

PREREVIEWBENCH: Benchmarking LLMs for Editorial Desk Screening in Scholarly Publishing

Ming Chen^{1,2*} and Zhiyuan Ning^{3*†} and Yi Du^{1,2,4‡}

¹Computer Network Information Center, Chinese Academy of Sciences

²University of Chinese Academy of Sciences

³Westlake University

⁴Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences

chenming@cnic.cn, ningzhiyuan@westlake.edu.cn, duyic@cnic.cn

Abstract

Before any reviewer is assigned, a journal editor decides whether a manuscript merits external review. Yet this routine desk screening has no benchmark. We present PREREVIEWBENCH, the first controlled benchmark for journal desk screening. Because real desk decisions are confidential, we pair Web of Science papers published in 2025 with journals to construct proxy labels: true matches, within-field caliber surpluses and deficits, and cross-field scope mismatches. The 7,292 balanced paper–journal instances span four broad areas and 58 journals, supporting desk triage, branch-specific assessment (caliber fit or rejection cause), and transfer-journal recommendation. Evaluating four frontier LLM families, we find average triage accuracy of 0.965 on cross-field scope mismatches but only 0.572 on within-field caliber deficits, where blind expert-led review reproduces the triage labels perfectly. Transfer recommendation is weaker still, with the best Hit@1 at just 0.369. Models are also sensitive to input perturbations, and a cue suggesting that most papers merit review sharply shifts GPT-5.5’s decisions. These results point to a clear division of labor: LLMs can already carry the scope side of desk screening, while caliber judgment stays with editors. PREREVIEWBENCH makes that harder half measurable: whether an in-scope paper is strong enough for its journal.

1 Introduction

Peer review is among the scarcest resources in science: by one estimate, more than 130 million hours went to reviewing manuscripts worldwide in 2020 alone (Aczel et al., 2021). Yet before any reviewer spends a single hour, the journal editor must judge whether a submission merits external review at all.

This desk screening is a routine part of journal publishing: surveys report desk-rejection rates from 5% to 76% across journals (Garand and Harman, 2021), and some journals filter out roughly half of all submissions before external review (Primack et al., 2019). As the most upstream step in the review pipeline, desk screening directly controls reviewer workload and author waiting time.

Research on automated scientific review almost entirely evaluates what happens after a manuscript enters review, mostly in conference settings, from review datasets and task surveys (Kang et al., 2018; Dycke et al., 2023; Staudinger et al., 2024) to studies using large language models (LLMs) to write or assess reviews (Du et al., 2024; Ou et al., 2025). The two research lines closest to desk screening stop short of this decision: acceptance prediction (Kang et al., 2018) targets the final outcome after external review, and journal recommendation systems (Schuemie and Kors, 2008; Pradhan and Pal, 2020; Pradhan et al., 2020) help authors choose venues but make no gatekeeping judgment. The decision itself, whether a manuscript should be sent for review, still lacks any benchmark.

Journal desk screening stands apart from conference peer review as a decision problem in its own right. (1) Evidence is limited: editors hold the full manuscript, but in practice desk decisions rest primarily on the title, abstract, keywords, and the target journal’s profile. (2) The judgment is coupled: an editor weighs topical fit and caliber simultaneously, and the same paper sent to a different journal can yield the opposite verdict; there is no journal-independent notion of “paper quality.” (3) The decision has a downstream consequence: a rejection immediately raises the question of where the paper should go instead. A benchmark that reflects real desk screening should therefore capture all three characteristics, beyond a single send-or-reject binary.

*Equal contribution.

†Project lead.

‡Corresponding author.

We present PREREVIEWBENCH, the first controlled benchmark for journal desk screening. Because real desk-screening records are confidential, we start from the metadata of papers published in 2025 and indexed in Web of Science, and use controlled pairing to construct proxy labels for editorial judgment: true matches, within-field caliber surpluses and deficits, and cross-field scope mismatches. The benchmark contains 7,292 balanced paper–journal instances across four broad areas, eight subfields, and 58 journals, and supports three progressive tasks that follow the editor’s decision path: desk triage (send or reject); branch-specific assessment (caliber fit on the send side, rejection cause on the reject side); and transfer-journal recommendation for rejected papers. To validate the proxy labels, four expert-led annotation teams, one per broad area, blind-reviewed 200 cases under the same information condition as the models (title, abstract, keywords, and target journal profile only); their agreement with the labels reached 96.5%.

We evaluate four frontier LLM families (GPT, Claude, DeepSeek, Qwen) on PREREVIEWBENCH across prompt variants, journal-context ablations, and stress tests, alongside non-LLM baselines. The results reveal a clear capability gap: current LLMs judge “wrong field” cases reliably, with average triage accuracy of 0.965 on cross-field scope mismatches. Judging whether an in-scope paper meets a journal’s caliber bar is far weaker: within-field caliber-deficit accuracy drops to 0.572, while the field-matched annotators, given the same information, agree perfectly with the triage labels on this scenario. The shortfall is not label noise: on the audited cases, humans make this judgment and current LLMs do not. Transfer recommendation is weaker still: the best model’s Hit@1 reaches only 0.369, and counting metadata-derived alternative venues lifts it only to 0.636. Under stress tests, every model struggles with degraded abstracts, most lose accuracy under injected noise keywords, and a prompt cue suggesting that most papers merit review sharply skews GPT-5.5’s decisions. Each family also shows stable decision tendencies across prompts, with some leaning conservative and others permissive.

These results point to a clear division of labor: models already screen scope mismatches reliably enough to support triage-level editorial assistance, while their unreliable caliber judgment requires that humans stay in the decision loop. We argue for journal-aware, human-AI editorial collaboration, in

which systems assess fit to a specific journal rather than quality in isolation. The harder problem to tackle next is within-field caliber judgment.

Our contributions are fourfold:

1. We construct PREREVIEWBENCH, the first controlled benchmark for journal desk screening, with 7,292 balanced paper–journal instances and three progressive tasks; code and data are available at https://github.com/kg4sci/PRB_db.
2. We design a controlled pipeline that constructs the caliber and scope proxy labels through journal stratification, outlier filtering, and citation-stratified pairing, validated by the 200-case expert blind review.
3. We systematically evaluate four LLM families across prompt variants, journal-context ablations, stress tests, and a send-prior intervention, with non-LLM baselines and contamination probes.
4. We map the capability boundary of LLMs on desk screening: reliable scope screening (0.965), within-field caliber judgment far below annotator agreement (0.572), and weak transfer recommendation. Until this gap closes, the reliable role for LLMs remains editorial assistance.

2 Related Work

Peer-review Benchmarks PEERREAD and NLPEER provide conference manuscripts, reviews, and accept/reject labels (Kang et al., 2018; Dycke et al., 2023); other work analyzes review argument structure, predicts review scores, or supports review-based decisions (Hua et al., 2019; Li et al., 2020; Dycke et al., 2021). This literature remains predominantly conference-centered (Staudinger et al., 2024). In materials science, recent resources organize the literature itself: an expert-annotated corpus of CO₂-reduction electrocatalysts and synthesis procedures (Chen et al., 2024) and CITE, a heterogeneous text-attributed citation graph of catalytic literature (Zhang et al., 2026). These materials resources serve information extraction and graph learning, peer-review benchmarks analyze post-submission review and acceptance processes, and our benchmark targets metadata-based journal triage before external review begins.

LLM Review Assistance Beyond benchmark construction, a growing body of work studies how LLMs can assist with scientific reviewing. Systems such as REVIEWROBOT, REVIEWERGPT,

and OPENREVIEWER generate or assist with scientific paper reviews (Wang et al., 2020; Liu and Shah, 2023; Idahl and Ahmadi, 2025). REVIEW-CRITIQUE, AGENTREVIEW, and TREEREVIEW examine review critique, simulated peer-review dynamics, and structured review generation (Du et al., 2024; Jin et al., 2024; Chang et al., 2025). Related work also studies LLM assistance for meta-reviewing (Hossain et al., 2025) and scientific paper edits made in response to peer reviews (D’Arcy et al., 2024). This literature primarily concerns the generation, analysis, or use of review feedback. Our benchmark instead considers the earlier journal-level decision of whether a manuscript should be sent for external review.

Review Quality and Robustness The growing use of LLMs in scientific reviewing has also raised questions about the reliability of their assessments. Studies identify aspects in peer reviews and detect weak or superficial reviewing (Lu et al., 2025; Purkayastha et al., 2025). Other work examines review utility for authors, the grounding of LLM-generated critiques, and imbalanced attention across manuscript content (Sadallah et al., 2025; Ou et al., 2025; Shin et al., 2025). Related studies introduce structured critical-assessment protocols or expose adversarial brittleness in automated reviewing (Dycke et al., 2025; Lin et al., 2025). Recent reviews synthesize broader reliability, security, and robustness concerns in LLM-based scientific review (Nguyen and Ahmadi, 2026; Wang et al., 2026), while model evaluations show that fluent outputs do not imply reliable paper assessment (Zhou et al., 2024). These findings motivate our controlled tests of keyword noise, synthetic degradation, and prompt-prior sensitivity.

Journal Recommendation Transfer recommendation connects our third task to prior work on manuscript–journal matching. JANE retrieves candidate journals through manuscript similarity (Schuemie and Kors, 2008), while CNAVER and HASVRec introduce learning-based recommendation methods (Pradhan and Pal, 2020; Pradhan et al., 2020). More recent work considers explainable journal recommendation (de Campos et al., 2024). Scientific text and document representation work includes SciBERT (Beltagy et al., 2019) and SPECTER (Cohan et al., 2020), while SciRepEval provides a multi-format benchmark for evaluating such representations (Singh et al., 2023). Reviewer recommendation and paper–reviewer matching ad-

dress a related allocation problem but target reviewers rather than publication venues (Anjum et al., 2019; Thorn Jakobsen and Rogers, 2022; Karan et al., 2025). Existing journal recommenders generally treat venue selection as a standalone matching problem; in our setting, transfer ranking follows and is conditioned on a desk-reject decision.

Together, these lines address individual components of scientific reviewing, but none evaluates the coupled decision an editor faces: whether to send, why, and where else. PREREVIEWBENCH casts these three questions as desk triage, branch-specific assessment, and transfer-journal recommendation in one controlled benchmark.

3 PREREVIEWBENCH

PREREVIEWBENCH tests two parts of journal pre-review: topical fit and journal caliber within a field. It contains four controlled scenarios that cover both send and reject decisions. These scenarios support three tasks: desk triage, branch-specific assessment, and transfer-journal recommendation after rejection. Figure 1 shows how we align journal pools and construct paper–journal pairs.

