AUTOSUPERVISION: CLOSING THE FEEDBACK LOOP IN SCIENTIFIC WORKFLOWS WITH GROUNDED REVISION VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in large language models (LLMs) have enabled AI systems to assist scientific research and peer review. However, an essential capability for reliable AI-assisted scientific workflows remains underexplored: verifying whether reviewer feedback leads to meaningful and evidence-supported manuscript improvements. We introduce AUTOSUPERVISION, which evaluates whether scientific manuscript revisions genuinely address reviewer concerns through grounded evidence. AUTOSUPERVISION leverages transparent peer-review records as a natural source of supervision, where reviewer comments specify scientific concerns, author responses describe claimed resolutions, and revised manuscripts provide evidence of changes. Given reviewer comments, author responses, and revised manuscripts, models must characterize reviewer concerns, determine whether concerns have been addressed, and identify supporting manuscript evidence. We construct 8,790 episode-level concerns from *Nature Communications* articles and corresponding review records. Then we conducted experiments on LLMs, an ablation study, and a case study. Our results show that while LLMs perform well in characterizing reviewer concerns, with GPT-5.5 achieving a score of 0.754, evidence-based verification remains the primary bottleneck, with the best-performing model reaching only 0.501.

1 INTRODUCTION

Scientific discovery is not a one-shot generation process, but an iterative cycle in which claims are proposed, challenged, revised, and re-evaluated in light of evidence and critique (Wang et al., 2023). In scholarly publishing, peer review institutionalizes this cycle: reviewers identify weaknesses, request additional evidence or clarification, authors revise their manuscripts, and reviewers or editors then assess whether the new version has actually addressed the raised concerns (Mulligan et al., 2013). Recent advances in LLMs have motivated a vision of AI-assisted scientific discovery, where AI systems participate not only in generating research artifacts but also in evaluating and revising them. This raises a question for iterative AI scientific workflows: can AI verify, with grounded manuscript evidence, whether feedback has actually been resolved in a subsequent revision?

Current AI-driven scientific workflows (Figure 1) have begun to address the first two parts of this cycle: generating scientific artifacts and producing feedback on them. AutoResearch systems aim to automate or assist activities such as literature analysis, hypothesis generation, experiment design, coding, and manuscript preparation (Lu et al., 2024; Schmidgall et al., 2025; Ghareeb et al., 2026). In parallel, AutoReview systems investigate whether LLM-based agents can evaluate manuscripts, identify weaknesses, generate reviewer comments, and assist human reviewers (Liu & Shah, 2023; Zhou et al., 2024; Latona et al., 2024; Cao et al., 2025; Jiang & Ng, 2025; PapeReview, 2026). However, these systems mostly stop at generating artifacts or critiques. They do not directly verify whether a subsequent revision substantively addresses the critique, whether the author's response is reflected in the revised manuscript, or whether the revised text provides sufficient evidence for the paper's scientific claims.

Revision verification poses challenges beyond lexical or semantic matching. Reviewer concerns may target experimental design, claim-evidence alignment, baseline coverage, figure clarity, sta-

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.

tistical reporting, or the scope of conclusions, while corresponding revisions may appear as local clarifications, new analyses, additional experiments, revised visualizations, added limitations, or distributed changes across sections. A verifier must therefore recover the intent of the original concern, interpret the author's claimed resolution, and assess whether the revised manuscript provides evidence that substantively addresses the underlying scientific issue. This assessment must also be grounded in the manuscript rather than in the response letter alone. Author responses describe claimed resolutions, but they do not by themselves establish that the manuscript has changed in the relevant way. A response may overstate the revision, address only part of the concern, or describe a change whose evidential support remains insufficient. Reliable verification must therefore jointly reason over the reviewer concern, the author response, and concrete manuscript evidence.

Existing peer-review datasets provide valuable resources for modeling review artifacts, but they do not directly support grounded revision verification. PeerRead, NLPeer, and MOPRD include combinations of papers, reviews, rebuttals, manuscript versions, meta-reviews, and decisions (Kang et al., 2018; Dycke et al., 2023; Lin et al., 2023). Related resources study meta-review generation, review discussion modeling, review-rebuttal argument relations, review quality assessment, and substantiation of review statements (Cheng et al., 2020; Shen et al., 2022; Kennard et al., 2022; Ghosal et al., 2022; Guo et al., 2023). These datasets have enabled substantial progress in computational peer-review research, but they are primarily organized around reviews, rebuttals, decisions, or discussion structures. They do not define concern-level revision verification as a benchmark task, where a model must jointly read the reviewer concern, the author response, and the revised manuscript to decide whether the concern was resolved and to identify the manuscript evidence supporting that judgment. This leaves the following task largely unaddressed:

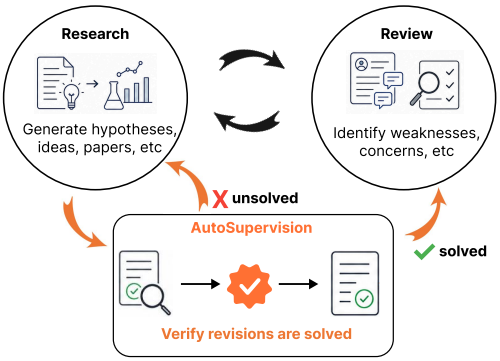


Figure 1: AUTOSUPERVISION for scientific revision verification. Black loop: workflow without AUTOSUPERVISION. Orange loop: workflow with AUTOSUPERVISION: verify whether reviewer concerns are truly addressed.

> Given a reviewer's concern, an author's response, and a revised manuscript, can a model characterize the concerns and whether the concerns have been addressed and identify the exact manuscript evidence supporting this judgment?

To make this question benchmarkable, we introduce AUTOSUPERVISION, a benchmark for AutoSupervision that evaluates concern-level, evidence-grounded verification of scientific manuscript revisions. Each instance is centered on a reviewer concern within a revision episode. Given the concern, the corresponding author response, and the revised manuscript, a model must characterize the concern, determine whether it has been resolved, and identify the manuscript evidence supporting its judgment. This formulation isolates revision assessment from concern discovery, so model performance reflects the ability to verify revisions rather than the ability to extract reviewer comments.

