on the separate 50-paper ICLR cohort. EC: evidence correctness; DD: diagnostic depth; Act.: actionability; PS: paper specificity; Pit.: pitfall severity. Overall sums the five dimensions.

Cutoff year	EC	DD	Act.	PS	Pit.	Overall
≤ 2022	1.28	1.80	2.00	1.96	−0.68	6.36
≤ 2023	1.40	1.82	1.98	2.00	−0.68	6.52
≤ 2024	0.78	1.70	1.98	1.98	−1.40	5.04

D.10 Temporal Cutoff Sensitivity

This sensitivity analysis investigates whether DynReviewer's related-paper retrieval path depends on temporally proximate literature. We evaluate a separately frozen set of 50 ICLR target papers, with fixed human references and the same Gemini judge across three year cutoffs. We restrict related-paper retrieval within the historical corpus D to papers published up to 2022, 2023, or 2024, while keeping the full 18-rubric recurring experience library unchanged.

As reported in Table 33, Q is 39.73% with sources through 2022 and 40.95% through 2023, compared with 39.02% under the 2024 full-source control. The ordering is nonmonotonic: allowing more recent sources does not consistently improve the review. SFP ranges from 82.74% to 86.71%, while native decision accuracy is 60.0%, 62.0%, and 62.0% and rating MAE is 1.37, 1.40, and 1.43. These are paired comparisons on the same targets with the backbone and recurring-rubric library held fixed.

E Experience Library and Prompts

E.1 The 18-Rubric Experience Library

Table 35 documents the frozen recurring-experience library used by DynReviewer. The identifiers S01–S18 exactly match the source nodes in Figure 8. Core criteria, applicability conditions, evidence requirements, and source venues are reproduced from the execution library; underscores in names are replaced by spaces for readability. These entries guide target-specific compilation rather than prescribe a negative finding for every paper. Related-paper rubrics are generated separately for each target and are not part of this static inventory. Complete records, including failure patterns and review guidance, are retained in the accompanying artifacts.

Table 35: The 18 recurring rubrics. Each entry lists its core criterion, followed by all applicability conditions, evidence requirements, and source venues in the frozen library.

ID	Field	Library content
S01	**clean comparison**	
S01	Criterion	Assess whether comparative claims are supported by relevant, competitive baselines and assessable comparisons in the claimed setting.
S01	Applicability	The paper claims improvement or novelty over prior work and reports comparative results.
S01		The evaluation spans multiple facets or subcategories where per-facet comparisons matter.
S01		Related work or baselines exist that are standard or state-of-the-art for the task.
S01		The paper claims improvement, effectiveness, or superiority for a new method or component.
S01		The evaluation section omits well-known or standard baselines from the field.
S01		The central performance claim depends on relative comparisons that exclude key alternatives.
S01		The paper frames its contribution as an improvement over existing methods or a cited baseline.
S01		Supporting evidence includes qualitative examples, appendix material, or analyses rather than standardized metrics.
S01		The evaluation setup differs from that of the prior work being compared against.
S01	Evidence	Explicit experimental comparison with strong, recently published baselines, not only ablations or self-comparisons.
S01		Per-facet or per-subcategory results showing the method's standing relative to relevant prior work.
S01		Related work coverage that systematically includes prior methods, including those integrated into experiments.
S01		Evidence that baselines were tuned or configured to a competitive level.

Continued on next page

Table 35 continued: the 18-rubric experience library

ID	Field	Library content
S01		Quantitative comparisons against standard and state-of-the-art baselines under the same setting.
S01		Justification for any omitted baseline that reviewers would expect.
S01		Results in the specific setting claimed (e.g., few-shot), not only easier adjacent settings.
S01		Head-to-head comparison with the cited prior work under a shared, clearly described evaluation setup and metrics.
S01		Quantitative results with sufficient detail (metrics, baselines, significance) to verify the claimed improvement.
S01		Representative examples, not selectively chosen, with enough detail for reviewers to independently assess them.
S01		Discussion of limitations or failure cases relative to the prior work.
S01	Source	ICLR 2023–2024; NeurIPS 2023; WWW 2024; EMNLP 2023; UAI 2024
S02	**clean novelty**	
S02	Criterion	Assess whether the claimed methodological contribution is clearly distinguished from the closest prior methods and whether the stated differences support the claimed added value.
S02	Applicability	The paper applies or adapts an established technique, possibly in a new domain, while claiming a methodological contribution.
S02		The method is described as a modification, calibration, or variation of a known pipeline or framework.
S02		The related work section is brief or omits the closest prior methods applying the same technique.
S02		The paper uses novelty language ('first', 'novel') without substantiation.
S02		The paper claims a novel method whose core mechanism overlaps with an established family of techniques.
S02		The title, abstract, or introduction uses terms like 'novel', 'first', or 'new' to describe the contribution.
S02		The related work section is absent, superficial, or lists prior methods without critical comparison.
S02		Papers claiming a novel method within an area where closely related techniques already exist.
S02		Papers whose contribution involves a known mechanism (e.g., interaction modeling, representation learning) applied or extended to a setting.
S02		Papers whose introduction omits discussion of prior work addressing the same core mechanism or problem.
S02		The paper claims a central method, metric, or idea as its key contribution.
S02		The method heavily relies on or borrows core components from prior published work.
S02		The contribution is framed as an integration or adaptation of known techniques without a new conceptual framework.
S02		Improvements over existing baselines are small or evaluated in limited settings.
S02	Evidence	Explicit identification in related work of prior methods sharing the core mechanism, with a clear statement of technical differences.
S02		A systematic comparison to the closest baseline(s), either algorithmically (component-by-component) or empirically, demonstrating added value over directly applying the prior method.
S02		Demonstration that the proposed modifications alter the fundamental mechanism rather than applying surface-level tweaks.
S02		For domain-transfer claims, evidence that the transfer itself introduces nontrivial new challenges or techniques.
S02		A dedicated passage identifying the specific technical differences between the proposed mechanism and the closest prior methods.
S02		Empirical or analytical comparisons against the most relevant baselines from the same technique family.
S02		A critical analysis of gaps in prior work that the proposed method addresses.