3.1 Data Sources and Scope

We build PREREVIEWBENCH from the metadata of journal articles published in 2025 and indexed in Web of Science (WoS).¹ It covers four broad areas: Biology, Chemistry, Materials Science, and Medicine. We select two WoS subfields from each area, giving eight subfields in total.

We keep WoS subfields that also have a Google Scholar Top Publications listing.² For each subfield, the candidate venue pool contains only journals listed in both systems. A paper instance includes its title, final indexed abstract, and keywords. These are public, publication-time records, not confidential abstracts from the original submission. A journal profile includes its name, Journal Impact Factor (JIF), Journal Citation Reports (JCR) quartile, Google Scholar h5-index, and official aims & scope. We use only papers published in 2025 to reduce variation over time.

3.2 Journal Stratification

We define the Scholar Synergy Index (SSI) to measure relative journal caliber within each subfield.

¹<https://clarivate.com/webofsciencegroup/solutions/web-of-science/>

²https://scholar.google.com/citations?view_op=top_venues&hl=en

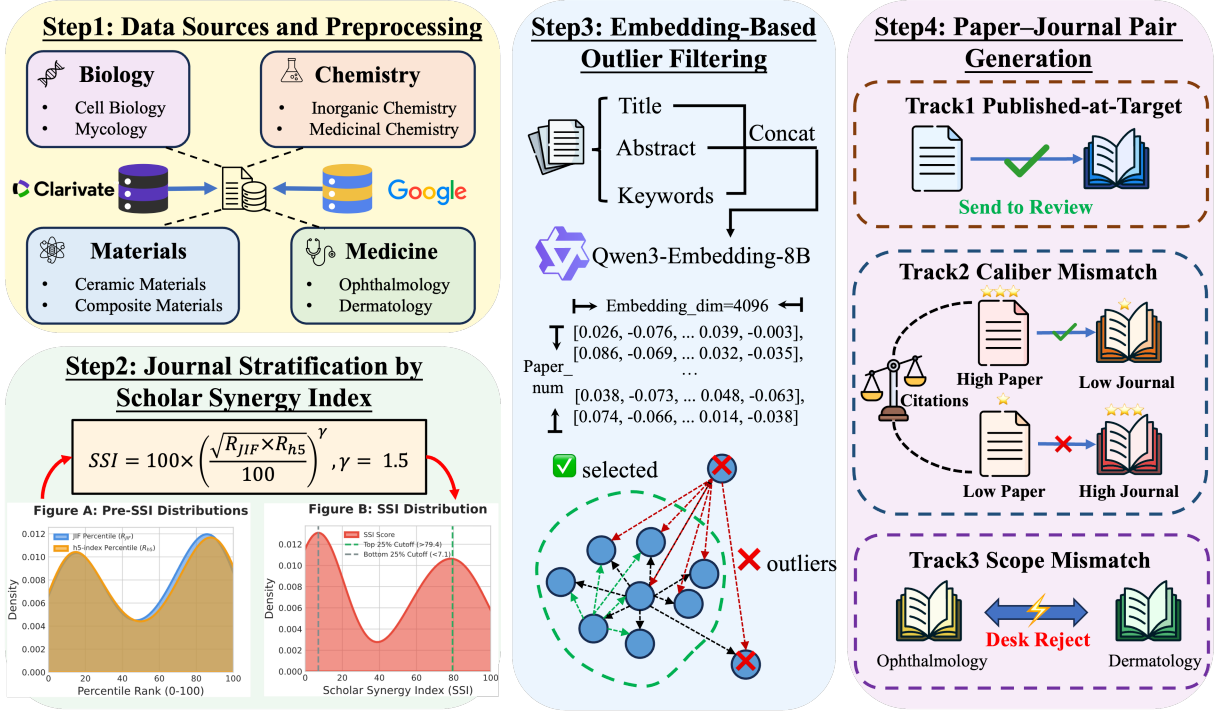


Figure 1: Construction pipeline for PREREVIEWBENCH. After aligning Web of Science and Google Scholar journal pools, we stratify journals by the Scholar Synergy Index (SSI) and filter topically isolated journals using a median-absolute-deviation (MAD) rule over embedding distances. Three tracks then produce four controlled proxy scenarios: Published-at-Target, Caliber-Surplus, Caliber-Deficit, and Scope-Mismatch.

We set $\gamma = 1.5$:

$$SSI = 100 \times \left(\frac{\sqrt{R_{JIF} \times R_{h5}}}{100} \right)^\gamma, \quad (1)$$

here, R_{JIF} is the within-subfield percentile rank of JIF and R_{h5} is the within-subfield percentile rank of the Google Scholar h5-index. Percentiles are computed within each subfield as

$$R_i = \left(1 - \frac{\text{rank}_i - 1}{N} \right) \times 100. \quad (2)$$

SSI is a proxy for journal caliber within a subfield. We retain only the top and bottom 25% of journals in each subfield. The geometric mean penalizes a journal that is strong on only one citation signal. The exponent γ increases the gap between higher- and lower-selectivity venues when we construct pairs.

3.3 Embedding-based Outlier Filtering

Before constructing paper-journal pairs, we remove journals that are topically isolated within their subfield. We use qwen3-embedding-8b (Zhang et al., 2025) to embed each paper from its title, keywords, and abstract. For each journal j , we average its paper embeddings and L2-normalize the result to obtain centroid v_j . We then score the

journal by its median cosine distance from other journals in the same subfield:

$$s_j = \text{median}_{k \in C, k \neq j} (1 - \cos(v_j, v_k)). \quad (3)$$

Within each subfield, we compute a median absolute deviation (MAD) z-score. If the subfield has at least four journals, we remove a journal when its score exceeds 3.0. This step makes within-subfield pairs more comparable in topic. It does not assign scope or caliber labels to individual papers.

3.4 Instance Construction

We organize paper-journal pair generation into three construction tracks, as shown in Figure 1. Track 1 preserves the published paper-journal pair. Track 2 creates two directions of within-subfield caliber mismatch. Track 3 creates a cross-subfield scope mismatch. Together, the three tracks produce four controlled scenarios: **Published-at-Target**, **Caliber-Surplus**, **Caliber-Deficit**, and **Scope-Mismatch**. Published-at-Target and Caliber-Surplus are constructed send cases. Caliber-Deficit and Scope-Mismatch are constructed desk-reject cases. These labels are controlled proxies for plausible editorial decisions, not records of actual desk decisions.

Track 1: Published-at-Target These instances pair each paper with the journal that published it.

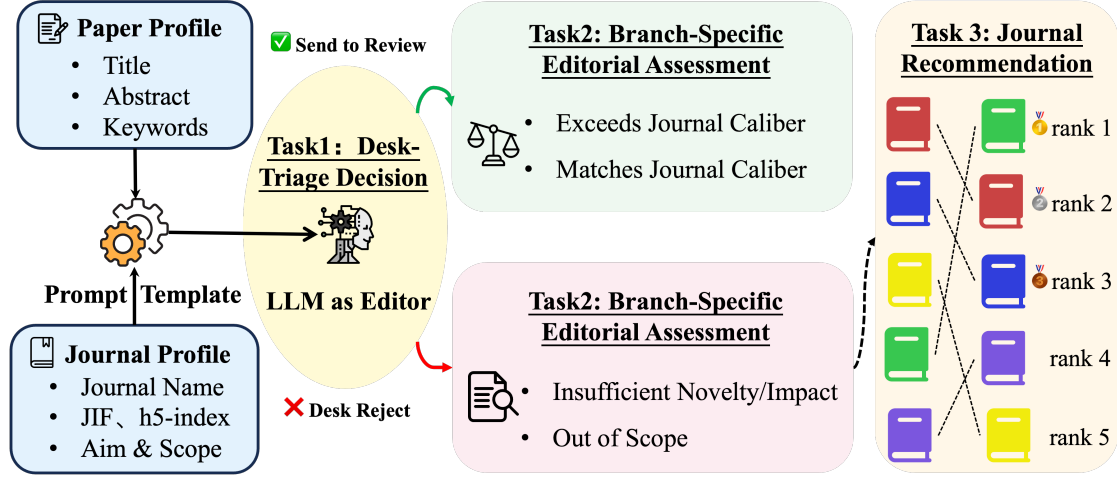


Figure 2: Three editor-facing tasks in PREREVIEWBENCH. The model first makes a desk-triage decision from paper metadata and the target-journal profile. It then predicts the branch-specific label associated with the send or reject branch. For constructed reject cases, it additionally ranks candidate transfer journals. Branch metrics are evaluated on the corresponding gold branch, and transfer metrics on gold reject cases.

They receive the label `Send for Review` and the branch label `Matches Journal Caliber`. They serve as proxies for papers that passed desk triage at their true venues.

Track 2: Caliber Mismatch This track uses each high- and low-SSI journal pair to construct two opposite directions of caliber mismatch: (1) **Caliber-Surplus Instances** pair a paper from a high-SSI journal with a lower-SSI target in the same subfield. We use citation counts to make this proxy contrast clearer. We start with the top 25% most cited papers from the high-SSI journal and the bottom 25% from the low-SSI journal. We then narrow both percentile bands in steps of 0.05. We stop when every paper in the retained high set has more citations than every paper in the retained low set. The instances receive the labels `Send for Review` and `Exceeds Journal Caliber`. (2) **Caliber-Deficit Instances** reverse this direction. A paper from the low-SSI journal is paired with the high-SSI target in the same pair. The instances receive the labels `Desk Reject` and `Insufficient Novelty/Impact`.

Track 3: Scope Mismatch These instances keep the broad area fixed but move a paper to the other WoS subfield in that area. We choose the target journal in the other subfield with the closest SSI. The instances receive the labels `Desk Reject` and `Out of Scope`. Transfer candidates come from the union of the two paired subfields.

Within each source subfield, we sample Scope-Mismatch cases to match the number of Caliber-Deficit cases. The final benchmark is exactly balanced across the four instance types.

3.5 Benchmark Tasks

Each instance supports three editor-facing tasks. The model must decide desk triage before seeing the candidate journal list. This order prevents transfer options from revealing the triage label. Figure 2 shows the full sequence.

Task 1: Desk-Triage Decision The model receives the manuscript metadata and target-journal profile. It predicts `Send for Review` or `Desk Reject` and gives a confidence score from 1 to 10. Published-at-Target and Caliber-Surplus instances are send cases; Caliber-Deficit and Scope-Mismatch instances are reject cases.