We construct AUTOSUPERVISION from transparent peer-review records of *Nature Communications*, which provide reviewer comments and author responses alongside published articles (Anonymous, 2022). From 56,000 article-level records, we extract revision episodes and concern-level verification instances by aligning reviewer concerns, author responses, and manuscript evidence. We then evaluate LLMs, supervised models, and agentic verification systems. The results show that modern models can often characterize reviewer concerns, but still struggle to determine whether revisions genuinely resolve them and to ground their judgments in manuscript evidence.

Our contributions are summarized as follows:

- We define AutoSupervision as a new evaluation problem for AI-assisted scientific workflows and introduce AUTOSUPERVISION, a benchmark for concern-level, evidence-grounded verification of scientific manuscript revisions.

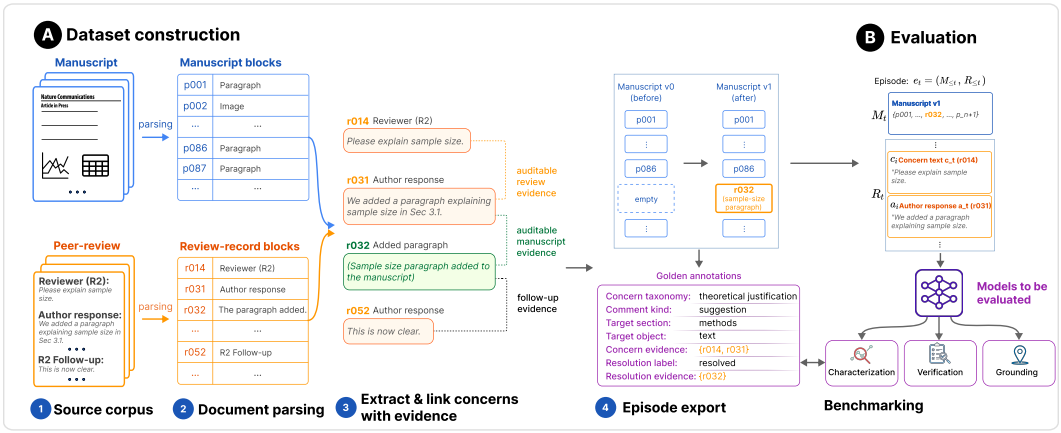


Figure 2: AUTOSUPERVISION construction and evaluation pipeline. First, we pair articles and peer reviews (A1), parse them into blocks with identifiers (A2), extract and link concerns with evidence (A3), and export revision episodes for concern characterization, resolution verification, and evidence grounding (A4). After building AUTOSUPERVISION, we benchmark existing LLMs on this task (B).

- We construct a large-scale dataset from 56,000 transparent peer-review article records from *Nature Communications* and develop a systematic pipeline to align reviewer concerns, author responses, and manuscript revisions, yielding 8,790 episode-level instances with resolution labels and evidence annotations.
- We provide comprehensive evaluations across LLMs, agentic systems, and ablation studies, revealing the remaining challenges in evidence-grounded revision assessment.

2 RELATED WORK

Resources for computational peer review. Computational studies of peer review have produced several valuable resources for modeling scientific evaluation. PeerRead pairs paper drafts with reviews and acceptance decisions, enabling research on review generation, review analysis, and acceptance prediction (Kang et al., 2018). NLPeer unifies peer-review datasets with structured representations, licensing information, and reviewing-assistance tasks (Dycke et al., 2023). MOPRD introduces a multidisciplinary open peer-review dataset containing review comments, rebuttal letters, manuscript versions, meta-reviews, and decisions (Lin et al., 2023). Peer Review Analyze further annotates ICLR reviews with paper-section correspondence, aspect categories, statement purposes, and significance labels (Ghosal et al., 2022).

These datasets enable computational studies of peer review, but primarily focus on review artifacts, paper-level decisions, or properties of reviewer feedback rather than whether feedback is successfully incorporated into revised manuscripts.

Modeling scientific feedback and revision processes. Prior work has studied how reviewers and authors interact through review comments and response documents. APE introduced argument pair extraction between peer reviews and rebuttals, enabling analysis of reviewer arguments and author responses (Cheng et al., 2020). DISAPERE annotated discourse structures in review-rebuttal pairs and characterized author stances toward reviewer arguments (Kennard et al., 2022). MReD studied structured meta-review generation by modeling review discussions and decision rationales (Shen et al., 2022). ARIES (D'Arcy et al., 2024) further connects reviewer comments to manuscript edits. However, it does not by itself establish whether the underlying concern has been resolved.

These efforts provide important foundations for understanding scientific feedback exchange. However, they mainly treat rebuttals as the endpoint. In contrast, AUTOSUPERVISION incorporates the revised manuscript as an additional component and evaluates the complete feedback chain from reviewer concern to author response and evidence of manuscript change.

LLMs for scientific research and review assistance. Recent advances in LLMs have motivated AutoResearch systems that automate or assist scientific research such as literature analysis, hypothesis generation, experimentation, coding, and manuscript preparation (Lu et al., 2024; Schmidgall et al., 2025; Ghareeb et al., 2026). In parallel, AutoReview systems explore the use of LLMs to evaluate manuscripts and provide scientific feedback. ReviewerGPT investigates LLM-based generation of reviewer comments and manuscript critiques (Liu & Shah, 2023). Other studies evaluate the quality and usefulness of LLM-generated reviews and research-paper feedback (Zhou et al., 2024; Liang et al., 2024). CSPaper Review provides fast, rubric-faithful feedback aligned with conference review criteria (Cao et al., 2025). PapeReview offers an agentic pre-submission review workflow that produces structured feedback on manuscript clarity, methodology, and contribution (PapeReview, 2026). The Stanford Agentic Reviewer uses an agentic workflow that retrieves relevant work from arXiv and synthesizes it with manuscript content to generate actionable feedback (Jiang & Ng, 2025). REVAS has also been deployed in ACL Rolling Review to provide reviewers with feedback on the quality of their reviews (ACL Rolling Review, 2026).