Continued on next page

Table 35 continued: the 18-rubric experience library

ID	Field	Library content
S02		Explicit identification of the closest prior methods and a point-by-point statement of what differs.
S02		Direct empirical comparison against the most relevant baselines, or a clear justification if omitted.
S02		Precise articulation of the specific novel contribution beyond incremental changes to existing frameworks.
S02		A thorough related-work analysis that explicitly identifies prior work using similar ideas, metrics, or methods.
S02		A clear statement of what aspects differ from the closest prior work and why those differences matter.
S02		Comparisons against baselines built from the same borrowed components or the closest prior method.
S02		Demonstration that the combination or extension yields unexpected value or synergy beyond the sum of existing parts.
S02	Source	ICLR 2023–2024; NeurIPS 2022–2024; WWW 2024; EMNLP 2023; UAI 2024
S03	**assumption verifiability**	
S03	Criterion	Check whether the paper's key theoretical assumptions are accompanied by a practical procedure for verifying them in real settings, rather than depending on unobservable quantities such as an optimal or oracle model.
S03	Applicability	Papers whose main theoretical guarantees rely on one or more stated assumptions.
S03		Assumptions formulated in terms of unknown optimal parameters, oracle quantities, or other unobserved objects.
S03		Papers claiming practical relevance or targeting real-world deployment.
S03	Evidence	A concrete diagnostic, test, or empirical procedure that practitioners can run to check whether each key assumption holds.
S03		Discussion of which assumptions are verifiable directly from data and which require proxy or surrogate checks.
S03		Demonstrations (e.g., simulations or case studies) showing the verification procedure in action.
S03	Source	NeurIPS 2022; UAI 2024
S04	**clear problem motivation**	
S04	Criterion	Verify that the paper explicitly states the problem it addresses, the specific technical challenges involved, and a clear motivation for why solving them matters.
S04	Applicability	The abstract or introduction does not explicitly state the problem or its importance.
S04		Technical challenges are implied or mentioned only vaguely rather than clearly identified.
S04		The paper proceeds to methods without framing a gap or need.
S04	Evidence	An explicit problem statement in the abstract or introduction.
S04		A motivation tying the problem to a concrete gap, need, or technical difficulty.
S04		Identification of specific technical challenges the method addresses.
S04	Source	NeurIPS 2023; WWW 2024
S05	**writing clarity and notation**	
S05	Criterion	Check whether the manuscript's prose, definitions, and mathematical notation are clear, consistent, and free of typos and grammatical errors so claims can be followed and verified.
S05	Applicability	Any manuscript presenting technical claims, methods, or mathematical notation
S05		Papers where reviewers must trace symbols, definitions, or logical arguments across sections
S05	Evidence	Precise, unambiguous sentences with clearly defined key terms
S05		Consistent, well-defined notation (e.g., a notation table or definitions at first use)

Continued on next page

Table 35 continued: the 18-rubric experience library

ID	Field	Library content
S05		Text evidently proofread, with minimal typos, grammatical errors, and no incomplete sentences
S05	Source	ICLR 2024; UAI 2024
S06	**related work completeness**	
S06	Criterion	Verify that the paper's related work section adequately covers directly relevant prior tasks and methods that could contextualize or challenge the claimed novelty.
S06	Applicability	Papers claiming novelty or contributions in areas where established or recent prior work exists
S06		Papers whose scope involves specific task families or architectures with adjacent lines of prior research
S06	Evidence	A related work discussion that names and positions the paper relative to the main established tasks, methods, or benchmarks in its claimed area
S06		An explicit justification if well-known adjacent works are deliberately excluded
S06		Reviewer-identified omitted works checked for direct relevance to the paper's scope
S06	Source	NeurIPS 2024; WWW 2024
S07	**design choice justification**	
S07	Criterion	Check whether the paper explicitly justifies its key design and methodological choices (including component ordering, baselines, metrics, hyperparameters, and pipeline structure) with reasoning, theory, or evidence, rather than merely stating what was done.
S07	Applicability	The method has multiple alternative design options but only one is used without discussion of why.
S07		The method involves sequential stages or components whose ordering or structure is non-obvious.
S07		Baselines, metrics, or hyperparameters are presented without motivation or acknowledgment of defaults.
S07	Evidence	Explicit textual rationale linking each major design choice to prior work, theory, or task requirements.
S07		Empirical comparisons or sensitivity analyses showing the chosen configuration (or ordering) outperforms plausible alternatives.
S07		Discussion of trade-offs and alternatives considered for key decisions.
S07	Source	WWW 2024; UAI 2024
S08	**fair efficiency comparison**	
S08	Criterion	Verify that efficiency and performance claims rest on comparisons where the proposed method and baselines compete under equal conditions, with any added resources or computational overhead accounted for.
S08	Applicability	Papers claiming speed or efficiency advantages over baselines
S08		Papers whose method consumes extra supervision, annotation, or data not available to baselines
S08		Papers adding computational components to an existing pipeline
S08	Evidence	A controlled inference-time or cost comparison against tuned baselines in matched settings
S08		Explicit accounting or justification for any extra resources (annotations, parameters, operations) used by the method
S08		Discussion of speed-accuracy or cost-benefit trade-offs supporting the claimed advantage
S08	Source	WWW 2024; EMNLP 2023

Continued on next page

Table 35 continued: the 18-rubric experience library

ID	Field	Library content
S09	**ambiguous problem formulation**	
S09	Criterion	Check that the paper explicitly and unambiguously states the problem setting—naming the learning paradigm, data access model, and primary objective—so readers need not infer them from context.
S09	Applicability	Papers whose method could apply to multiple learning paradigms, data access regimes, or competing objectives, but which never explicitly declare which one is addressed.
S09		Papers where the evaluation setup cannot be mapped back to a clearly stated formulation.
S09	Evidence	An explicit problem statement, ideally in the introduction, naming the paradigm, setting (e.g., data access model, offline/online distinction), and primary objective versus any secondary constraints.
S09		Consistency between the stated formulation and the terminology, method description, and experimental setup.
S09		Definitions of any generic or overloaded terms (e.g., 'unlabeled data', 'policy optimization') needed to disambiguate the setting.
S09	Source	NeurIPS 2023; UAI 2024
S10	**baseline improvement evidence**	
S10	Criterion	Verify that the proposed method demonstrates substantial, competitive improvement over strong state-of-the-art baselines on real-world benchmark data.
S10	Applicability	The paper proposes a new method whose contribution is framed as a performance improvement.
S10		The paper reports quantitative comparisons against multiple baselines on real-world datasets.
S10	Evidence	Comparisons against properly tuned, state-of-the-art baselines under a standard evaluation protocol.
S10		Reported gains that are meaningful in magnitude and statistically supported, not marginal.
S10		Consistent performance across real datasets, not only synthetic or simplified settings.
S10	Source	ICLR 2024; EMNLP 2023; UAI 2024
S11	**claimed but undemonstrated execution**	
S11	Criterion	Verify that every claimed execution, implementation, or analysis is backed by actually reported results (e.g., hardware measurements, deployment details, tables, or metrics) rather than only simulation, theory, or omitted content.
S11	Applicability	Paper claims implementation, deployment, or analysis of a system or component in abstract, introduction, or conclusion
S11		Claims rely on simulation, emulation, or theoretical arguments where physical execution is implied
S11	Evidence	Concrete measurements, specifications, or results directly tied to the claimed implementation or analysis
S11		Details of the platform or setup used for the claimed execution
S11		Reported coverage matching the stated scope of the study
S11	Source	NeurIPS 2022; WWW 2024
S12	**code availability reproducibility**	
S12	Criterion	Verify that the paper provides (or commits to providing) code and sufficient implementation details so that reported experimental results can be independently reproduced.
S12	Applicability	The paper presents an algorithmic or computational method with empirical results or benchmarks.
S12		The paper's claims depend on specific implementations, architectures, training procedures, or custom data preprocessing.
S12		The paper omits any statement about code availability or incomplete implementation details.