Task 2: Branch-Specific Editorial Assessment The model then selects a controlled label for the relevant editorial branch. For a send decision, it predicts `Matches Journal Caliber` or `Exceeds Journal Caliber`. For a reject decision, it predicts `Insufficient Novelty/Impact` or `Out of Scope`. This task tests whether the triage decision rests on the correct caliber or scope judgment.

Task 3: Transfer-Journal Recommendation For constructed desk-reject cases, the model ranks candidate journals for transfer. The realized publication venue is the primary target. Candidates come from the source subfield for Caliber-Deficit cases and from the union of the paired subfields for Scope-Mismatch cases.

3.6 Data Statistics and Quality

Dataset Statistics PREREVIEWBENCH contains 7,292 instances from 58 source journals and eight WoS subfields (Table 1). Each of the four instance

Broad Area	Source Subfield	Instances		Candidate Journals	
		Per Scenario	Total	Within	Paired Union
Biology	Cell Biology	468	1,872	10	15
	Mycology	69	276	5	15
Chemistry	Inorganic Chemistry	234	936	7	14
	Medicinal Chemistry	237	948	7	14
Materials Science	Ceramics	189	756	9	16
	Composites	297	1,188	7	16
Medicine	Dermatology	121	484	7	13
	Ophthalmology	208	832	6	13
Total/Range	8 Subfields	1,823	7,292	5–10	13–16

Table 1: Composition of PREREVIEWBENCH across 58 journals, four broad areas, and eight WoS source subfields. Each subfield contributes the same number of instances to each of the four constructed scenarios. Within denotes the source-subfield candidate pool, while Paired union denotes the combined pool used for Scope-Mismatch transfer ranking.

Scenario	<i>N</i>	Triage	Branch cond.	Joint
Published-at-Target	40	100.00	100.00	100.00
Caliber-Surplus	60	93.33	98.21	91.67
Caliber-Deficit	60	100.00	96.67	96.67
Scope-Mismatch	40	100.00	100.00	100.00
Overall	200	98.00	98.47	96.50

Table 2: Annotator agreement (%) with constructed labels. Branch-specific agreement is conditional on agreement with the desk-triage label; Joint requires both labels to agree.

types has 1,823 cases. The dataset also has equal numbers of send and reject cases (3,646 each). Candidate lists contain 5–16 journals, depending on the subfield pair. An abstract has 203.8 words on average, and a paper has 8.8 keywords on average.

Expert Validation We evaluate the constructed labels through a blind, stratified review of 200 cases. Table 2 reports agreement between the human annotations and the constructed labels. The review was led by four domain experts outside the author team, one for each broad area. Each expert participated directly in annotation and coordinated graduate students from the same field who also took part in labeling. Each area team reviewed 50 cases using the same title, abstract, keywords, and target-journal profile shown to the models. The annotators did not see the constructed labels, source venue, or model outputs. The sample covers all eight subfields and gives more weight to the harder caliber cases: 40 Published-at-Target, 60 Caliber-Surplus, 60 Caliber-Deficit, and 40 Scope-Mismatch cases.

Annotators agreed with 196 of 200 desk-triage labels (98.00%). Conditional agreement on the branch label was 193 of 196 (98.47%), and joint agreement was 193 of 200 (96.50%). Joint agreement by broad area ranged from 90% to 100%.

All four triage disagreements were Caliber-Surplus cases judged out of scope: two in Chemistry and two in Medicine. Among the agreed triage cases, one Chemistry surplus case was judged Matches instead of Exceeds. Two Chemistry deficit cases were judged Out of Scope instead of Insufficient Novelty/Impact. The results support the constructed distinctions but also show some ambiguity near their boundaries. They do not make the editorial labels objective or noise-free. Appendix Figure 3 presents a representative case at the boundary between within-field caliber transfer and journal-specific scope.

4 Experiments

4.1 Experimental Setup

Benchmark and Inputs The main experiment uses all 7,292 PREREVIEWBENCH instances, balanced across the four scenarios, with the full journal-context condition. Each model receives only the manuscript title, abstract, and keywords together with the target journal’s name, JIF, JCR quartile, h5-index, and aims & scope. Models never see source venues, realized venues, SSI, or constructed labels. For constructed reject cases, the model receives a candidate-journal list for transfer ranking only after forming its desk-triage decision, following the task protocol in Section 3.5.

Evaluation For desk triage, we report accuracy, Macro-F1, balanced accuracy, false-reject rate, and false-send rate. For branch-specific assessment, we report send-side quality Macro-F1 (Q-F1), reject-side cause Macro-F1 (R-F1), invalid-prediction rate, and joint editorial correctness. Joint correctness requires both the triage and branch labels to

Model	Prompt	Triage Decision				Branch Assessment			Transfer Ranking	
		Acc.	F1	F-Rej.	F-Send	Q-F1	R-F1	Joint	MRR	NDCG@5
GPT-5.5	Direct	0.828	0.828	<u>0.123</u>	0.220	0.588	0.782	0.649	<u>0.523</u>	0.576
	Rationale-first	<u>0.827</u>	<u>0.826</u>	0.129	0.217	<u>0.562</u>	<u>0.777</u>	<u>0.638</u>	0.524	0.576
DeepSeek-V4-Pro	Direct	0.808	0.809	0.191	<u>0.191</u>	0.371	0.747	0.566	0.465	0.526
	Rationale-first	0.807	0.807	0.206	0.179	0.354	0.744	0.560	0.473	<u>0.532</u>
Claude-4.7-Opus	Direct	0.772	0.775	0.211	0.233	0.495	0.744	0.594	0.447	0.492
	Rationale-first	0.779	0.781	0.213	0.217	0.452	0.731	0.578	0.460	0.504
Qwen3.6-Plus	Direct	0.790	0.788	0.119	0.301	0.330	0.675	0.532	0.422	0.470
	Rationale-first	0.789	0.789	0.145	0.270	0.327	0.697	0.535	0.437	0.486

Table 3: Overall benchmark results across the eight main model–prompt configurations. Branch assessment is split into quality prediction on gold send cases (Q-F1) and rejection-cause prediction on gold reject cases (R-F1). Joint requires both the correct desk-triage decision and the correct branch label. Transfer ranking uses candidate-projected MRR and NDCG@5. Best and second-best comparable results are shown in bold and underlined, respectively; lower is better for F-Rej. and F-Send, and higher is better otherwise.

be correct. Transfer recommendation is evaluated only on constructed reject cases using candidate-projected MRR, Hit@ k , and NDCG@ k against the realized publication venue. Appendix A gives the complete metric definitions.

Models and Inference We evaluate four frontier LLM families: GPT-5.5 (OpenAI, 2026), Claude-4.7-Opus (Anthropic, 2026), Qwen3.6-Plus (Qwen Team, 2026), and DeepSeek-V4-Pro (DeepSeek, 2026). Each model is tested with two prompt modes on all three coupled tasks. *Direct* requests only the structured answer; *Rationale-first* first requests a concise 2–4 sentence editorial justification and then the same structured answer. This is a prompt-level verbalized-justification condition, not an open-ended reasoning trace. We set decoding temperature to 0 and report one finalized run per model–prompt configuration. We do not enable or tune provider-specific deliberation controls such as reasoning effort, extended thinking, or thinking budgets; each provider endpoint therefore uses its default internal-deliberation behavior. Across the main benchmark, journal-context ablations, two robustness stress tests, and the GPT-specific send-prior intervention, we report 42 finalized runs. Exact prompt templates are provided in Appendix B.

4.2 Experimental Results

The results are organized into the main benchmark, diagnostic analyses, and controlled ablation and robustness experiments.

4.2.1 Main Benchmark Results

Overall Performance and Decision Tendencies Table 3 compares the eight main configurations. GPT-5.5 is strongest overall, but even its

Group	Slice	Triage	Joint	MRR
Scenario avg.	Pub.-Target	0.849	0.843	–
	Cal.-Surp.	0.814	0.124	–
	Cal.-Def.	0.572	0.414	0.347
	Scope-Mis.	0.965	0.946	0.591
Broad-area avg.	Biology	0.878	0.671	0.503
	Chemistry	0.745	0.554	0.536
	Mat. Sci.	0.763	0.485	0.370
	Medicine	0.806	0.619	0.464

Table 4: Scenario-wise and broad-area summaries, macro-averaged over the eight main model–prompt configurations. MRR is undefined for the two all-send scenarios. These slices characterize variation in task difficulty rather than rank competing systems.

Direct setting reaches only 0.649 joint correctness, indicating substantial error in the coupled triage-and-assessment workflow. It has the best triage accuracy (0.828) and send-side quality F1, while *Rationale-first* is effectively tied on triage and transfer MRR (0.524 vs. 0.523). The requested short justification therefore does not consistently improve performance. Error direction also differs by family: DeepSeek-V4-Pro has the lowest false-send rate, Claude-4.7-Opus the highest false-reject rate, and Qwen3.6-Plus the highest false-send rate. We describe these as family-specific decision tendencies rather than editorial competence.

Scope Screening versus Caliber Judgment Table 4 exposes the benchmark’s central capability gap. Averaged over the eight main configurations, triage accuracy is 0.965 on cross-field Scope-Mismatch but only 0.572 on within-field Caliber-Deficit. The latter also has the lowest joint correctness and transfer MRR. This is not merely an artifact of observed disagreement in the audited subset: annotators agreed with all 60 Caliber-Deficit triage labels in Table 2. Current models therefore

System	Triage	Q-F1	R-F1	Joint
Logistic regression	0.706	0.549	0.635	0.527
GPT-5.5 Direct	0.828	0.588	0.782	0.649
Transfer system	MRR		NDCG@5	
Jaccard	0.369		0.402	
BM25	0.378		0.416	
SciBERT	0.403		0.449	
SPECTER	<u>0.430</u>		<u>0.475</u>	
GPT-5.5 Direct	0.523		0.576	
Target view	MRR	Hit@1	Soft-NDCG@5	
Exact venue	0.523	0.369	0.576	
Metadata multi-gold	0.694	0.636	0.602	

Table 5: Baselines and transfer-target sensitivity. The upper block compares desk triage and branch assessment; the middle block compares candidate-projected transfer rankers. Bold and underlining mark the best and second-best comparable systems within these blocks. The lower block re-evaluates the same GPT-5.5 *Direct* outputs under a different, metadata-based relevance definition and is therefore not ranked against the exact-venue view. Standalone rankers are not full coupled systems.

screen clear scope mismatches much more reliably than they judge whether an in-scope paper clears a journal’s caliber bar. Difficulty also varies by area: Biology has the strongest triage and joint results, while Materials Science has the lowest transfer MRR (0.370). Appendix Tables 15 and 16 give the complete scenario results; Tables 13 and 14 give the per-model area breakdowns.