These systems automate or assist the generation and improvement of scientific feedback. However, they do not evaluate whether reviewer feedback is subsequently incorporated into a revised manuscript. AUTOSUPERVISION addresses this complementary stage by characterizing reviewer concerns, verifying their resolution, and grounding the resulting judgments in manuscript evidence.

3 TASK DEFINITION

3.1 SCIENTIFIC REVISION LOOP

We formalize scientific revision as an interaction among three roles: an author process, a reviewer process, and an AutoSupervision model. Let $\mathcal{A}$ denote the author, $\mathcal{R}$ denote the reviewer process, and $\mathcal{S}_\theta$ denote the AutoSupervision model to be evaluated. At revision round t, the manuscript before revision is M_{t-1}. A reviewer observes the manuscript and produces a set of reviewer concerns:

$$C_t = \mathcal{R}(M_{t-1}) = \{c_{t,1}, \ldots, c_{t,n_t}\}, \tag{1}$$

where n is the number of concerns at round t. The author then responds to these concerns and revises the manuscript:

$$(M_t, A_t) = \mathcal{A}(M_{t-1}, C_t), \tag{2}$$

where M_t is the post-revision manuscript and $A_t = \{a_{t,1}, \ldots, a_{t,n_t}\}$ denotes the author responses or rebuttals corresponding to the reviewer concerns.

For each concern $c_{t,i}$, we define a review-response record

$$r_{t,i} = (c_{t,i}, a_{t,i}), \tag{3}$$

and let $R_t = \{r_{t,1}, \ldots, r_{t,n_t}\}$ be the set of records associated with revision round t. In the benchmark, $\mathcal{A}$ and $\mathcal{R}$ are not models to be trained. They are observed through public transparent peer-review records: reviewer comments instantiate C_t, author responses instantiate A_t, and the post-revision manuscript M_t. The task is to evaluate whether $\mathcal{S}_\theta$ can audit this interaction.

3.2 ROUND-LEVEL EPISODES

Peer-review data contains multiple versions across revision rounds. We therefore define a *revision episode* as a transition between two consecutive manuscript versions, such as `v0_to_v1` or `v1_to_v2`. Formally, an episode is

$$e_t = (M_{\leq t}, R_{\leq t}). \tag{4}$$

This formulation evaluates each concern together with the author response and manuscript state available at the corresponding revision stage, rather than aggregating information across the entire review history.

3.3 AUTOSUPERVISION PREDICTION TASK

Given an episode e_t, the AutoSupervision model reads the post-revision manuscript together with the review-response history up to and including round t and predicts a set of structured outputs $\hat{Y}_t$:

$$\hat{Y}_t = \mathcal{S}_\theta(e_t) = \{\hat{y}_{t,i}\}_{i=1}^{n_t}. \tag{5}$$

Each predicted element corresponds to one record $r_{t,i}$. It contains concern-characterization fields, revision-verification fields, review-record evidence block IDs, and revised-manuscript grounding block IDs. The gold output for the same episode Y_t^* is

$$Y_t^* = \{y_{t,i}^*\}_{i=1}^{n_t}. \tag{6}$$

Evaluation compares $\hat{Y}_t$ with Y_t^* for every episode, aligning predictions by the fixed review-response record identifiers.

This task is different from automatic review generation and automatic manuscript revision. The model is not asked to generate reviewer concerns C_t, write author responses A_t, or produce a revised manuscript M_t. Instead, $\mathcal{S}_\theta$ audits the closed feedback loop: given what the reviewer requested, what the author claimed, and what appears in the revised manuscript, it must determine whether the concern was addressed and cite the exact evidence supporting that judgment. The full prediction schema and label definitions are provided in Appendix A.

4 Dataset Construction

4.1 Pipeline

Figure 2 (A) summarizes the dataset construction pipeline.

Source Corpus and Splits We collect 56,000 published articles and corresponding transparent peer-review files from *Nature Communications*. Within each publication year, papers are assigned to training, validation, and test partitions using an approximately 8:1:1 ratio. All review rounds and manuscript states from the same paper remain in the same partition, preventing cross-round information leakage while preserving the publication-year distribution across splits.

Document Parsing Both the published articles and their corresponding peer-review files are downloaded in PDF format. Because PDFs primarily encode visual layout rather than structured document content, we first parse them into ordered, typed blocks using MinerU (Wang et al., 2024). The resulting blocks include paragraphs, section headings, figures, tables, equations, references, captions, and other scientific content. Each manuscript block receives a stable identifier with the prefix `p`, while each peer-review block receives an identifier with the prefix `r`.

Concern Extraction For each sampled paper, we use GPT-5.5 (OpenAI, 2026) to process the published article and its peer review. A peer review may contain multiple rounds of reviewer comments, author responses, and reviewer follow-up messages. The model extracts candidate concern threads, links each reviewer concern to its corresponding author response and follow-up discussion, and assigns the resulting review-response record to the revision episode in which the response was made.

During benchmark construction, the annotation model has access to the complete review history. Reviewer follow-ups from subsequent rounds may help align concern-response pairs and filter instances by confirming whether a concern was addressed or remained unresolved. In the standard full-context evaluation setting, however, models receive only the review-response history up to and including the current round. Information from subsequent rounds is excluded from both the model input and the reference evidence block IDs used for scoring.

Each extracted record preserves the original reviewer concern, the corresponding author response, the normalized reviewer identifier, the assigned revision round, and the supporting review-block identifiers. It also includes the concern-characterization, resolution, and target-scope annotations defined by the benchmark schema.

LLM-Assisted Manuscript Reconstruction. To construct round-specific manuscript inputs, we provide the GPT-5.5 with the published article and the complete public peer-review record, including reviewer comments, author responses, and follow-up discussions. The LLM uses these materials to retrospectively reconstruct the manuscript states associated with individual revision rounds. These reconstructed states serve as the manuscript component of the exported revision episodes.

Table 1: Main AUTOSUPERVISION performance results. Coverage is concern-level prediction coverage, and Overall is the equal-weight average of characterization, verification, and manuscript grounding. Bold and underlined values indicate the best and second-best, respectively.