Continued on next page

Table 35 continued: the 18-rubric experience library

ID	Field	Library content
S12	Evidence	A clear statement of code availability (repository link, supplementary code, or a commitment to release upon acceptance).
S12		Sufficient implementation details: hyperparameters, random seeds, hardware, and preprocessing steps, whether or not code is released.
S12		For unavailable code, an explanation of why it cannot be shared and enough specification for faithful reimplementation.
S12	Source	NeurIPS 2024; WWW 2024; EMNLP 2023; UAI 2024
S13	**result explanation depth**	
S13	Criterion	Verify that the paper explains why observed performance differences occur, not merely that they occur.
S13	Applicability	Papers reporting comparative performance across multiple models, encoders, or evaluation metrics
S13		Papers whose central claims rest on interpreting performance variation
S13	Evidence	Analysis linking performance differences to underlying causes such as architecture, training data, or task properties
S13		Qualitative examples or error breakdowns illustrating why results vary
S13		Discussion of variance, confounds, or significance behind reported differences
S13	Source	NeurIPS 2022, 2024
S14	**claim generality vs evaluation breadth**	
S14	Criterion	Check that the breadth of empirical evaluation (datasets, tasks, models, configurations) matches the generality of the claims made.
S14	Applicability	The paper makes generalization or capability claims about a method's effectiveness in some domain.
S14		Claims are supported by experiments covering a narrow set of inputs, tasks, models, or conditions.
S14		The review focus is on evaluation breadth rather than the method's internal logic.
S14	Evidence	Testing across multiple datasets, tasks, models, or setups proportional to the claimed generality.
S14		Explicit statement of the claim's scope or boundary conditions.
S14		Adequate variation in baselines, hyperparameters, and random seeds for strong performance claims.
S14	Source	ICLR 2023–2024
S15	**baseline comparison consistency**	
S15	Criterion	Verify that the proposed method is compared against a consistent, appropriate set of baselines (including the original reference model) across all evaluation tables, using identical metrics and protocols.
S15	Applicability	Papers claiming improvement or preservation of qualities (e.g., diversity, novelty, uniqueness) over a reference or baseline model
S15		Evaluations spread across multiple comparison tables or figures
S15		Alignment or adaptation pipelines evaluated against the model they were derived from
S15	Evidence	Quantitative comparisons of each proposed variant against the reference or original model on all claimed metrics
S15		The same baseline set and evaluation protocol used consistently across all comparison tables
S15		Explicit rationale for which baselines are included and excluded
S15	Source	WWW 2024; UAI 2024
S16	**missing variance reporting**	
S16	Criterion	Verify that reported performance claims are supported by variance or uncertainty measures (e.g., error bars, standard deviations, confidence intervals, or significance tests) for both the proposed method and all baselines.
S16	Applicability	Papers reporting comparative performance across multiple runs or random seeds

Continued on next page

Table 35 continued: the 18-rubric experience library

ID	Field	Library content
S16		Papers claiming superiority over baselines where differences are small or marginal
S16		Papers showing result tables or plots based only on mean values
S16	Evidence	Repeated runs or cross-validation with variability reported
S16		Confidence intervals, error bars, or standard deviations for proposed and baseline results
S16		Statistical significance testing when performance gaps are small
S16		A description of the number of runs, seeds, or splits underlying each reported number
S16	Source	ICLR 2024; UAI 2024
S17	**evaluation scale sufficiency**	
S17	Criterion	Verify that the empirical evaluation uses datasets and experiment scales large enough to support the paper's effectiveness and generalizability claims, accounting for variance and false-positive risk.
S17	Applicability	The paper makes empirical effectiveness or generalizability claims based on experiments
S17		The evaluation involves models prone to unstable behavior or false positives at small scale
S17		The paper reports experiments on small or few datasets
S17	Evidence	Dataset sizes and experimental scale should be reported with justification that they suffice for the claims made
S17		Statistical significance testing, confidence intervals, or multiple runs/cross-validation to account for variance
S17		Results on larger or more diverse benchmarks, or ablations showing robustness beyond the small-scale setting
S17	Source	ICLR 2024; EMNLP 2023
S18	**incremental theoretical novelty**	
S18	Criterion	Verify that the paper's theoretical results (theorems, propositions, proofs) represent a substantive advance beyond existing work rather than a minor extension or direct application of prior results.
S18	Applicability	The paper claims novel theorems, propositions, or theoretical frameworks
S18		Theoretical results build on or cite prior theoretical frameworks
S18		The abstract or introduction emphasizes theoretical novelty
S18	Evidence	A clear statement of how each main theoretical result differs from the closest prior work
S18		Proofs that introduce new techniques or assumptions, not just borrowed ones
S18		Demonstration of new insights, broader applicability, or significant improvements over existing results
S18	Source	ICLR 2024; UAI 2024

E.2 Main-Method Prompt Templates

The prompts below construct review experience and generate the final review. The recurring-rubric and related-paper paths share target-specific compilation, inspect the target paper separately, and aggregate their findings. The final subsection specifies one-call zero-shot and raw-related few-shot review generation. The evaluation and Top-K selection prompts appear in Appendix C.4.

E.3 Rubric-Experience Construction

Historical weakness extraction

BASE INSTRUCTION

Split one historical human-review weakness into atomic concerns. Return JSON only with the closed shape {"atoms":[{"review_span":"...","normalized_concern":"..."}]}. Return at most three atoms. Each atom must express exactly one concern. Each review_span must be one contiguous exact substring of weakness_text; do not paraphrase the span. normalized_concern must preserve the weakness direction and must not turn the concern into positive compliance. Do not invent identifiers, source facts, paper evidence, outcomes, ratings, or decisions.

INPUT FIELDS (DYNAMIC VALUES)

source_paper_id: <SOURCE_PAPER_ID>; weakness_id: <WEAKNESS_ID>; weakness_text: <WEAKNESS_TEXT>

Review-mechanism signature

BASE INSTRUCTION

Compile one historical weakness atom into a causal review-mechanism signature. Return JSON only and exactly these fields: observed_failure, evidence_obligation, affected_claim_type, affected_claim_description, causal_rationale, applicability_cues, known_risk_patterns, risk_category. The four fields observed_failure, evidence_obligation, affected_claim_type, and causal_rationale must each be an object with exactly two keys, text and support, where support is review_explicit, review_entailed, or model_inferred. Use review_explicit only when the field text is one contiguous verbatim substring of review_span. Use review_entailed for a conservative paraphrase whose full meaning follows from review_span without adding a new premise. Never label a paraphrase review_explicit. Use model_inferred for any plausible bridge, mechanism, consequence, or evidence demand that the review does not state or entail. When uncertain, use model_inferred. Never label an invented claim review_explicit or review_entailed. A review_entailed field may not introduce new entities, datasets, baselines, metrics, or causal consequences absent from review_span. In particular, affected_claim_type is normally model_inferred unless the exact category phrase appears in review_span. If any word in a proposed paraphrase is not supported by review_span, use model_inferred rather than review_entailed. affected_claim_description is a plain non-empty string only; it must not be an object and must not contain a support field. applicability_cues and known_risk_patterns are each a list of one to four strings. Preserve the weakness direction; do not rewrite a criticism as a positive compliance statement. These lists are non-exhaustive investigation hints. They must not determine the final verdict, are not closed failure conditions, and may not prevent discovery of a new failure mechanism. Use generic entities rather than copying paper-specific names. Do not emit IDs, provenance, recurrence eligibility, or a final pass/fail verdict.