4.2.2 Analysis Experiments

Baselines and Transfer-Target Sensitivity Table 5 compares GPT-5.5 *Direct* with supervised and retrieval baselines. Logistic regression uses grouped five-fold cross-validation by paper with target-profile similarity, target-journal metrics, and basic manuscript statistics; it excludes the source venue, realized venue, SSI, and transfer options. Jaccard and BM25 rank full journal profiles, SciBERT mean-pools 512-token profiles, and SPECTER uses a 512-token scope representation. GPT-5.5 exceeds logistic regression by 12.2 points on both triage and joint correctness and exceeds SPECTER by 0.093 MRR. However, logistic regression surpasses the other three LLM families on Q-F1, so the gains cannot be attributed solely to LLM reasoning. The lower block tests target sensitivity: crediting same-subfield, same-SSI-stratum alternatives raises GPT-5.5 MRR from 0.523 to 0.694 and Hit@1 from 0.369 to 0.636. Thus, even the best exact-venue top-1 result is

Model	Prompt	Output		Transfer MRR		
		Out.	Comp.	Proj.	Strict	Valid
GPT-5.5	Direct	0.780	0.779	<u>0.523</u>	<u>0.523</u>	0.671
GPT-5.5	Rat.-first	0.783	0.783	0.524	0.524	<u>0.669</u>
DeepSeek	Direct	<u>0.808</u>	0.648	0.465	0.373	0.576
DeepSeek	Rat.-first	0.820	0.628	0.473	0.363	0.579
Claude	Direct	0.759	0.759	0.447	0.447	0.588
Claude	Rat.-first	0.776	0.776	0.460	0.460	0.593
Qwen	Direct	0.699	0.448	0.422	0.269	0.601
Qwen	Rat.-first	0.724	0.435	0.437	0.261	0.600

Table 6: Transfer-output diagnostics for the eight main runs. Out. is the nonempty-ranking rate, and Comp. is the complete-permutation rate. Proj. removes out-of-candidate entries while preserving order; Strict assigns zero to invalid permutations; Valid conditions on complete permutations. Best and second-best results in each column are shown in bold and underlined, respectively.

only 0.369, while some apparent misses remain metadata-compatible alternatives. These alternatives are metadata-derived rather than expert-confirmed, so the two views diagnose strict-target sensitivity rather than establish transfer suitability or a model improvement.

Transfer Quality and Output Compliance Table 6 separates ranking quality from format reliability. GPT-5.5 and Claude produce complete permutations for nearly all nonempty rankings, and their projected and strict MRR are identical to three decimals. DeepSeek-V4-Pro and especially Qwen3.6-Plus often return incomplete permutations: for Qwen *Direct*, projected MRR is 0.422 but strict MRR is 0.269, while valid-only MRR reaches 0.601. Transfer failure therefore reflects both candidate ordering and output compliance; treating the two as one error source would understate the useful signal in partial rankings.

4.2.3 Ablation and Robustness Experiments

Journal-context Ablation Table 7 tests whether models use the supplied journal profile. The no_metrics condition removes JIF, quartile, and h5-index while retaining the journal name and aims & scope; only_name removes aims & scope and all numeric indicators. Removing metrics costs GPT-5.5 and DeepSeek-V4-Pro 1.0–3.1 triage points and 1.8–2.7 joint points. Full context also produces the best transfer MRR except for Claude *Direct*’s small no-metrics gain. Name-only losses are largest for GPT-5.5 and DeepSeek-V4-Pro, showing that richer journal context generally helps both triage and transfer, although the magnitude is model-dependent.

Model	Prompt	Full			No-Metrics			Name-Only		
		Acc.	Joint	MRR	Acc.	Joint	MRR	Acc.	Joint	MRR
GPT-5.5	Direct	0.828	0.649	0.523	0.818 $\downarrow$	0.623 $\downarrow$	0.508 $\downarrow$	0.808 $\downarrow$	0.613 $\downarrow$	0.465 $\downarrow$
	Rationale-first	0.827	0.638	0.524	0.817 $\downarrow$	0.615 $\downarrow$	0.503 $\downarrow$	0.803 $\downarrow$	0.595 $\downarrow$	0.460 $\downarrow$
DeepSeek-V4-Pro	Direct	0.808	0.566	0.465	0.789 $\downarrow$	0.548 $\downarrow$	0.437 $\downarrow$	0.778 $\downarrow$	0.545 $\downarrow$	0.409 $\downarrow$
	Rationale-first	0.807	0.560	0.473	0.776 $\downarrow$	0.534 $\downarrow$	0.425 $\downarrow$	0.776 $\downarrow$	0.536 $\downarrow$	0.409 $\downarrow$
Claude-4.7-Opus	Direct	0.772	0.594	0.447	0.774 $\uparrow$	0.570 $\downarrow$	0.457 $\uparrow$	0.770 $\downarrow$	0.557 $\downarrow$	0.409 $\downarrow$
	Rationale-first	0.779	0.578	0.460	0.772 $\downarrow$	0.562 $\downarrow$	0.450 $\downarrow$	0.770 $\downarrow$	0.556 $\downarrow$	0.422 $\downarrow$
Qwen3.6-Plus	Direct	0.790	0.532	0.422	0.771 $\downarrow$	0.514 $\downarrow$	0.391 $\downarrow$	0.770 $\downarrow$	0.517 $\downarrow$	0.382 $\downarrow$
	Rationale-first	0.789	0.535	0.437	0.775 $\downarrow$	0.519 $\downarrow$	0.397 $\downarrow$	0.769 $\downarrow$	0.519 $\downarrow$	0.389 $\downarrow$

Table 7: Journal-context ablations. Arrows indicate changes relative to the corresponding Full setting: upward arrows denote improvements and downward arrows denote degradations.

Model	Prompt	Keyword noise		Degradation	
		Clean	Noise	Reject	R-F1
GPT-5.5	Direct	<u>0.879</u>	0.849	0.508	0.647
	Rationale-first	0.871	0.828	0.532	0.665
DeepSeek	Direct	0.811	0.835	<u>0.547</u>	0.650
	Rationale-first	0.796	0.810	0.569	<u>0.661</u>
Claude	Direct	0.858	0.708	0.244	0.345
	Rationale-first	0.826	0.814	0.406	0.535
Qwen	Direct	0.890	0.853	0.431	0.563
	Rationale-first	0.861	0.833	0.419	0.551

Table 8: Robustness stress tests. Clean/Noise report send accuracy; Reject/R-F1 report performance on degraded cases. Best and second-best results are shown in bold and underlined.

Input Perturbations Table 8 reports two complementary stress tests on Published-at-Target papers. Keyword noise adds five hot-topic and cross-domain terms without changing the send label; degradation weakens innovation while preserving topic and target fit, changing the gold label to Desk Reject. Noise reduces accuracy in six of eight configurations, most sharply for Claude *Direct* (15.0 points), while DeepSeek improves slightly. Every model struggles with degradation (best reject accuracy: 0.569), confirming that within-scope caliber judgment is harder than scope matching. Appendix Figure 8 gives the direct-prompt breakdown.

Sensitivity to a Send-Prior Cue Table 9 tests an active prompt intervention that tells GPT-5.5 roughly 90% of the batch is plausibly sendable. The cue raises Published-at-Target accuracy from 0.879 to 0.961 under *Direct*, but lowers Caliber-Deficit accuracy from 0.603 to 0.351 and MRR from 0.523 to 0.430; ECE worsens from 0.028 to 0.071. *Rationale-first* shows the same shift, so the cue changes the operating point rather than underlying capability. Appendix Table 12 provides the fixed-prediction control and the full intervention

Mode	Setting	Acc.	Pub.	Cal.-Def.	MRR	ECE
Direct	Base	0.828	0.879	0.603	0.523	0.028
Direct	+Prior	0.787	0.961	0.351	0.430	0.071
Rationale-first	Base	0.827	0.871	0.605	0.524	0.021
Rationale-first	+Prior	0.787	0.962	0.357	0.433	0.069

Table 9: GPT-5.5 send-prior intervention. The cue states that roughly 90% of the batch is sendable. Pub. and Cal.-Def. are scenario-specific accuracies. The paired rows expose the trade-off between send-scenario accuracy, caliber-deficit detection, transfer quality, and ECE.

results; Appendix Figures 6–9 provide confidence diagnostics. Because stress-test curves can be non-monotonic and provider scales can drift, calibration is exploratory.

5 Conclusion

We presented PREREVIEWBENCH, the first controlled benchmark for journal desk screening: 7,292 balanced paper–journal instances built from Web of Science metadata by controlled pairing, with expert-validated proxy labels and three progressive tasks. Across four frontier LLM families, the benchmark exposes a consistent capability boundary. Models detect cross-field scope mismatches reliably (average triage accuracy 0.965) but falter on within-field caliber deficits (0.572), where blind expert-led review reproduces every triage label; transfer recommendation is weaker still, and stress tests reveal further fragility. These results support a clear division of labor: LLMs can already provide scope-level triage assistance, while caliber judgment stays with editors. We see journal-aware, human-AI editorial collaboration as the practical path, and within-field caliber judgment, now measurable, as the problem worth solving next.

Limitations

PREREVIEWBENCH evaluates constructed proxies rather than observed confidential desk decisions. Publication at a venue and citation-stratified pairings enable controlled comparisons, but publication does not reveal the original desk-triage decision. SSI is also a venue-level proxy, not a direct measure of manuscript quality or an editor’s threshold. The 200-case study shows high agreement between the constructed labels and field-matched annotations. However, each case was reviewed by one annotator, the sample covers only the selected areas, and disagreements cluster near scope and caliber boundaries. Real desk decisions may also depend on full-text evidence, reviewer availability, recent submissions, portfolio considerations, special issues, and editor-specific judgment, which the benchmark does not observe.

The benchmark uses final indexed abstracts rather than confidential submission versions. It covers only 2025 records from four broad areas, eight WoS subfields, and 58 journals. The findings may therefore not generalize to other fields, years, editorial cultures, or future model versions. We retain one finalized temperature-0 output per configuration and do not estimate run-to-run or API-version variance. Post-stratification illustrates sensitivity to class prevalence but does not identify a deployment threshold. The prompt-prior intervention changes model predictions and can reduce sensitivity on difficult reject cases. Closed-book probes weaken a simple memorization explanation but cannot rule out overlap with model training data.