System	Coverage	Characterization	Verification	Grounding	Overall
Claude-Opus-4.8 (Anthropic, 2026)	0.977	0.691	**0.501**	**0.719**	**0.637**
GPT-5.5 (OpenAI, 2026)	1.000	**0.754**	0.455	0.702	**0.637**
GLM-5.2 (Z.ai, 2026)	1.000	0.698	0.475	<u>0.711</u>	<u>0.628</u>
Gemini-3.1-Pro (Google DeepMind, 2026)	0.997	0.698	0.473	0.700	0.624
Qwen3.7-max (Qwen Team, 2026b)	1.000	0.701	0.463	0.699	0.621
SFT Qwen3.5-9B	1.000	<u>0.708</u>	<u>0.490</u>	0.643	0.614
DeepSeek-V4-Pro (Xu et al., 2026)	1.000	0.668	0.451	0.575	0.565
Kimi-K2.6 (Moonshot AI, 2026)	1.000	0.653	0.455	0.571	0.560
DeepSeek-V3.2 (Liu et al., 2025)	1.000	0.677	0.396	0.551	0.541
MiniMax-2.7 (MiniMax, 2026)	1.000	0.626	0.373	0.563	0.521
DeepSeek-V4-Flash (Xu et al., 2026)	1.000	0.629	0.399	0.443	0.490
Intern-S2-35B (InternLM Team, 2026)	1.000	0.594	0.351	0.470	0.472
Qwen3.5-9B (Qwen Team, 2026a)	1.000	0.563	0.376	0.451	0.463
GPT-4o-mini agentic (OpenAI, 2024)	0.998	0.573	0.300	0.336	0.403
GPT-4o-mini (OpenAI, 2024)	0.933	0.497	0.216	0.085	0.266

Episode Export We apply conservative filtering to ensure that retained instances represent actionable and auditable scientific revision requests. We exclude: 1) praise-only comments and broad paper-level assessments without a specific revision request, 2) concerns without verifiable reviewer-comment evidence, and 3) concerns whose targets exist only in the rebuttal letter, an external repository, or other unavailable artifacts.

When a concern recurs across multiple revision rounds, we retain the reviewer comment and corresponding author response for each relevant episode while storing later follow-up discussion. This prevents short follow-up statements, such as "OK" or "Addressed," from replacing the original scientific concern and response context.

For each valid transition from manuscript version $t-1$ to version t, we export one revision episode e_t. Each episode includes the manuscripts and review-response records up to and including round t. Although earlier rounds are retained as contextual information, only the concern-response pairs associated with the current round are prediction and evaluation targets. Information from subsequent rounds may support annotation and filtering, but is excluded from the evaluation input and the reference evidence used for scoring.

4.2 DATASET STATISTICS

For the benchmark evaluation in this paper, we sample 10% of the test papers from each year using a fixed random seed, yielding 557 candidate papers. After review-response extraction, revision alignment, and quality filtering, the resulting benchmark contains 847 revision episodes. Across these episodes, AUTOSUPERVISION contains 6,455 unique point-level reviewer concerns. Because a concern may remain active or recur across multiple revision rounds, the unique concerns correspond to 8,790 episode-level concern instances. These episode-level instances are the primary units used for benchmark evaluation. More details about dataset distribution are provided in Appendix B.1. To evaluate the robustness of our annotations to the choice of model, we measure inter-model agreement among the annotation outputs, as reported in Appendix B.2.

5 EVALUATION

5.1 EVALUATION PIPELINE

Figure 2 (B) summarizes the Evaluation pipeline. We evaluate models under a unified structured prediction setting. All models must output JSON conforming to the benchmark schema. Missing concern IDs are scored as missing predictions. The evaluator aligns predictions by the fixed input concern IDs.

The benchmark evaluates three complementary capabilities: Characterization, Verification, and Grounding. The Characterization score averages comment-kind macro-F1, point-type macro-F1, and a hierarchical target-scope score over target domain, section, and object. The Verification score averages resolution-label macro-F1, paper-status macro-F1, and micro-F1 over evidence block IDs supporting the resolution judgment. The Grounding score is computed as micro-F1 over target paper block IDs, which identify the revised-manuscript evidence blocks where a reviewer concern is addressed. Please refer to Appendix A for details. The overall score assigns equal weights to the three components:$S_{\text{overall}} = \frac{1}{3}\left(S_{\text{char}} + S_{\text{ver}} + S_{\text{ground}}\right)$.

5.2 Evaluated Models

We evaluate AutoSupervision on a diverse set of recent LLMs, covering both closed-source and open-source models. The evaluated models span different model families and scales, enabling a comprehensive assessment of current capabilities for reviewer concern analysis.

Closed-source models. We include several proprietary frontier models, including Claude-Opus-4.8 (Anthropic, 2026), GPT-5.5 (OpenAI, 2026), Qwen3.7-max (Qwen Team, 2026b) and Gemini-3.1-Pro (Google DeepMind, 2026). These models represent state-of-the-art general-purpose language models with strong reasoning and instruction-following capabilities. We additionally evaluate GPT-4o-mini (OpenAI, 2024) and its agentic variant to examine the performance gap between earlier-generation compact models and recent frontier systems.

Open-source models. We evaluate a broad range of open-source models, including GLM-5.2 (Z.ai, 2026), DeepSeek-V4-Pro (Xu et al., 2026), Kimi-K2.6 (Moonshot AI, 2026), DeepSeek-V3.2 (Liu et al., 2025), MiniMax-2.7 (MiniMax, 2026), DeepSeek-V4-Flash (Xu et al., 2026), Intern-S2-35B (InternLM Team, 2026), and Qwen3.5-9B (Qwen Team, 2026a). We also include an instruction-tuned Qwen3.5-9B model to investigate the impact of supervised fine-tuning on reviewer concern understanding.

5.3 Main Results

Table 1 summarizes benchmark performance across closed-source and open-source models. Overall, the results reveal a substantial capability gap between identifying reviewer concerns and performing evidence-based assessment of revisions. While modern models can almost always extract and characterize reviewer concerns, accurately verifying whether a revision resolves the concern remains a significant challenge.