INPUT FIELDS (DYNAMIC VALUES)

source_atom_id: <SOURCE_ATOM_ID>; source_paper_id: <SOURCE_PAPER_ID>; review_span: <REVIEW_SPAN>; normalized_concern: <NORMALIZED_CONCERN>; polarity: <POLARITY>

Recurring rubric synthesis

BASE INSTRUCTION

Synthesize one generic peer-review rubric from the following mechanism signatures. These signatures come from at least two independent papers. Return JSON only with exactly these fields: name, criterion, applicability, evidence_requirements, failure_patterns, review_guidance. name must be a short lowercase snake_case label. criterion must be one actionable review check. applicability, evidence_requirements, failure_patterns, and review_guidance must each be a list of 1-4 non-empty strings. Do not return more than four items in any list. Abstract away paper names, datasets, authors, and equations; preserve the causal mechanism and weakness direction. Do not invent a publication decision or score.

INPUT FIELDS (DYNAMIC VALUES)

cluster_id: <CLUSTER_ID>; signatures: <SIGNATURES>

E.4 Review Generation

Target-paper profile

BASE INSTRUCTION

Build a whole-paper audit profile from the complete Markdown paper. Do not infer facts not present in the paper. Cover methods, evidence, experiments, baselines, ablations, limitations, and risks; do not reduce the profile to a list of claims. Return JSON only with the PaperProfile fields. Every section heading must be copied exactly from the supplied Markdown. Do not use reviews, decisions, or labels. Keep list items concise (one grounded sentence each), without including a Markdown heading in the item text. Always include a sections array with one entry per major Markdown heading (Markdown levels 1-2 only; do not enumerate every checklist or appendix subheading), and include contribution_types even when it is an empty array. Use this shape: {paper_id,title,primary_area, research_question,contribution_types,method_components,conclusion_evidence_links, experiment_objects,baseline_and_ablation_objects,generalization_and_limitations,risk_exposures, sections:[{heading,level,summary,evidence_focus}]}.

INPUT FIELDS (DYNAMIC VALUES)

paper_id: <PAPER_ID>; title: <TITLE>; abstract: <ABSTRACT>; primary_area: <PRIMARY_AREA>; complete_body_markdown: <COMPLETE_BODY_MARKDOWN>

Rubric search

BASE INSTRUCTION

You are a restricted rubric-search agent, not a conversational assistant and not a paper reviewer. You have exactly four tools: glob, grep, read, and submit_selection. The only paper information available to you is the PROFILE below; do not ask for the original paper and do not invent another target.

RUNTIME REMINDER: this is a profile-only atomic rubric search. Your first search tool call must be read with path "index.md"; then call glob("*") once to confirm the library root. Only the recurring branch is available. Read recurring/index.md before choosing categories. Use the PROFILE to decide which category descriptions match this paper's methods, evidence, experiments, baselines, generalization, or limitations; do not inspect every category. For recurring, call glob("recurring/<category>/*") to reveal atomic JSON. Select complementary applicable rubrics from the recurring branch only. Indexes and generation_manifest.json are metadata, not selectable rubrics. Build a complementary non-redundant set grounded in exact profile phrases. Copy rubric_id and path exactly from read atomic files. submit_selection is terminal. Do not use a fixed rubric count; stop when profile-relevant coverage is sufficient. Search latency guard: at most <SEARCH_ACTION_BUDGET> exploratory glob/grep calls are available. This is not a limit on the number of selected rubrics. Once the guard is reached, stop exploring, read already discovered atomic files, and call submit_selection.

Rubric search (continued)

PATH DISCIPLINE: A path is valid only if it was returned by a successful glob or grep during this run and then successfully read. Never invent or import paths. A selectable file is an atomic JSON object containing rubric_id and criterion. PROFILE ANCHORS: <PROFILE_ANCHORS> SEARCH OBJECTIVE: derive the paper's evidence needs from the PROFILE, use the branch indexes to choose relevant categories, read the corresponding atomic files, and select a complementary non-redundant set. Do not impose a fixed category checklist or fixed count. Use exact profile phrases for grep and for every selection reason. Reject an atomic rubric when its original audit object is absent from the PROFILE; mentioning profile words in the reason does not make an unrelated rubric applicable. Before submission, remove rubrics whose criterion duplicates another selected rubric.

Only after the search is complete, call submit_selection with the choices. submit_selection is the only final output; after it succeeds, stop immediately and emit no prose, Markdown, or questions. The submit_selection tool accepts only ids and paths returned by read. If no applicable atomic rubric was read, call submit_selection with {"choices":[]}. PROFILE: <PAPER_PROFILE>

USER INSTRUCTION Execute the hierarchical search now: read index.md, glob the root once, read the available branch indexes, choose profile-relevant categories, read atomic rubrics, and finish with submit_selection. Never answer in prose.

INPUT FIELDS (DYNAMIC VALUES)

paper_profile: <PAPER_PROFILE>; allowed_branches: <ALLOWED_BRANCHES>; search_action_budget: <SEARCH_ACTION_BUDGET>

Related-paper review distillation

BASE INSTRUCTION

From this one related paper and its corresponding human reviews, distill reusable review rubrics for the target profile. Do not copy paper-specific names or numbers. Return JSON {rubrics:[{criterion, applicability,evidence_requirement,reason_for_target}]} and keep every rubric grounded in the supplied review pair.

INPUT FIELDS (DYNAMIC VALUES)

target_profile: <TARGET_PROFILE>; related_paper: <RELATED_PAPER>

Target-specific rubric compilation

BASE INSTRUCTION

Construct a set of target-specific review criteria for the supplied paper. Use the whole paper profile and the supplied source material. Source material consists of either a frozen general rubric file or one related-paper/human-review distilled rubric. Return JSON with exactly one key, ‘rubrics‘, whose value is an array. Each item must contain criterion, applicability, evidence_requirement, and source_refs (an array of source labels). Preserve every distinct, materially applicable failure mode found in the sources. Do not merge criteria merely because they share a broad category such as evaluation, novelty, or reproducibility. Merge only when resolving one criterion with the same target-paper evidence would necessarily resolve the other. A criterion may combine its diagnosis, consequence, and direct remedy, but must not combine independently testable concerns. Do not impose a small fixed count or keep only the most severe criteria.