Transfer evaluation treats the realized publication venue as the primary target, although several destinations may be reasonable. The metadata-based multi-gold analysis reduces this single-target penalty, but its alternatives remain weak labels rather than editor-validated recommendations. Strict and candidate-projected metrics also measure different mixtures of ranking quality and format compliance. Finally, the requested 1–10 confidence score is a constrained diagnostic signal, not a calibrated probability.

Ethics Statement

This benchmark is intended for decision support and auditing, not autonomous editorial replacement. Automated screening may amplify topical, disciplinary, prestige-related, linguistic, or demographic bias and may homogenize journal portfolios. False

rejects can deny authors access to review, while false sends consume scarce reviewer labor. Transfer suggestions may also reinforce existing venue hierarchies. Any use therefore requires accountable human decisions, appeal paths, subgroup and drift audits, and clear disclosure that model outputs are advisory. The released inputs are publication-time records rather than confidential submissions, which reduces confidentiality risk relative to releasing full manuscripts.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants T2322027 and 62442204. We thank Zhengyin Hu and Zihao Zhang for their assistance with data preparation. We also thank the domain experts for their valuable annotation work.

References

- Balazs Aczel, Barnabas Szasz, and Alex O. Holcombe. 2021. [A billion-dollar donation: estimating the cost of researchers’ time spent on peer review](#). *Research Integrity and Peer Review*, 6:14.
- Omer Anjum, Hongyu Gong, Suma Bhat, Wen-Mei Hwu, and JinJun Xiong. 2019. [PaRe: A paper-reviewer matching approach using a common topic space](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 518–528, Hong Kong, China. Association for Computational Linguistics.
- Anthropic. 2026. Introducing Claude Opus 4.7. <https://www.anthropic.com/news/claude-opus-4-7>. Official model announcement and documentation.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. [SciBERT: A pretrained language model for scientific text](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.
- Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. 2025. [TreeReview: A dynamic tree of questions framework for deep and efficient LLM-based scientific peer review](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15651–15682, Suzhou, China. Association for Computational Linguistics.
- Xueqing Chen, Yang Gao, Ludi Wang, Wenjuan Cui, Jiamin Huang, Yi Du, and Bin Wang. 2024. [Large language model enhanced corpus of CO₂ reduction](#)

- electrocatalysts and synthesis procedures. *Scientific Data*, 11:347.
- Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. **SPECTER: Document-level representation learning using citation-informed transformers**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.
- Mike D’Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2024. **ARIES: A corpus of scientific paper edits made in response to peer reviews**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6985–7001, Bangkok, Thailand. Association for Computational Linguistics.
- Luis M. de Campos, Juan M. Fernández-Luna, and Juan F. Huete. 2024. **An explainable content-based approach for recommender systems: a case study in journal recommendation for paper submission**. *User Modeling and User-Adapted Interaction*, 34:1431–1465.
- DeepSeek. 2026. DeepSeek V4 preview release. <https://api-docs.deepseek.com/news/news260424>. Official release note for the public DeepSeek V4 family, including V4-Pro.
- Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, and 21 others. 2024. **LLMs assist NLP researchers: Critique paper (meta-)reviewing**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099, Miami, Florida, USA. Association for Computational Linguistics.
- Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. **NLPeer: A unified resource for the computational study of peer review**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.
- Nils Dycke, Edwin Simpson, Ilia Kuznetsov, and Iryna Gurevych. 2021. **Assisting decision making in scholarly peer review: A preference learning perspective**. *arXiv preprint arXiv:2109.01190*.
- Nils Dycke, Matej Zečević, Ilia Kuznetsov, Beatrix Suess, Kristian Kersting, and Iryna Gurevych. 2025. **STRICTA: Structured reasoning in critical text assessment for peer review and beyond**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22687–22727, Vienna, Austria. Association for Computational Linguistics.
- James C. Garand and Moriah Harman. 2021. **Journal desk-rejection practices in political science: Bringing data to bear on what journals do**. *PS: Political Science & Politics*, 54(4):676–681.
- Eftekhari Hossain, Sanjeev Kumar Sinha, Naman Bansal, R. Alexander Knipper, Souvika Sarkar, John Salvador, Yash Mahajan, Sri Ram Pavan Kumar Guttikonda, Mousumi Akter, Md. Mahadi Hassan, Matthew Freestone, Matthew C. Williams Jr., Dongji Feng, and Santu Karmaker. 2025. **LLMs as meta-reviewers’ assistants: A case study**. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7763–7803, Albuquerque, New Mexico. Association for Computational Linguistics.
- Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. **Argument mining for understanding peer reviews**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2131–2137, Minneapolis, Minnesota. Association for Computational Linguistics.
- Maximilian Idahl and Zahra Ahmadi. 2025. **OpenReviewer: A specialized large language model for generating critical scientific paper reviews**. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 550–562, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. **AgentReview: Exploring peer review dynamics with LLM agents**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1208–1226, Miami, Florida, USA. Association for Computational Linguistics.
- Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. **A dataset of peer reviews (PeerRead): Collection, insights and NLP applications**. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.
- Vanja M. Karan, Stephen McQuistin, Ryo Yanagida, Colin Perkins, Gareth Tyson, Ignacio Castro, Patrick G.T. Healey, and Matthew Purver. 2025. **A dataset for expert reviewer recommendation with large language models as zero-shot rankers**. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11422–11427, Abu Dhabi, UAE. Association for Computational Linguistics.
- Jiyi Li, Ayaka Sato, Kazuya Shimura, and Fumiyo Fukumoto. 2020. **Multi-task peer-review score prediction**. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 121–126, Online. Association for Computational Linguistics.
- 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. 2025. **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*, pages 4819–4839, Suzhou, China. Association for Computational Linguistics.
- Ryan Liu and Nihar B. Shah. 2023. [ReviewerGPT? an exploratory study on using large language models for paper reviewing](#). *arXiv preprint arXiv:2306.00622*.
- Sheng Lu, Iliia Kuznetsov, and Iryna Gurevych. 2025. [Identifying aspects in peer reviews](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6145–6167, Suzhou, China. Association for Computational Linguistics.
- Thi Huyen Nguyen and Zahra Ahmadi. 2026. [LLM-based scientific peer review: Methods, benchmarks, and reliability challenges](#). *arXiv preprint arXiv:2606.25057*.
- OpenAI. 2026. GPT-5.5 system card. <https://openai.com/index/gpt-5-5-system-card/>. Model system card and product documentation.
- Jiefu Ou, William Walden, Kate Sanders, Zhengping Jiang, Kaiser Sun, Jeffrey Cheng, William Jurayj, Miriam Wanner, Shaobo Liang, Candice Morgan, Seunghoon Han, Weiqi Wang, Chandler May, Hannah Recknor, Daniel Khashabi, and Benjamin Van Durme. 2025. [CLAIMCHECK: How grounded are LLM critiques of scientific papers?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21712–21735, Suzhou, China. Association for Computational Linguistics.
- Tribikram Pradhan, Abhinav Gupta, and Sukomal Pal. 2020. [HASVRec: A modularized hierarchical attention-based scholarly venue recommender system](#). *Knowledge-Based Systems*, 204:106181.
- Tribikram Pradhan and Sukomal Pal. 2020. [CNAVER: A content and network-based academic venue recommender system](#). *Knowledge-Based Systems*, 189:105092.
- Richard B. Primack, Tracey J. Regan, Vincent Devictor, Lucy Zipf, Laurent Godet, Rafael Loyola, Bea Maas, Robin J. Pakeman, Graeme S. Cumming, Amanda E. Bates, Liba Pejchar, and Lian Pin Koh. 2019. [Are scientific editors reliable gatekeepers of the publication process?](#) *Biological Conservation*, 238:108232.
- Sukannya Purkayastha, Zhuang Li, Anne Lauscher, Lizhen Qu, and Iryna Gurevych. 2025. [LazyReview: A dataset for uncovering lazy thinking in NLP peer reviews](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3280–3308, Vienna, Austria. Association for Computational Linguistics.
- Qwen Team. 2026. [Qwen 3.6 Plus model launch](#). Qwen Code documentation.
- Abdelrahman Sadallah, Tim Baumgärtner, Iryna Gurevych, and Ted Briscoe. 2025. [The good, the bad and the constructive: Automatically measuring peer review’s utility for authors](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28991–29021, Suzhou, China. Association for Computational Linguistics.
- Martijn J. Schuemie and Jan A. Kors. 2008. [Jane: suggesting journals, finding experts](#). *Bioinformatics*, 24(5):727–728.
- Hyungyu Shin, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung Hong, Moontae Lee, and Juho Kim. 2025. [Mind the blind spots: A focus-level evaluation framework for LLM reviews](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35630–35656, Suzhou, China. Association for Computational Linguistics.
- Amanpreet Singh, Mike D’Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. [SciRepEval: A multi-format benchmark for scientific document representations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566, Singapore. Association for Computational Linguistics.
- Moritz Staudinger, Wojciech Kusa, Florina Piroi, and Allan Hanbury. 2024. [An analysis of tasks and datasets in peer reviewing](#). In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 257–268, Bangkok, Thailand. Association for Computational Linguistics.
- Terne Thorn Jakobsen and Anna Rogers. 2022. [What factors should paper-reviewer assignments rely on? community perspectives on issues and ideals in conference peer-review](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4810–4823, Seattle, United States. Association for Computational Linguistics.
- Jialiang Wang, Yuchen Liu, Hang Xu, Kaichun Hu, Shimin Di, Wangze Ni, Linan Yue, Min-Ling Zhang, Kui Ren, and Lei Chen. 2026. [When AI reviews science: Can we trust the referee?](#) *The Innovation Informatics*, 2(1):100030.
- Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. [ReviewRobot: Explainable paper review generation based on knowledge synthesis](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 384–397, Dublin, Ireland. Association for Computational Linguistics.
- Chenghao Zhang, Qingqing Long, Ludi Wang, Wenjuan Cui, Jianjun Yu, and Yi Du. 2026. [CITE: Benchmarking heterogeneous text-attributed graph models](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31426–31448, San Diego, California, United States. Association for Computational Linguistics.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. [Qwen3 Embedding: Advancing text embedding and reranking through foundation models](#). *arXiv preprint arXiv:2506.05176*.
- Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. [Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, Torino, Italia. ELRA and ICCL.