Among all evaluated systems, Claude-Opus-4.8 and GPT-5.5 achieve the highest overall performance (both 0.637), followed by GLM-5.2 (0.628) and Gemini-3.1-Pro (0.624). Interestingly, the performance difference between the strongest closed-source and open-source models is relatively small, with GLM-5.2 approaching frontier proprietary systems. However, their remaining limitations are primarily concentrated in deeper reasoning tasks that require evaluating the relationship between a claimed revision and the underlying manuscript evidence.

Coverage Coverage is concern-level prediction coverage. Nearly all models achieve perfect or near-perfect coverage, indicating that modern LLMs rarely fail to identify the presence of reviewer concerns. The only notable degradation occurs for smaller or earlier-generation models, such as GPT-4o-mini. Given the fixed input concern set, most models achieve near-complete prediction coverage, indicating reliable structured-output completion.

Characterization Characterization measures whether a model correctly identifies the type and scope of a reviewer concern. GPT-5.5 achieves the strongest performance (0.754), while most frontier models exceed 0.65. The high characterization scores, combined with near-perfect coverage, indicate that models possess strong semantic understanding of reviewer feedback.

Verification Verification is the most challenging dimension. Claude-Opus-4.8 obtains the highest verification score (0.501), followed by the Qwen3.5-9B SFT model (0.490) and GLM-5.2 (0.475). This pattern is notable because verification does not simply require language understanding; it requires multi-step reasoning over the reviewer concern, the claimed revision, and the manuscript

Table 2: Ablation study. GPT-5.5 performance under different input configurations.

Setting	Characterization	Verification	Grounding	Overall
Concern-only	0.676	0.042	0.000	0.239
Concern+Response	0.696	0.236	0.000	0.311
Concern+Manuscript	0.719	0.394	0.633	0.582
Full context	0.754	0.455	0.702	0.637

content. The results suggest that current models often recognize the intent of a revision but struggle to determine whether the provided evidence is sufficient, complete, and scientifically relevant.

Grounding Manuscript grounding evaluates whether models can identify supporting evidence. Claude-Opus-4.8 achieves the highest grounding score (0.719), followed closely by GLM-5.2 (0.711), GPT-5.5 (0.702), and Gemini-3.1-Pro (0.700). In contrast, smaller models show substantial degradation, with GPT-4o-mini achieving only 0.085. The large performance gap highlights that grounding requires more than generating plausible explanations: models must accurately retrieve and connect manuscript-level evidence to reviewer concerns.

Overall, these results demonstrate that AUTOSUPERVISION evaluates capabilities beyond reviewer feedback understanding. Existing LLMs already achieve strong performance in detecting and characterizing concerns, but they remain limited in evidence-based verification and grounding.

5.4 IMPACT OF AGENTIC INFERENCE

Beyond scaling model size, we evaluate whether a fixed structured inference pipeline can improve performance. The agentic baseline uses GPT-4o-mini in a three-stage pipeline for concern characterization, manuscript grounding, and resolution verification. Between characterization and grounding, we add a deterministic manuscript-block retriever that narrows the full manuscript to a high-recall candidate evidence set for each concern. The model then performs grounding and verification over this smaller candidate context and outputs the same schema as all other systems. Appendix C provides the full pipeline details.

This pipeline improves GPT-4o-mini substantially over the single-shot setting: overall score increases from 0.266 to 0.403, coverage from 0.933 to 0.998, and manuscript grounding from 0.085 to 0.336. The improvement shows that explicit retrieval, decomposition, and schema normalization help models produce more complete and better-grounded predictions.

5.5 IMPACT OF SUPERVISED FINE-TUNING

We further examine the effect of task-specific supervised fine-tuning by comparing Qwen3.5-9B with its supervised variant. 500 papers are randomly sampled from the training dataset split for training. Appendix D provides reproducibility details, including the data construction, model, training, and inference details.

SFT Qwen3.5-9B improves overall performance from 0.463 to 0.614, with gains across all three components: characterization improves from 0.563 to 0.708, verification from 0.376 to 0.490, and grounding from 0.451 to 0.643. The improvements indicate that supervised adaptation helps models better capture the knowledge of AUTOSUPERVISION, particularly the alignment between reviewer concerns, author responses, and manuscript evidence. The substantial gain in grounding further suggests that task-specific supervision improves evidence localization, enabling models to better identify relevant manuscript support for reviewer feedback.

5.6 ABLATION STUDY

Table 2 analyzes how different sources of information contribute to AUTOSUPERVISION performance. We progressively remove components from the full-context setting to understand which capabilities depend on reviewer information, author responses, and manuscript evidence.

When only reviewer concerns are provided, GPT-5.5 achieves a relatively strong characterization score (0.676), but verification performance drops substantially (0.042), and grounding is not valid

Table 3: Closed-loop revision case study. Outcomes are concern-level pairwise comparisons between the first and second revision. Net is the improved minus the worse.

Supervisor	Improved	Same	Worse	Net
No supervision	18.0%	66.5%	15.5%	0.025
DeepSeek-V4-Flash	29.2%	70.8%	0.0%	0.292
GLM-5.2	35.4%	64.6%	0.0%	0.354

by design because no manuscript evidence is available. This indicates that reviewer text contains sufficient information to identify the type and scope of a concern, but is insufficient for determining whether the concern has been addressed. Adding author responses improves verification to 0.236 while maintaining strong characterization (0.696). However, grounding remains unavailable without manuscript evidence. This suggests that response letters provide useful information about the claimed resolution, but these claims cannot be validated without examining the actual revision. Providing concerns and manuscript inference substantially improves grounding, achieving 0.633 grounding and 0.582 overall performance. This demonstrates that manuscript evidence is the primary source for locating revision outcomes. Nevertheless, it remains below full-context performance, particularly on verification (0.394 vs. 0.455), because the model lacks the complete context, such as the author's explanation of the change.