Each output item must be a target-specific rubric, not a generic checklist heading. ‘criterion‘ must state the precise risk or unresolved question for this paper. ‘applicability‘ must cite the target claim, method component, dataset, baseline family, theorem, table, figure, or section that makes the criterion relevant. ‘evidence_requirement‘ must state what concrete evidence would resolve the risk and what is presently missing or ambiguous in the supplied paper. The existence of a citation, experiment, baseline, ablation, theorem, or limitation paragraph does not by itself resolve the corresponding failure mode. Use ‘source_refs‘ to retain every source supporting the criterion. Omit a source concern only when it is demonstrably inapplicable or already fully resolved by specific target-paper evidence. Do not copy

Target-specific rubric compilation (continued)

source-paper-specific names, numbers, or decisions, and do not invent evidence or publication decisions.

INPUT FIELDS (DYNAMIC VALUES)

target: <TARGET>; profile: <PROFILE>; sources: <SOURCES>

Branch evidence inspection

BASE INSTRUCTION

Inspect this paper independently using ONLY the supplied rubric branch. The paper is the sole factual authority; the profile and rubrics are inspection requests, not evidence of defects. Return JSON matching the supplied schema, no Markdown fences. For EVERY rubric return one short decision: retained (a supported concern or unresolved material question), resolved by actual paper evidence, or inapplicable. Give a brief conclusion and a short verbatim paper quote; do not output hidden reasoning or a proof of your decision. When expected evidence is absent, quote the closest actual claim or experimental setup and state what is missing in the finding, never invent a quote. For every retained rubric include at least one finding linked by rubric_id. Preserve independently testable failure mechanisms. Related topics alone are not duplicates. Inspect central claims independently as well; additional material findings may have an empty rubric_ids array. Do not fabricate problems to fill a quota or force a finding for a resolved rubric. Do not impose an arbitrary number of findings. A finding must state a precise issue/question, a short verbatim evidence quote and locator, its implication for a claim, and an actionable request. Test the benign interpretation; ambiguity should be a calibrated question, not an asserted contradiction. Do not treat acknowledged limitations as automatically harmless, or demand unrelated experiments. No human-review gold is provided or should be assumed. Use IDs F1,F2,... within this response. Copy concise evidence excerpts faithfully from the paper and preserve their meaning. Literal occurrence is recorded as a diagnostic; it is not a test of whether the evidence supports the concern.

INPUT FIELDS (DYNAMIC VALUES)

target: <TARGET>; profile: <PROFILE>; rubrics: <RUBRICS>; output_schema: <OUTPUT_SCHEMA>

Final aggregation

BASE INSTRUCTION

Assemble a peer review from independently verified branch findings. Return JSON only matching the supplied schema. Write a brief factual summary and paper-supported strengths. Partition ALL finding IDs into groups exactly once. Group findings only when they have the same failure mechanism and evidence obligation. Semantic-similarity candidate pairs are hints, NOT confirmed duplicates. Keep distinct mechanisms separate even if they concern the same topic. Prefer the specific technical issue over a broad category as group title. Every finding must survive; you cannot remove, replace, weaken, or introduce a finding. The final concern text will be rendered verbatim from the input fields by code, so only specify grouping and concise headings. Do not refer to memory sources or this assembly process in summary/strengths.

ICLR RECOMMENDATION CONDITION

When the target venue is ICLR, also give an overall recommendation based on the paper and findings. Use rating 1, 3, 5, 6, 8, or 10: 1 strong reject; 3 reject; 5 marginally below acceptance; 6 marginally above acceptance; 8 accept; 10 strong accept. Give decision Accept or Reject. This is the reviewer's recommendation, not a prediction of the conference outcome or other reviewers' scores. Judge contributions and evidence rather than counting concerns.

INPUT FIELDS (DYNAMIC VALUES)

target: <TARGET>; findings: <FINDINGS>; possible_duplicate_pairs: <POSSIBLE_DUPLICATE_PAIRS>; output_schema: <OUTPUT_SCHEMA>

E.5 One-step Zero-shot and Few-shot Controls

This template defines one generation call per target. Zero-shot supplies the target paper and venue; raw-related few-shot adds up to three retrieved papers and their original human reviews. Both arms use the same prompt builder. The venue condition appends Rating and Decision sections only for ICLR; other venues receive neither section nor a numeric recommendation.

Zero-shot and raw-related few-shot review

SINGLE SYSTEM PROMPT

Write a substantive Markdown peer review of the supplied paper. Use exactly these level-2 sections in order: Summary, Strengths, Specific and Decisive Concerns, Questions, and Limitations. If the target venue is ICLR, add Rating and Decision. Review the complete target paper. Ground each substantive statement in a paper heading, table/figure/equation/algorithm label, or short verbatim span. State uncertainty where the paper leaves evidence incomplete. Examine claims, assumptions, method components, datasets, baselines, ablations, compute, and limitations. Prioritize material concerns and give an actionable request for each. When related papers and original human reviews are supplied, treat them as background data that suggests candidate questions. Determine each finding from evidence in the target paper. If the target venue is ICLR, end with '## Rating' containing exactly one rating in the form '<value>/10', where <value> is one of 1, 3, 5, 6, 8, 10. Use the ICLR experimental anchors: 1 strong reject; 3 reject; 5 marginally below acceptance; 6 marginally above acceptance; 8 accept, good paper; 10 strong accept. Then write '## Decision' containing exactly 'Accept' or 'Reject'. Ratings 1, 3, 5 require Reject; ratings 6, 8, 10 require Accept. Give your own paper-grounded recommendation based on contributions and evidence. For other venues, do not provide a numeric rating or publication decision.

INPUT FIELDS (DYNAMIC VALUES)

target: <TARGET>; venue: <VENUE>; related_paper_review_pairs: <RELATED_PAPER_REVIEW_PAIRS>

F Cases and Provenance

We present four domain cases with explicit rubric attribution and a paired cross-system review comparison, followed by the aggregate provenance diagram. Appendix F.5 reproduces a separate complete output example.

F.1 Cases Across Four Domains

The four cases below are drawn from the same DeepSeek ICLR200 run. The Graph case is the original example; the other three are selected closest to their domain median atomic-finding count among papers linked to both paths (paper-ID tie-break; no quality-score selection). Green denotes related-paper provenance and blue recurring-rubric provenance. Bracketed source annotations are added for presentation, while weakness excerpts and linked criteria retain their saved wording.

Reading the Annotations S01–S18 are global IDs in the static library. In contrast, `history:R4` means target-specific check 4 in the recurring-rubric branch for this paper; it is not static rubric 4. A source such as S07 is compiled into that local check, which is then linked to the quoted finding. `related:R3` similarly denotes local check 3 from related-paper retrieval. Local IDs reset for each paper; a local check may have multiple static sources. These mappings record attribution, not causal benefit.