A Metric Definitions

This appendix reports the exact metric definitions used by the released evaluator. Let N be the number of evaluated instances, y_i the gold label, and $\hat{y}_i$ the model prediction for instance i .

Task 1: Desk-Triage Decision Metrics For the binary send/reject decision, accuracy (Acc.) is

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\hat{y}_i = y_i]. \quad (4)$$

For each class $c \in \{\text{Send}, \text{Reject}\}$, precision, recall, and the F1 score (F1) are

$$\begin{aligned} P_c &= \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}, \\ R_c &= \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}, \\ F1_c &= \frac{2P_c R_c}{P_c + R_c}, \end{aligned} \quad (5)$$

where TP_c , FP_c , and FN_c denote true positives, false positives, and false negatives for class c , respectively. The macro-averaged F1 score (Macro-F1) is

$$\text{Macro-F1} = \frac{1}{2} \sum_c F1_c. \quad (6)$$

Balanced accuracy (Bal. Acc.) averages recall across the two decision classes:

$$\text{BalancedAcc} = \frac{R_{\text{Send}} + R_{\text{Reject}}}{2}. \quad (7)$$

The false-reject rate (F-Rej.) and false-send rate (F-Send) are branch-specific editorial errors:

$$\text{FalseReject} = \frac{\sum_i \mathbf{1}[y_i = \text{Send} \wedge \hat{y}_i = \text{Reject}]}{\sum_i \mathbf{1}[y_i = \text{Send}]}, \quad (8)$$

$$\text{FalseSend} = \frac{\sum_i \mathbf{1}[y_i = \text{Reject} \wedge \hat{y}_i = \text{Send}]}{\sum_i \mathbf{1}[y_i = \text{Reject}]}. \quad (9)$$

The evaluator additionally reports the parsed-decision rate (ParsedRate),

$$\text{ParsedRate} = \frac{\sum_i \mathbf{1}[\hat{y}_i \in \{\text{Send}, \text{Reject}\}]}{N}. \quad (10)$$

This rate is reported separately to distinguish decision quality from schema-compliance failures.

Task 2: Branch-Specific Editorial Assessment Metrics The branch-specific editorial assessment is evaluated conditionally on the gold desk-triage branch. On gold send cases, the label space is $\{\text{Matches}, \text{Exceeds}\}$; on gold reject cases, it

is $\{\text{Novelty}, \text{Scope}\}$. We refer to send-branch Macro-F1 as quality F1 (Q-F1) and reject-branch Macro-F1 as rejection-cause F1 (R-F1). For either branch b ,

$$\text{Accuracy}_b = \frac{1}{N_b} \sum_{i=1}^{N_b} \mathbf{1}[\hat{z}_i = z_i], \quad (11)$$

$$\text{Macro-F1}_b = \frac{1}{|L_b|} \sum_{\ell \in L_b} F1_\ell, \quad (12)$$

where z_i and $\hat{z}_i$ are the gold and predicted branch labels and L_b is the admissible label set for branch b . Invalid-prediction rate is

$$\text{InvalidRate}_b = \frac{\sum_i \mathbf{1}[\hat{z}_i \notin L_b]}{N_b}. \quad (13)$$

Joint editorial correctness (Joint) measures end-to-end usefulness:

$$\text{Joint} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\hat{y}_i = y_i \wedge \hat{z}_i = z_i]. \quad (14)$$

Task 3: Transfer-Journal Recommendation Metrics

Transfer-journal recommendation is evaluated only on gold reject cases. Let g_i be the realized publication venue used as the single gold target, r_i the predicted ranking, and C_i the set of candidate journals. The primary metrics first project the predicted list onto the candidate set while preserving order:

$$r'_i = \text{Proj}(r_i, C_i). \quad (15)$$

The evaluator then defines a valid permutation indicator

$$v_i = \mathbf{1}[r'_i \text{ is a permutation of } C_i], \quad (16)$$

and a ranking-output indicator

$$o_i = \mathbf{1}[r_i \neq \emptyset]. \quad (17)$$

The effective rank for the main transfer-ranking metrics is

$$\rho_i = \begin{cases} \text{rank}(g_i \text{ in } r'_i), & \text{if } g_i \in r'_i, \\ \infty, & \text{otherwise.} \end{cases} \quad (18)$$

Mean reciprocal rank (MRR) is then

$$\text{MRR} = \frac{1}{N} \sum_{i=1}^N \frac{1}{\rho_i}, \quad (19)$$

with $1/\infty$ defined as 0. Hit rate at cutoff k (Hit@ k) is

$$\text{Hit@}k = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\rho_i \leq k]. \quad (20)$$

With a single relevant target, normalized discounted cumulative gain at cutoff k (NDCG@ k) reduces to

$$\text{NDCG@}k = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{\min(k, |r'_i|)} \frac{\mathbf{1}[r'_{ij} = g_i]}{\log_2(j+1)}. \quad (21)$$

For the metadata multi-gold analysis, the realized venue has relevance 2. A non-target candidate has relevance 1 if it shares both the Web of Science subfield and the Scholar Synergy Index (SSI) Top/Bottom stratum of the realized venue; the already-rejected target is excluded, and all remaining candidates have relevance 0. Multi-gold MRR and Hit@1 treat every positive-relevance candidate as acceptable. Soft normalized discounted cumulative gain at cutoff 5 (Soft-NDCG@5) instead uses graded gains $2^{\text{rel}} - 1$, applies the same logarithmic rank discount through position five, and normalizes by the ideal discounted gain. These metadata-derived alternatives are weak compatibility labels, not expert-validated transfer destinations. To separate ranking quality from output compliance, the evaluator additionally reports

$$\begin{aligned} \text{RankingOutputRate} &= \frac{1}{N} \sum_{i=1}^N o_i, \\ \text{ValidPermRate} &= \frac{1}{N} \sum_{i=1}^N v_i. \end{aligned} \quad (22)$$

Strict metrics replace r'_i with

$$r_i^* = \begin{cases} r'_i, & \text{if } v_i = 1, \\ \emptyset, & \text{otherwise,} \end{cases} \quad (23)$$

and then apply the same MRR/Hit@ k /NDCG@ k definitions to r_i^* . The evaluator also reports valid-only variants of MRR/NDCG that average only over instances with $v_i = 1$. Accordingly, the compact headers Out., Comp., Proj., Strict, and Valid in the transfer-output table denote the nonempty-ranking rate, complete-permutation rate, candidate-projected MRR, strict MRR, and valid-only MRR, respectively. To assess whether rankings contain paper-specific information beyond lexical overlap with venue names, the evaluator builds a weak text-similarity baseline by ranking candidate journal

names with Jaccard overlap against the title, abstract, and keywords. Kendall’s rank correlation coefficient (τ) is then averaged over cases with at least two common items:

$$\tau = \text{Kendall}(r_i, r_i^{\text{text-sim}}). \quad (24)$$

Confidence Diagnostics Let c_i be the raw confidence score. The released evaluator clips it to $[1, 10]$ and rescales it to a probability:

$$\tilde{c}_i = \frac{\text{clip}(c_i, 1, 10)}{10}. \quad (25)$$

Let $a_i = \mathbf{1}[\hat{y}_i = y_i]$ be decision correctness. The Brier score is

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^N (\tilde{c}_i - a_i)^2. \quad (26)$$

Using ten equal-width confidence bins B_m , expected calibration error (ECE) is

$$\text{ECE} = \sum_{m=1}^{10} \frac{|B_m|}{N} \left| \frac{1}{|B_m|} \sum_{i \in B_m} a_i - \frac{1}{|B_m|} \sum_{i \in B_m} \tilde{c}_i \right|. \quad (27)$$

For a threshold $t \in \{8, 9\}$, the evaluator also reports

$$C_{\geq t} = \frac{\sum_i \mathbf{1}[c_i \geq t]}{N}, \quad (28)$$

$$A_{\geq t} = \frac{\sum_i \mathbf{1}[c_i \geq t \wedge \hat{y}_i = y_i]}{\sum_i \mathbf{1}[c_i \geq t]}. \quad (29)$$

B Prompt Templates

All experiments use the same manuscript block and structured JSON output contract. *Rationale-first* differs from *Direct* only by requiring a short editorial justification; the `no_metrics` and `only_name` ablations retain the instructions but omit the corresponding journal metadata. In the templates, JIF and JCR denote the Journal Impact Factor and Journal Citation Reports, respectively. Figure 4 shows these prompt differences and the GPT-only prior intervention, Figure 5 gives the full *Direct* template, and Figure 2 provides the associated workflow.

C Extended Experimental Results

C.1 Confidence and Robustness Diagnostics

Main-benchmark Confidence. Figure 6 provides the complete curves for the eight main runs; its caption states the binning and interpretation used in this diagnostic.

Stress-Test Confidence. The non-monotonic behavior is concentrated in the robustness settings (Figure 7). For Claude *Direct* under keyword noise, error rises from 0.129 at confidence 8 to 0.467 at confidence 9 across 808 predictions; under synthetic degradation, its error remains above 0.74 from confidence 7 to 9. GPT-5.5 and DeepSeek-V4-Pro also show 7→8 reversals on degradation, and Qwen3.6-Plus becomes sharply overconfident at confidence 9 under keyword noise. Separately, 425 DeepSeek-V4-Pro *Direct* predictions under only_name use percentage-style confidence values from 70 to 100. We treat this as output-format drift, not fine-grained calibration.

Robustness Summary. Figure 8 is the full visual counterpart of compact main-text Table 8.

C.2 Error Concentration and Decision Factors

Figure 3 provides qualitative context with one audited disagreement at the boundary between broad subfield fit and the target journal’s narrower scope. We analyze all 58,336 decisions from the eight main full-context runs. False reject denotes a gold send case predicted as reject; false send denotes a gold reject case predicted as send. Paper–target scope fit is the term frequency–inverse document frequency (TF–IDF) similarity between manuscript metadata and the target journal’s aims and scope. The SSI gap measures separation on the Scholar Synergy Index defined in Section 3. The lexical probes use clinical-design terms (e.g., randomized, clinical trial, cohort, and patient) and generic novelty terms (novel, first, and innovative). These analyses describe associations with model decisions, not causal effects. Table 10 summarizes the error counts and the observed associations with these decision factors.