5.7 Closed-Loop Revision Case Study

We also test whether the benchmark is useful for choosing a verifier that improves a downstream revision loop. This case study asks whether systems that are stronger under the benchmark-style supervisor task provide more useful feedback when inserted between two author-revision calls. The pipeline starts from an initial submission draft, uses fixed human reviewer concerns from the review record, asks GPT-5.5 to produce a first revision, applies a AutoSupervision-style supervisor to the first revision, renders the supervisor output as author-facing feedback, asks GPT-5.5 for a second revision, and finally uses a blinded Claude Opus 4.8 judge to compare concern-level outcomes. We evaluate 10 sampled papers. The no-supervision condition receives the raw reviews and author-response context but no structured concern-level supervisor output. The results are shown in Table 3.

GLM-5.2 produces the largest number of useful second-revision changes. DeepSeek V4 Flash is weaker but still improves over the no-supervisor condition. Without supervisor feedback, some concerns are improved while more regressions are introduced. This supports the practical claim that AutoSupervision-style verification can provide more useful revision guidance.

6 Conclusion

We introduced AutoSupervision, a benchmark for scientific manuscript revision verification and concern-to-revision grounding. AutoSupervision provides structured reviewer concerns, author responses, revised manuscript evidence, resolution labels, and exact evidence block annotations. Our evaluation shows that existing LLMs achieve strong characterization ability, but still struggle with verifying whether a revision genuinely addresses a reviewer concern and identifying the supporting manuscript evidence.

Limitations. The current benchmark is based on Nature Communications transparent peer-review records, which mainly include papers that successfully completed the revision process. This may limit the diversity of revision trajectories and overrepresent well-addressed concerns. Future extensions could include broader venues and review outcomes to evaluate models under more diverse scenarios. The round-specific manuscript states are reconstructed by an LLM and may not exactly reproduce the original historical submissions. Our benchmark therefore assesses verification over reconstructed revision episodes rather than an exact replay of the original review process. Moreover, AutoSupervision focuses on whether models can assess revision resolution from available evidence. In practice, additional capabilities such as generating revision suggestions, prioritizing reviewer concerns, and interacting with authors or reviewers are also important directions beyond the current benchmark scope.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data sets, implement methods, assist with translation, clean and reformat dataset. We have not used generative AI tools to formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, support qualitative and thematic data analysis, or interpret results. Providing critical ingredients for proving mathematical claims, helping develop theoretical models or conceptual frameworks, assisting in the writing of proofs are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

AUTOSUPERVISION is built from publicly available transparent peer-review records and published articles. These records were released by the source journal as part of its transparent review process, and our benchmark uses them only to study concern-level revision verification and evidence grounding. These materials are used to assess model behavior, rather than the competence or conduct of individual authors, reviewers, or editors. The code adheres to ethical principles and has no known ethical concerns.

REPRODUCIBILITY STATEMENT

In addition to the details we provide in Section 4 and Appendix D, we will also publicly release the benchmark dataset and code needed to reproduce the reported results.

REFERENCES

ACL Rolling Review. Using the REVAS review assistant tool in the ARR-May cycle. https://aclrollingreview.org/revas-may26, 2026. ACL Rolling Review blog post.

Anonymous. Transparent peer review for all. *Nature Communications*, 13:6173, 2022. doi: https://doi.org/10.1038/s41467-022-33056-8.

Anthropic. Introducing Claude Opus 4.8. https://www.anthropic.com/news/claude-opus-4-8, May 2026. Accessed: 2026-07-27.

Lele Cao, Lei You, and R&D Team. CSPaper review: Fast, rubric-faithful conference feedback. In Lucie Flek, Shashi Narayan, Lê Hong Phuong, and Jiahuan Pei (eds.), *Proceedings of the 18th International Natural Language Generation Conference: System Demonstrations*, pp. 3–7, Hanoi, Vietnam, October 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.inlg-demos.2/.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7000–7011, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.569. URL https://aclanthology.org/2020.emnlp-main.569/.

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. ARIES: A corpus of scientific paper edits made in response to peer reviews. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6985–7001, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.377. URL https://aclanthology.org/2024.acl-long.377/.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. NLPeer: A unified resource for the computational study of peer review. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pp. 5049–5073, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.277. URL https://aclanthology.org/2023.acl-long.277/.

Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J Szostkiewicz, Dmytro Shved, Gavin J Gyimesi, Jon M Laurent, Samantha M Wright, Muhammed T Razzak, et al. A multi-agent system for automating scientific discovery. *Nature*, pp. 1–3, 2026.

Tirthankar Ghosal, Sandeep Kumar, Prabhat Kumar Bharti, and Asif Ekbal. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *Plos one*, 17(1):e0259238, 2022. doi: https://doi.org/10.1371/journal.pone.0259238.

Google DeepMind. Gemini 3.1 Pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Pro-Model-Card.pdf, February 2026. Accessed: 2026-07-27.

Yanzhu Guo, Guokan Shang, Virgile Rennard, Michalis Vazirgiannis, and Chloé Clavel. Automatic analysis of substantiation in scientific peer reviews. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10198–10216, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.684. URL https://aclanthology.org/2023.findings-emnlp.684/.

InternLM Team. Intern-S2-Preview: An efficient 35b scientific multimodal foundation model. https://huggingface.co/internlm/Intern-S2-Preview, 2026. Model card; accessed: 2026-07-27.

Yixing Jiang and Andrew Ng. Stanford agentic reviewer: Tech overview. https://paperreview.ai/tech-overview, 2025. Accessed: 2026-07-22.

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 Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1647–1661, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1149. URL https://aclanthology.org/N18-1149/.

Neha Nayak Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. DISAPERE: A dataset for discourse structure in peer review discussions. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1234–1249, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.89. URL https://aclanthology.org/2022.naacl-main.89/.

Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R Davidson, Veniamin Veselovsky, and Robert West. The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates, 2024.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196, 2024. doi: https://doi.org/10.48550/arXiv.2310.01783.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. Moprd: A multidisciplinary open peer review dataset. *Neural Computing and Applications*, 35(34):24191–24206, 2023. doi: https://doi.org/10.1007/s00521-023-08891-5.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models, 2025.

Ryan Liu and Nihar B Shah. Reviewergpt? an exploratory study on using large language models for paper reviewing, 2023.

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.