F.1.1 Graph: Port-Hamiltonian Architectural Bias for Long-Range Propagation in Deep Graph Networks

Both excerpts concern the scope of the theoretical guarantees. The related-paper check focuses on gradient bounds for the full model; the recurring check asks whether the claimed theoretical extension is supported. Their overlap records how two inspection paths address the same topic, without establishing that either criticism is correct.

Related-paper path

The full port-Hamiltonian PH-DGN with driving forces lacks theoretical guarantees for non-vanishing gradients and stable training; the lower bound in Theorem B.1 depends on step size and driving force bounds, and the upper bound in Theorem A.1 suggests potential gradient explosion, which is not systematically controlled or analyzed. **[Rubrics: paper-related; local check: related:R1]**

Linked criterion (related:R1): The theoretical guarantees on non-vanishing gradients and long-range propagation are proven only for the purely conservative PH-DGN (Theorems 2.1-2.4, A.2), while the full port-Hamiltonian version with driving forces lacks such guarantees; the lower bound in Theorem B.1 depends on step size and driving force bounds, and the upper bound in Theorem A.1 suggests potential gradient explosion, which is not empirically controlled or analyzed.

Recurring rubric: S18

The theoretical guarantees for the full port-Hamiltonian version are weaker and depend on additional assumptions, but the paper's abstract claims general guarantees. **[Rubrics: S18; local check: history:R13]**

Linked criterion (history:R13): Does the paper provide sufficient evidence to support the claim that the theoretical guarantees on long-range propagation hold for the full port-Hamiltonian version (with driving forces), given that the main theorems (2.1-2.4) are proven only for the conservative case?

F.1.2 FM / NLP: COLLAB: CONTROLLED DECODING USING MIXTURE OF AGENTS FOR LLM ALIGNMENT

Both paths discuss the greedy switching approximation: one excerpt asks about the suboptimality gap, while the other also requests comparison with alternative switching strategies.

Related-paper path

The greedy switching policy is acknowledged as a crude proxy for the infeasible dynamic programming solution, but the suboptimality gap introduced by this approximation is not quantified. **[Rubrics: paper-related; local check: related:R3]**

Linked criterion (related:R3): The greedy switching policy is a crude approximation to the optimal dynamic programming solution, and its suboptimality is not quantified or empirically justified.

Recurring rubric: S07

The greedy switching policy is acknowledged as a 'crude proxy' for the optimal dynamic programming solution, but the paper does not analyze the approximation error or compare to alternative switching strategies. **[Rubrics: S07; local check: history:R4]**

Linked criterion (history:R4): Is the greedy switching policy (Algorithm 1) explicitly justified as a design choice, given that it is acknowledged as a 'crude proxy' for the optimal dynamic programming solution?

F.1.3 GENERAL ML: RELCHANET: NEURAL NETWORK FEATURE SELECTION USING RELATIVE CHANGE SCORES

Both paths question whether the reported efficiency comparison separates the method from architecture and training-budget differences.

Related-paper path

The computational efficiency comparison uses a different architecture for NeuroFS (3-layer, 1000 neurons) than for RelChaNet (1-layer, 100 neurons), and does not report epoch counts or per-epoch costs, so the efficiency claim may be misleading. **[Rubrics: paper-related; local check: related:R3]**

Linked criterion (related:R3): The claimed computational efficiency of RelChaNet is not fully transparent because the paper reports wall-clock time but does not account for the overhead of gradient accumulation over mini-batches, the cost of reinitializing candidate weights, or the potential need for more epochs due to random regrowth, and the comparison with NeuroFS uses a different architecture (3-layer vs. 1-layer).

Recurring rubric: S08

The efficiency comparison uses different architectures for NeuroFS (3-layer, 1000 neurons) and LassoNet (1 hidden layer, 100 neurons) without reporting the number of epochs or total FLOPs, so the wall-clock times do not isolate the method's efficiency from architectural differences. **[Rubrics: S08; local check: history:R2]**

Linked criterion (history:R2): Are the efficiency claims (RelChaNet faster than NeuroFS and LassoNet) based on matched conditions that account for the dense network overhead and the extra training epochs required by random regrowth?

F.1.4 TRUSTWORTHY AI: ACCURATE SPLIT LEARNING ON NOISY SIGNALS

The related-paper excerpt concerns privacy accounting over training; the recurring-rubric excerpt concerns the justification of scaling, masking, and layer-placement choices. These illustrate distinct inspection topics within the same final review.

Related-paper path

The paper provides DP bounds only for a single forward pass (Table 9) and acknowledges that direct composition yields large privacy bounds, leaving privacy accounting for the entire training process as an open problem. This limits the practical privacy guarantee of the proposed method. **[Rubrics: paper-related; local check: related:R2]**

Linked criterion (related:R2): The privacy guarantee is only established for a single forward pass, and the total privacy budget for the entire training process is acknowledged to be large and impractical; the paper does not provide a rigorous privacy accounting method for noisy SplitNN training.

Recurring rubric: S07

The paper does not provide systematic justification for the chosen scaling/masking ratios and denoising layer placement, and lacks sensitivity analysis. **[Rubrics: S07; local check: history:R5]**

Linked criterion (history:R5): Are the key design choices (e.g., scaling factor, masking ratio, denoising layer placement) explicitly justified with reasoning, theory, or evidence, rather than merely stated?

F.2 A PAIRED REVIEW COMPARISON: IRIS

We compare the three DeepSeek-based reviews of *IRIS: LLM-Assisted Static Analysis for Detecting Security Vulnerabilities* (ICLR 2025; paper ID `9LdJDU7E91`) against the same 16 human concerns using the joint evaluator described in Appendix B. The weighted Union coverage, Union_W, gives a full match weight 1 and a partial match weight 0.5, normalized by the number of human concerns. All three systems receive the same paper and reference inventory.

Table 36: Paired case-study results for IRIS using the same 16-item human concern reference for weighted coverage and supported-finding precision (SFP). This selected case does not estimate average performance.

System	Union_W (↑)	SFP (↑)
DynReviewer (DeepSeek)	**46.9%**	95.7%
ReviewGrounder (DeepSeek)	9.4%	**98.5%**
AI Scientist-v2 (DeepSeek)	15.6%	85.0%

DynReviewer matches nine of the 16 human concerns (weighted mass 7.5/16), compared with two for ReviewGrounder (1.5/16) and three for AI Scientist-v2 (2.5/16). The excerpts show three matched concerns; yellow marks a match, while gray marks no match under the shared evaluator. ReviewGrounder has the highest SFP in this case.

Human concern: causes of missed detections.

Matched **DynReviewer:** *"The paper lists causes of undetected vulnerabilities but does not quantify how often each cause occurs, making it unclear which limitation dominates."*

ReviewGrounder: No finding matched to this concern. AI Scientist-v2: No finding matched to this concern.

Human concern: reliability of the 50-alarm FDR audit.

Matched **DynReviewer:** *"The manual analysis of 50 alarms is a small sample with no detailed methodology or statistical confidence intervals."*

ReviewGrounder: No finding matched to this concern.