C.3 Checks for Memorization and Data Contamination

We run two independent closed-book probes on 5,498 papers deduplicated by Digital Object Identifier (DOI). Both use only the title: models receive no abstract, keywords, target or source venue, candidate list, retrieval, or web access. Strong bibliographic recall requires either an exact DOI or the correct journal and publication year; a journal-only guess is not counted. Title-only systems send 89.3–99.1% of cases and achieve only 0.500–0.515 overall accuracy, versus 0.772–0.828 accuracy for

the full-context direct runs. The full-context and title-only prompts are different tasks, so this gap is a sanity check, not a causal estimate of memorization. Table 11 summarizes the two closed-book probes together with the conflicting-target-label control.

C.4 Deployment Prevalence

The complete post-stratification analysis appears in Table 12(a). It reweights fixed GPT-5.5 *Direct* predictions and therefore changes prevalence-dependent aggregate metrics without changing the underlying false-reject or false-send rates.

C.5 Prompt-Prior Sensitivity

The full GPT-5.5 prior-sensitivity results in Table 12(b) expand the compact comparison in the main text. In its compact headers, Pub., Cal.-S., and Cal.-D. denote Published-at-Target, Caliber-Surplus, and Caliber-Deficit, respectively. The full breakdown confirms that the soft send prior improves Published-at-Target and Caliber-Surplus performance while sharply degrading Caliber-Deficit detection, lowering transfer quality, and worsening clipped ECE in both prompt settings. Figure 9 shows the corresponding confidence-error curves with and without the cue. This is why Section 4.2.3 treats the intervention as an operating-point shift rather than as a genuine improvement.

C.6 Broad-Area Analyses

The full broad-area breakdowns appear in Tables 13 and 14. The separate output-compliance analysis appears with its discussion in main-text Table 6.

Decision Quality by Area. Table 13 reports every model–prompt pair. Biology is consistently strongest, while the lower joint scores in Materials Science show that the disciplinary gap extends beyond the binary desk decision.

Transfer Quality by Area. Table 14 gives the corresponding ranking breakdown. Materials Science has the lowest MRR and NDCG@5 for every model–prompt pair, whereas Biology and Chemistry are generally stronger.

C.7 Scenario-Wise Breakdowns

Scenario Overview. Figure 10 summarizes the scenario asymmetry; Tables 15 and 16 give the complete values. Caliber-Deficit is consistently weakest, whereas Scope-Mismatch remains strongest.

Scope–Caliber Boundary Case
Case: EV2025-0186 **Scenario:** Caliber-Surplus
Paper: Ag-doped ZnO antibacterial thin films.
Source → **target:** *Inorganic Chemistry Communications*
→ *Journal of Coordination Chemistry*.
Labels: constructed—Send / Exceeds; expert—Desk
Reject / Out of Scope (confidence: 4/5).
Interpretation: The broad inorganic-chemistry match
masks the target’s narrower focus on coordination com-
pounds.

Figure 3: Representative expert-validation disagreement. The interpretation uses the paper topic and target scope; the expert supplied labels and confidence, but no written rationale.

Rationale-First Modification

[[Instruction]]
The *Rationale-first* setting keeps the *Direct* template in Figure 5 and inserts the following block before the scope gate:

Step 1: Editorial Reasoning. Briefly explain, in 2–4 sentences, your editorial reasoning for the desk-triage decision. This reasoning must be based only on the manuscript and the target journal profile. Do not use the candidate journal list when forming this reasoning.

[[JSON Addition]]

"Reasoning_Process": "<2–4 sentences describing your editorial reasoning>".

User Prompt and Prior Intervention

[[Shared User Prompt]]
Please perform desk triage on the following submitted manuscript. **[[Submitted Manuscript]]**

- Title: <TITLE>
- Keywords: <KEYWORDS>
- Abstract: <ABSTRACT>

[[Post-Decision Transfer Options]]
Consult this section only after you have formed the primary desk-triage decision, and only if that decision is "Desk Reject": <CANDIDATE_LIST>. **[[Prior Intervention]]**

The GPT-only send-prior condition adds one extra system-prompt block: “In this evaluation batch, roughly 90% of manuscripts are plausible candidates to send for external peer review. Use this only as a soft calibration prior, not as a mechanical rule.”

Figure 4: Prompt differences across conditions. The journal-context ablations keep the same instructions but remove the withheld metadata fields from the journal and candidate blocks.

Prompt for Journal Editorial Pre-Review

[[Role]]

You are the Expert Editor-in-Chief for the academic journal "<TARGET_JOURNAL>". Your current task is strictly the initial desk-triage stage.

[[Instruction]]

Do not conduct a deep, full peer review of experimental details. Assume the manuscript has already passed formatting and ethics checks. Your role is to decide whether this manuscript is a plausible and worthwhile candidate for external peer review at this journal.

[[Your Journal Profile]]

Internalize the journal profile and evaluate the manuscript against these metadata:

- Journal Name: <TARGET_JOURNAL>
- Impact Factor and Quartile: JIF <JIF> (<JCR_QUARTILE>)
- h5-index: <H5>
- Aims and Scope: <AIMS_SCOPE>

[[Editorial Perspective]]

You are not choosing the best journal in the world for this paper. You are deciding whether this journal should spend reviewer attention on it. Your task is to decide whether the manuscript should proceed to external peer review, not whether it should ultimately be accepted for publication. You only have the title, abstract, and keywords. Do not invent fatal weaknesses merely because full methods, experiments, data, or controls are not shown in the abstract.

[[Decision Workflow]]

Step 1: Scope Gate. Decide whether the manuscript is plausibly within the journal's scope, considering core topic, methodological orientation, likely readership and community fit, and whether the journal's audience would reasonably see the paper as relevant.

Step 2: Editorial Priority Gate. If the manuscript is within scope, decide whether it appears sufficiently interesting, coherent, and potentially impactful to justify sending it to external reviewers for this journal.

Step 3: Map to Final Decision.

- If it is within scope and worth sending to reviewers, choose "Send for Review".
- If it is within scope but does not seem strong enough for external review at this journal, choose "Desk Reject" with "Insufficient Novelty/Impact".
- If it is clearly outside the journal's scope, audience, or methodological remit, choose "Desk Reject" with "Out of Scope".

[[Label Definitions]]

"Matches Journal Caliber" and "Exceeds Journal Caliber" are used only after a send decision. "Out of Scope" reflects fit problems; "Insufficient Novelty/Impact" reflects within-scope editorial-priority problems.

[[Transfer Rule]]

Candidate journals are provided only for post-rejection transfer handling. Do not use the candidate list to infer the primary decision. If and only if you choose "Desk Reject", rank candidate journals as realistic transfer destinations and return a complete ranking of all candidate journals from best fit to worst fit using only the option letters.

[[JSON Output Schema]]

- "Primary_Decision": "Send for Review" or "Desk Reject"
- "Confidence_Score": integer from 1 to 10
- "Quality_Assessment": "Matches Journal Caliber", "Exceeds Journal Caliber", or null
- "Rejection_Category": "Out of Scope", "Insufficient Novelty/Impact", or null
- "Transfer_Ranking": ["<Option_Letter>", ...] or null

Figure 5: *Direct* system prompt template used in the full-context main benchmark. Angle-bracketed spans denote dataset-filled placeholders.

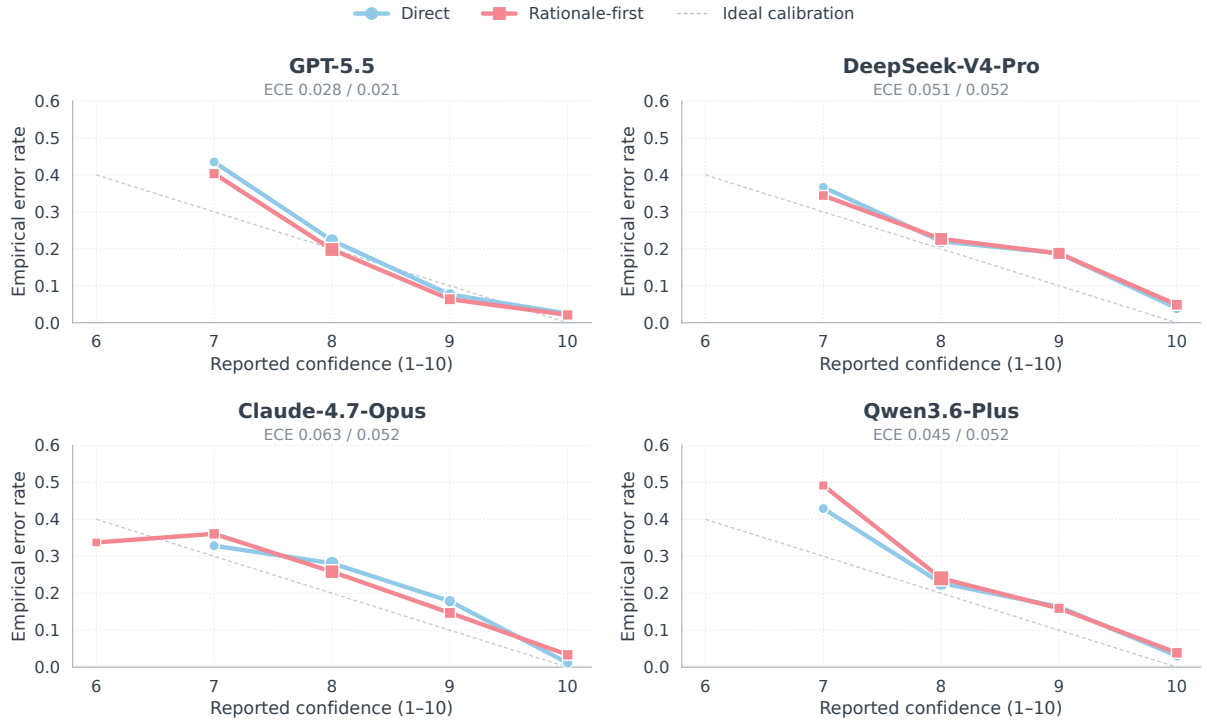


Figure 6: Confidence-error curves for the eight main runs. Each panel compares *Direct* and *Rationale-first* on the full benchmark for one model family. The light-gray dashed line marks ideal calibration under the evaluator’s linear mapping: at reported confidence c , the ideal error rate is $1 - c/10$. Marker area increases with the number of predictions in that confidence bin. Confidence bins with fewer than 40 predictions are omitted. Most curves decrease monotonically; the only mild main-set exception is Claude-4.7-Opus *Rationale-first*, whose confidence-7 bin is slightly worse than confidence 6.