MiniMax. The MiniMax-M2 series: Mini activations unleashing max real-world intelligence, 2026.

Moonshot AI. Kimi K2.6 tech blog: Advancing open-source coding. https://www.kimi.com/blog/kimi-k2-6, April 2026. Accessed: 2026-07-27.

Adrian Mulligan, Louise Hall, and Ellen Raphael. Peer review in a changing world: An international study measuring the attitudes of researchers. *Journal of the American Society for Information Science and Technology*, 64(1):132–161, 2013.

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, July 2024. Accessed: 2026-07-27.

OpenAI. GPT-5.5. https://openai.com/index/gpt-5-5-system-card/, April 2026. Accessed: 2026-07-22.

PapeReview. PapeReview: Agentic ai peer-review for research manuscripts. https://papereview.com/, 2026. Accessed: 2026-07-22.

Qwen Team. Qwen3.5-9B model card. https://huggingface.co/Qwen/Qwen3.5-9B, 2026a. Accessed: 2026-07-27.

Qwen Team. Qwen3.7: The agent frontier. https://qwen.ai/blog?id=qwen3.7, May 2026b. Accessed: 2026-07-27.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using LLM agents as research assistants. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5977–6043, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.320. URL https://aclanthology.org/2025.findings-emnlp.320/.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. MReD: A meta-review dataset for structure-controllable text generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2521–2535, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.198. URL https://aclanthology.org/2022.findings-acl.198/.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. Mineru: An open-source solution for precise document content extraction, 2024.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

Z.ai. GLM-5.2: Built for long-horizon tasks. https://z.ai/blog/glm-5.2, June 2026. Accessed: 2026-07-27.

Ruiyang Zhou, Lu Chen, and Kai Yu. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9340–9351, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.816/.

A Prediction Schema and Evaluation Details

Given a revision episode $e_t = (M_{\leq t}, R_{\leq t})$, the model produces one structured prediction for each review-response record $r_{t,i}$. Each prediction consists of three groups of fields: *Characterization*, *Verification*, and *Grounding*. Table 4 summarizes all scored fields and their valid options.

Characterization. Characterization evaluates whether a model correctly identifies the nature and scope of a reviewer comment. It includes three aspects: `comment_kind`, which describes whether a comment is a critique, question, or suggestion; `point_type`, which identifies the underlying issue category; and the target scope of the comment. We define *target scope* as the joint specification of `target_domain`, `target_section`, and `target_object`. Accordingly, S_{scope} denotes the hierarchical target-scope score computed from these three fields.

Verification. Verification evaluates whether a model can determine whether a reviewer concern has been addressed in the revised manuscript. It consists of `paper_status`, which describes the observed degree of change in the revised manuscript; `resolution_label`, which indicates whether the concern is resolved, partially resolved, or unresolved; and `resolution_evidence_block_ids`, which identifies the evidence blocks supporting the resolution judgment. The reference IDs for this field are restricted to review-record blocks included in the episode's full-context evaluation input. Subsequent-round blocks consulted only during benchmark construction are not scoring targets.

Grounding. Grounding evaluates whether a model can identify the exact revised-manuscript evidence supporting where a reviewer concern is addressed. It is measured using `target_paper_block_ids`, which specifies the set of revised-manuscript evidence blocks where the concern is addressed or grounded in the paper.

Component scores. Categorical fields are evaluated using macro-F1 to reduce the influence of label-frequency imbalance. For ontology labels, macro-F1 is computed over labels with positive gold support. Concerns with the `unverifiable` resolution label are excluded according to the benchmark evaluation protocol.

Predicted evidence block ID sets are evaluated using micro-F1 because block identifiers are document-specific references rather than shared categorical labels.

The target scope is evaluated using S_{scope}, a hierarchical score that assigns partial credit according to the granularity of the matched target:

$$S_{\text{scope}} = \frac{1}{N} \sum_{i=1}^{N} \left(0.2 I_d + 0.4 I_s + 0.4 I_o\right),$$

where I_d, I_s, and I_o indicate exact matches of the target domain, target section, and target object, respectively.

The three component scores are:

$$C_{\text{char}} = \frac{1}{3}\left(F_{\text{kind}} + F_{\text{type}} + S_{\text{scope}}\right),$$

$$C_{\text{ver}} = \frac{1}{3}\left(F_{\text{res}} + F_{\text{status}} + F_{\text{evidence}}\right),$$

$$C_{\text{ground}} = F_{\text{ground}}.$$

Here, F_{kind} and F_{type} denote the macro-F1 scores for `comment_kind` and `point_type`, respectively. F_{res} and F_{status} denote the macro-F1 scores for `resolution_label` and `paper_status`. F_{evidence} denotes the micro-F1 score for `resolution_evidence_block_ids`, measuring whether the model identifies evidence

Table 4: Scored prediction fields and options. Target scope refers jointly to target domain, target section, and target object. Manuscript grounding refers to exact revised-manuscript evidence blocks supporting where a concern is addressed.

Field	Options
Characterization	
Comment kind	critique; question; suggestion
Point type	novelty/significance; claim-evidence support; methodological soundness; experimental adequacy; baseline coverage; ablation coverage; statistical/reporting quality; reproducibility/implementation detail; writing/organization; figure/table presentation; theoretical/mechanistic justification; artifact/confound; clinical/application relevance; dataset/evaluation protocol; citation/related work; notation/internal consistency; supplementary/raw-evidence sufficiency; editorial/journal compliance; conclusion/generalization scope; limitation/scope
Target domain	main text; supplement
Target section	title; abstract; introduction; methods; results; discussion; conclusion; whole paper
Target object	text; figure; table; equation; reference; data
Verification	
Paper status	clear change; partial change; no visible change
Resolution label	resolved; partially resolved; unresolved
Resolution evidence block IDs	Set of evidence block IDs supporting the resolution judgment.
Grounding	
Target paper block IDs	Set of exact revised-manuscript evidence blocks where the concern is addressed or grounded in the paper.

blocks that support its resolution judgment. Finally, F_{ground} denotes the micro-F1 score for `target_paper_block_ids`, measuring whether the model identifies the exact revised-manuscript evidence blocks where the concern is addressed.