Matched **AI Scientist-v2:** *"The manual FDR estimate of 46% is based on only 50 random alarms (Section 5.2); no details on the true/false positive breakdown or criteria for 'potential attack surfaces' are provided, making this estimate potentially unreliable."*

Human concern: heuristic baselines for the specification ablation.

Matched **DynReviewer:** *"The ablation studies only replace GPT-4 specifications with CodeQL's, not with random or heuristic baselines, so they do not isolate the semantic contribution of the LLM."*

ReviewGrounder: No finding matched to this concern. AI Scientist-v2: No finding matched to this concern.

A Paper-Grounded Concern Beyond the Human Reference DynReviewer also questions the verification of IRIS's four previously unknown vulnerabilities:

> The paper verifies that CodeQL alone did not detect the 4 previously unknown vulnerabilities but does not verify other baselines or provide full details of the validation process.

The human reference contains no corresponding concern, and the joint evaluator assigns this finding full paper support (2/2). Section 5.3 states, "we verified that CodeQL alone did not detect them," but reports neither the other tools' results on these four cases nor external confirmation. The excerpt therefore identifies a gap in the evidence reported for the discovery claim; it does not establish that the vulnerabilities are false.

F.3 Full-Cohort Provenance

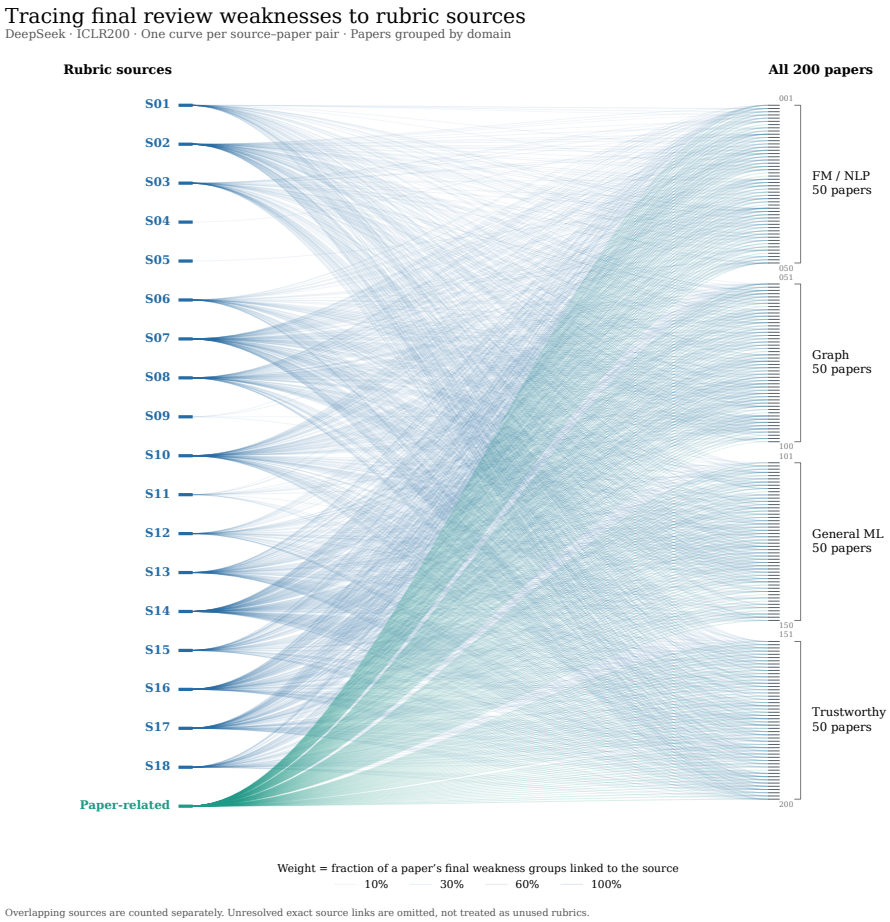


Figure 8: Recorded rubric provenance across all 200 ICLR papers for DynReviewer (DeepSeek). Left: 18 recurring rubrics (S01–S18, blue) and the related-paper path (green). Right: papers grouped into four domains (50 each). Curve weight is the fraction of a paper's weakness groups linked to the source. Multiple sources may link to one group; unresolved exact links are omitted. This diagram records attribution, not review quality or an ablation effect.

F.4 Provenance Counting Units

We inspect the saved finding identifiers used to assemble the DeepSeek ICLR200 final reviews. At the paper level, 199 reviews contain both history-prefixed and related-prefixed findings; one contains history-prefixed findings only. Of 2,167 final weakness groups, 1,040 combine the two prefixes, 881 contain related findings only, and 246 contain history findings only. A group can contain multiple atomic findings. Thus, a group containing both paths does not imply that an individual atomic finding has two independently verified source chains. These counts measure recorded branch provenance, not semantic correctness, unique-source attribution, or causal benefit.

F.5 Complete Review Output Example

The following is a saved DynReviewer review for *Transformers Are Universal In-Context Learners* (Qwen 3.7 Plus; paper ID `6S4WQD1LZR`). Its three weakness sections address theorem scope, quantitative bounds, and learnability. The Accept recommendation matches the official decision.

Complete Review: Transformers Are Universal In-Context Learners

Summary

The paper establishes universal approximation theorems for deep transformers operating on an arbitrary (even infinite) number of tokens by modeling contexts as probability distributions and using the Wasserstein distance to define continuity. It proves that transformers can approximate continuous in-context mappings uniformly over compact token domains for both unmasked and masked (causal) settings, with fixed embedding dimension and number of heads independent of precision. The masked setting requires additional regularity (Lipschitz contexts) and causal identifiability. The results are non-quantitative and rely on specific architectural choices (no normalization, scalar-output heads).

Strengths

• Rigorous measure-theoretic formalization of transformer expressivity that naturally handles arbitrary (including infinite) numbers of tokens via probability distributions and Wasserstein continuity.

• Universal approximation theorems (Theorems 1 and 2) for both unmasked and masked settings with fixed embedding dimension and number of heads independent of approximation precision.

• Novel proof technique leveraging the Stone-Weierstrass theorem on an algebra of cylindrical functions built from elementary attention mappings, overcoming the lack of algebraic structure in shallow attention architectures by using depth.

• Clear identification of the causal identifiability and Lipschitz context conditions required for the masked setting, with proofs that these conditions are preserved under composition of transformer layers.