The overall benchmark score is the average of the three component scores:

$$C_{\text{overall}} = \frac{1}{3}\left(C_{\text{char}} + C_{\text{ver}} + C_{\text{ground}}\right).$$

B Details of Dataset

B.1 Concern Attribute Distributions

Tables 5 and 6 summarize the concern-level composition of the final evaluation subset. As expected for accepted papers, most concerns are ultimately resolved, but the dataset still retains partially resolved and unresolved cases that require models to distinguish complete revision from incomplete response. The target-object distribution shows that revision verification is not a prose-only task: 1,231 concerns target figures, and additional concerns target data or analysis blocks, references, tables, and equations. The point-type distribution further shows that the benchmark spans a broad taxonomy of scientific revision needs, including writing and organization, claim-evidence support, statistical reporting, mechanistic justification, reproducibility details, figure/table presentation, experimental adequacy, methodology, scope, related work, and dataset or evaluation protocol concerns.

Table 5: Gold-label distributions in the full benchmark.

Field	Label	Count
comment_kind	critique	4,419
comment_kind	suggestion	2,843
comment_kind	question	1,528
resolution_label	resolved	8,265
resolution_label	partially resolved	434
resolution_label	unresolved	91
target_object	text	6,759
target_object	figure	1,231
target_object	data	365
target_object	reference	276
target_object	table	78
target_object	equation	81

Table 6: Point-type distribution in the full benchmark.

Point type	Count	Point type	Count
writing/organization	1,242	notation/internal consistency	433
claim-evidence support	944	methodological soundness	412
theoretical/mechanistic justification	796	conclusion/generalization scope	399
statistical/reporting quality	786	citation/related work	383
reproducibility/implementation detail	706	dataset/evaluation protocol	355
figure/table presentation	643	limitation/scope	316
experimental adequacy	581	artifact/confound	290
clinical/application relevance	212	novelty/significance	96
baseline coverage	68	supplementary/raw-evidence sufficiency	63
editorial/journal compliance	39	ablation coverage	26

B.2 Agreement Experiment

Because the gold construction uses LLM-assisted structured extraction, we add an agreement check on 100 randomly sampled papers from the benchmark test split. The concern set is fixed to the GPT-5.5 output before running the second model. It covers 160 revision episodes and 1,600 fixed concern instances. Table 7 summarizes agreement between GPT-5.5 and Claude Opus 4.8.

The two models agree strongly on the benchmark construction. Target section reaches 0.873 exact agreement and 0.796 Cohen's κ, point type reaches 0.562 exact agreement and 0.525 κ, and the resolution label reaches 0.848 exact agreement. For target manuscript evidence, the teachers achieve 0.710 block micro-F1 and 0.673 mean IoU. Resolution evidence is harder because it can include several adjacent rebuttal in the review record. Across all such blocks, agreement reaches 0.441 micro-F1 and 0.308 mean IoU.

C Agentic Baseline Details

The agentic baseline uses the same base model, GPT-4o-mini, for all LLM calls and produces the same JSON prediction schema used by the single-shot baselines.

Concern characterization. The first stage reads the fixed review-response records for an episode and predicts concern-level characterization fields, including comment kind, point type, target domain, target section, target object, and response status. It also generates short search phrases used by the downstream manuscript retriever.

Table 7: Agreement on 100 sampled benchmark papers. Claude Opus 4.8 re-annotates fixed GPT-5.5 concern ids using the same structured schema.

Categorical labels Field	Acc.	κ	AC1
Comment kind	0.776	0.649	0.716
Point type	0.562	0.525	0.540
Target section	0.873	0.796	0.860
Resolution label	0.848	0.261	0.837
Evidence block sets Field	Precision	F1	IoU
Target manuscript blocks	0.773	0.710	0.673
Resolution evidence	0.648	0.441	0.308

Deterministic manuscript-block retrieval. For each concern, it builds a query from the reviewer concern, author response, current-round discussion, the preliminary characterization labels, and the generated search phrases. It then scores every revised-manuscript block using lexical overlap, object cues such as figure/table/equation/reference mentions, and section/object consistency cues. The top 12 ranked blocks are retained as seed candidates, and a one-block neighbor window is added around each selected block to preserve local context. This top-12 setting is a recall-oriented candidate budget: these blocks are not the final predicted evidence, but the candidate pool from which the grounding stage selects final manuscript evidence IDs.

Grounding. The grounding stage receives each concern, its preliminary characterization, and the retrieved candidate manuscript blocks. It selects the target manuscript block IDs, predicts paper-side change status, and records whether manuscript grounding is available.

Verification. The verification stage receives the concern, author response, preliminary characterization, preliminary grounding, and selected manuscript evidence. It predicts response status, paper status, resolution label, and review-record evidence block IDs. If the grounding stage selects no manuscript blocks, the verification stage receives the top retrieved candidates as fallback context.

Normalization and cost. Finally, deterministic normalization merges the characterization, grounding, and verification outputs into one prediction per concern and enforces the benchmark schema. Grounding and verification are run in chunks of four concerns. The agentic pipeline improves structured-output reliability but is substantially more expensive than single-shot GPT-4o-mini, using 91.9M total tokens compared with 20.0M tokens for the single-shot baseline.

D Supervised Fine-tuning Details

Data preparation. The supervised model is trained from Qwen3.5-9B using the same structured prediction format as the main benchmark. The training data for SFT are generated from the dataset training split. We used 809 training episodes from 500 papers with 9,211 concern instances for SFT.

Training. Training was conducted on two NVIDIA H100 GPUs using full-parameter fine-tuning. The configuration uses the qwen3-5 model type, DeepSpeed ZeRO-3, bfloat16 precision, a maximum sequence length of 32,768, AdamW optimization, a learning rate of (1×10^{-5}), cosine learning-rate scheduling, a warmup ratio of 0.03, weight decay of 0.1, gradient checkpointing, a per-device batch size of 1, gradient accumulation of 2, and a random seed of 42.

Inference and evaluation. The checkpoint is served with an OpenAI-compatible vLLM endpoint, tensor parallel size 8, and maximum model length 32,768.