Weaknesses

1. Restrictive assumptions and limited practical applicability

The Dirac mass constraint at time zero ($\bar{\mu}(\{0\}) \geq \sigma$) excludes measures with Lebesgue density, and the suggested relaxation to a fixed time marginal is stated without proof or detailed justification. As stated in Section 4.1, Remark 2, "We impose that contexts have Dirac masses at 0, namely $\bar{\mu}(\{0\}) \geq \sigma$. While this is not restrictive for discrete measures, this prevents for instance measures with density with respect to Lebesgues. It is possible to lift this constraint, and instead impose that the time marginal $\bar{\mu}$ is fixed, i.e. replace the set Lip_C^$\sigma(\tilde{\Omega})$ by $\{\mu \in$ Lip_C$(\tilde{\Omega})$: $\bar{\mu} = \nu\}$ for any $\nu \in$ P([0,1]) satisfying $0 \in$ supp(ν). One can check that the proof of Theorem 2 carries over to this setting with minor modifications." This constraint significantly restricts the scope of practical applicability, as many real-world sequential data distributions may have continuous time marginals. The claim that the proof carries over with 'minor modifications' is not substantiated. Provide a detailed proof or at least a sketch showing how the compactness argument in Lemma 5 and the Stone-Weierstrass application in Proposition 2 extend to the case of a fixed time marginal ν with $0 \in$ supp(ν). This concern has medium severity and high confidence.

The exclusion of RoPE and other modern positional encodings limits practical relevance, and the claim that extension requires only 'slight adjustments' is not substantiated. As stated in Section 2.3, "The space-time lifting can incorporate positional information, however, recent positional encoding methods, such as Rotary Positional Embedding (RoPE) (Kazemnejad et al., 2024), are not included in the current formulation. This extension is left for future work." RoPE is widely used in modern transformers and modifies all attention layers to account for relative positional information. The claim in the Conclusion that this would require 'slight adjustments to the proof' may underestimate the complexity of the extension. Provide a more detailed analysis of what changes would be required to extend the framework to RoPE, including whether the measure-theoretic formulation and the algebra of elementary mappings would need to be fundamentally modified. This concern has low severity and medium confidence.

The framework's reliance on specific architectural choices (softmax attention, no normalization, specific MLP structure) limits generality, and potential extensions are not adequately discussed. As stated in Section 2, "For the sake of simplicity, we omit normalization in the following analysis." Real-world transformers typically include normalization layers (LayerNorm, BatchNorm) and may use different attention mechanisms (linear attention, sparse attention) or activation functions. The omission of these components limits the practical applicability of the results. Discuss whether the measure-theoretic framework and the universality results can be extended to transformers with normalization layers, and identify any technical obstacles that would need to be overcome. This concern has low severity and medium confidence.

Complete Review: Transformers Are Universal In-Context Learners (continued)

Lack of empirical diagnostics for theoretical assumptions in the masked setting. As stated in Section 4.1, "It is worth noting that these conditions are automatically satisfied by a discrete measure, $\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{(x_i,t_i)}$, with distinct times $\delta := \min_{i\neq j} |t_i - t_j| > 0$ (using $C = \text{Radius}(\Omega)/\delta$)." Practitioners cannot easily verify if their real-world sequential data satisfies the Wasserstein-Lipschitz and causal identifiability conditions required for the universality theorem to hold, limiting the practical applicability of the masked setting results. Provide practical heuristics, empirical tests, or surrogate metrics to estimate whether a given real-world dataset or context satisfies the Lipschitz and identifiability assumptions. This concern has medium severity and high confidence.

2. Non-quantitative bounds and architectural constraints

The non-quantitative nature of the approximation results means there is no explicit control over how the number of MLP parameters or depth depends on precision ε, limiting practical utility. As stated in Section 3.1, "A weakness is that the theorem is 'non-quantitative', meaning that that we have no explicit control over the dependency of the number of MLP parameters ξ_ℓ on ε." Without quantitative bounds, practitioners cannot determine the architectural requirements (depth, width, parameter count) needed to achieve a desired approximation precision, making the universality results difficult to apply in practice. Provide at least rough asymptotic bounds on how depth L and MLP parameter count scale with ε, or clearly delineate the specific technical obstacles that prevent such bounds and propose concrete approaches to overcome them. This concern has medium severity and high confidence.

The construction requires H = d' heads each outputting only a scalar (d_head = 1), resulting in an unbalanced architecture that may not reflect practical transformer designs. As stated in Section 3.1, "A limitation of our proof technique is that the number of heads grows proportionally to the output dimension while each head only outputs a scalar d_head(θ_ℓ) = 1. Obtaining a better balance between these two parameters is an interesting problem." This architectural constraint may limit the practical relevance of the results, as real-world transformers typically use fewer heads with larger head dimensions. It is unclear whether this is a fundamental limitation or an artifact of the proof technique. Discuss whether alternative proof techniques (e.g., approximating vector-valued functions directly rather than dimension-by-dimension) could yield constructions with fewer heads and larger head dimensions, and provide evidence or arguments for why this is or is not possible. This concern has low severity and medium confidence.

3. Gap between universality and learnability

The gap between existence of universal approximators and actual learnability through gradient-based training is acknowledged but not adequately addressed. As stated in Conclusion and Discussion, "It is important to note that universality results like ours do not directly translate into conclusions about the learning capabilities of transformers. However, our proof techniques, which leverage ideas from optimal transport, share similarities with those used in the analysis of two-layer MLPs (Chizat & Bach, 2018). Thus, we believe that future work could build on this approach to investigate both convergence properties of the training of transformers." Without results on learnability, the practical significance of the universality theorems is limited, as practitioners need to know whether transformers can be trained to achieve the approximations, not just that such approximations exist. Provide a more detailed discussion of the specific challenges in extending the optimal transport-based analysis of two-layer MLPs to deep transformers, and outline concrete research directions for establishing training convergence results. This concern has medium severity and high confidence.

The paper is purely theoretical with no numerical experiments to illustrate or support the universality claims on concrete tasks. As stated in Section 3.1 and Appendix D, "Appendix D demonstrates that this result can be applied to two key settings: permutation-invariant mappings over a fixed number of tokens and in-context learning of regression operators." The absence of empirical validation makes it difficult to assess the practical relevance of the theoretical results and whether the constructed transformers can actually approximate target mappings in practice, especially given the non-quantitative nature of the bounds. Include numerical experiments demonstrating the approximation of at least one non-trivial in-context mapping (e.g., a simple regression or classification task) with transformers of varying depths and widths, showing how approximation error decreases with architectural capacity. This concern has medium severity and high confidence.

Questions

No additional questions.

Rating

6/10.

Decision

Accept

G Artifacts and Reproducibility

G.1 Resource Availability and Access Conditions

Local experiment artifacts include sample manifests, reference inventories, generated reviews, and evaluation outputs. The rubrics and core prompts are documented in this appendix. A public archive is not yet available; redistribution of full paper and review texts depends on the source permissions.

Artifact	Current status	Access and redistribution conditions
ReviewAtlas-40K+ corpus	Used internally for resource construction and experience extraction.	Paper/review redistribution depends on source terms and rights; no corpus-wide license is asserted.
Frozen benchmark IDs	Stored with sample manifests and group assignments.	To be released as an anonymized identifier manifest, subject to source conditions.
Reference concern inventories	Versioned evaluation artifacts.	To be released subject to source-text permissions.
18 rubrics and core prompts	Documented in this appendix.	Available in this manuscript.
Evaluation and aggregation code	Scripts with result provenance.	To be released in a public repository.