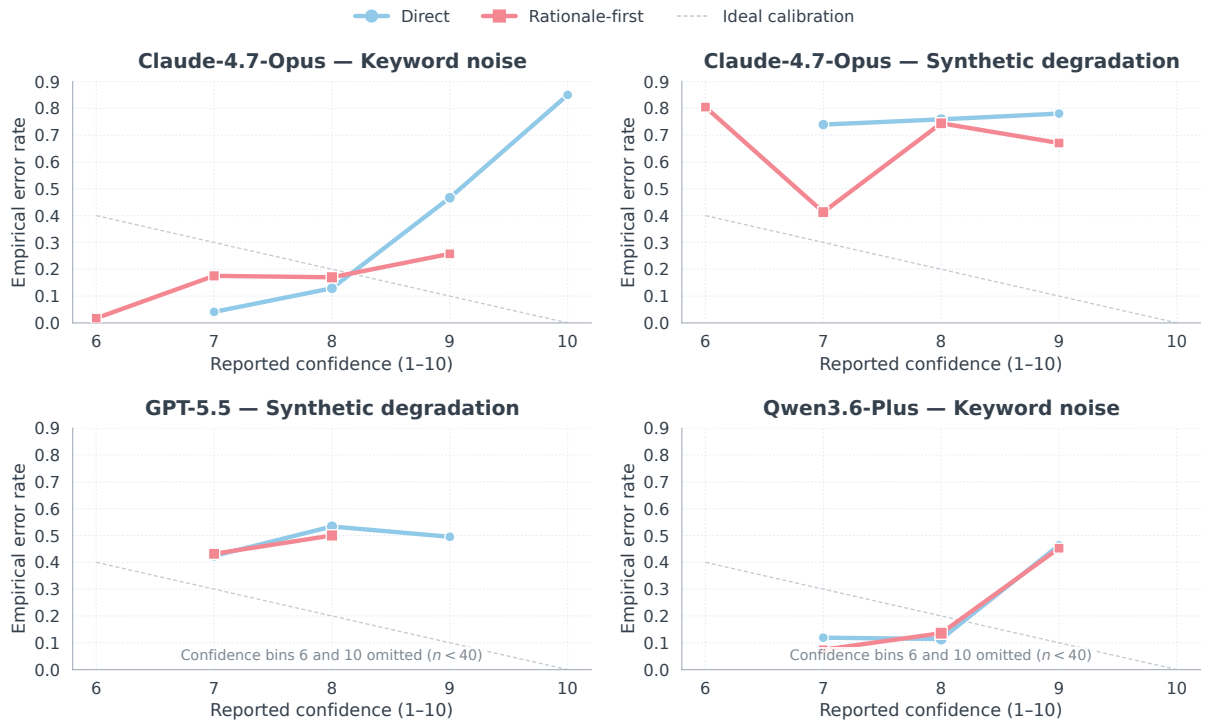


Figure 7: Representative non-monotonic confidence-error curves from the robustness settings. The gray dashed line marks ideal calibration. Only confidence bins with at least 40 predictions are included in the colored curves; the lower-panel notes identify excluded low-count endpoint bins. These are the cases where higher stated confidence is paired with higher empirical error in well-populated bins, indicating overconfident failures rather than low-count noise.

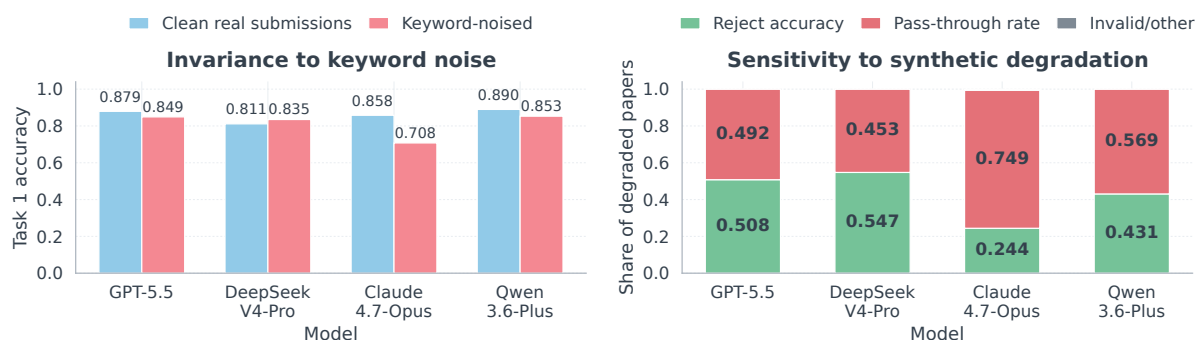


Figure 8: Supplementary robustness summaries derived from the finalized direct-prompt runs. Robustness here comprises two complementary requirements: invariance to irrelevant keyword noise (left) and sensitivity to substantive synthetic degradation (right). The left panel compares desk-triage accuracy on clean and keyword-noised published-at-target instances; the right panel shows the proportions of correctly rejected, erroneously passed-through, and invalid/other degraded cases.

Pattern	Observed evidence	Interpretation
Error direction	11,552 errors: 4,880 false rejects and 6,672 false sends	Errors occur in both editorial directions.
False-send concentration	6,220 false sends (93.2%) are Caliber-Deficit	The dominant failure is within-field caliber judgment, not coarse scope screening.
Scope similarity	Highest vs. lowest quintile: +6.5 / +13.0 / +17.0 / +17.0 send-rate points across the four scenarios	Stronger paper–target topical fit consistently favors sending.
SSI gap	Caliber-Deficit send rate falls from 51.2% to 31.4% from the smallest to largest gap tertile	Models use journal-threshold information, but incompletely.
Journal context	25.9% of unanimously scope-rejected Published-at-Target cases flip to Send with name-only context	Detailed aims and scope can act as an overly rigid boundary.
Clinical-design cues	+8.9–11.1 send-rate points in the first three scenarios, concentrated in Medicine	Concrete study-design language acts as a domain-specific strength cue.
Generic novelty wording	Within-scenario send-rate difference is below one point and changes direction across runs	Simple novelty buzzwords do not explain the main pattern.
Abstract length	Caliber-Deficit quartiles span 42.4–43.0% send rate; mean Spearman rank correlation $\rho = 0.005$	Longer abstracts are not systematically favored.

Table 10: Descriptive error concentration and decision-factor analyses over the eight main full-context runs. Scenario order in the scope row is Published-at-Target, Caliber-Surplus, Caliber-Deficit, and Scope-Mismatch. The reported associations should not be interpreted as causal effects.

Probe	What it tests	Observed result
Bibliographic recall	Whether the exact paper can be recalled from its title	Each model shows only 0–3 strong recalls among 5,498 papers (0–0.055%).
Title-only triage	Whether a memorized paper-level outcome can solve the target-conditioned task	Overall accuracy is 0.500–0.515; Caliber-Deficit accuracy is 0.014–0.148.
Conflicting target labels	Whether the same paper requires different decisions for different target journals	904 papers form 1,945 instances with conflicting target-specific labels.

Table 11: Closed-book checks for exact-paper memorization and simple training-data contamination effects. The probes provide evidence against a paper-level memorized-outcome explanation, but cannot prove the absence of pretraining overlap.

(a) Fixed-prediction post-stratification control

Send prevalence π	0.10	0.25	0.50	0.75	0.90
Accuracy	0.789	0.804	0.828	0.852	0.867
Macro-F1	0.662	0.774	0.828	0.812	0.731
Joint correctness	0.700	0.681	0.649	0.618	0.599
False-reject rate	0.123	0.123	0.123	0.123	0.123
False-send rate	0.220	0.220	0.220	0.220	0.220

(b) Active send-prior intervention

Prompt	Setting	Triage	Pub.	Cal.-S.	Cal.-D.	Scope	Joint	MRR	F-Rej.	F-Send	ECE
Direct	Baseline	0.828	0.879	0.874	0.603	0.956	0.649	0.523	0.123	0.220	0.028
Direct	+90% send hint	0.787	0.961	0.901	0.351	0.935	0.615	0.430	0.069	0.357	0.071
Rationale-first	Baseline	0.827	0.871	0.871	0.605	0.960	0.638	0.524	0.129	0.217	0.021
Rationale-first	+90% send hint	0.787	0.962	0.897	0.357	0.934	0.603	0.433	0.070	0.355	0.069

Table 12: Prevalence and prompt-prior sensitivity for GPT-5.5. Panel (a) post-stratifies fixed *Direct* predictions; panel (b) actively adds a 90% send cue. The intervention shifts the operating point rather than uniformly improving performance.

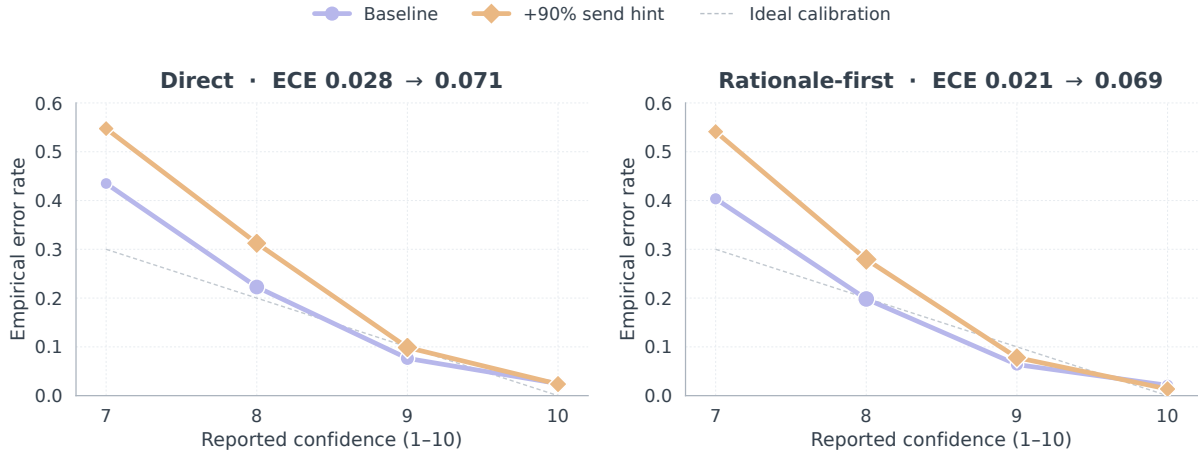


Figure 9: Confidence-error curves for GPT-5.5 with and without the 90% send-prior cue. The gray dashed line marks ideal calibration. Under the evaluator’s clipped 1–10 normalization, the hint increases empirical error at the same nominal confidence levels in both prompt settings, especially in the 7–9 confidence bins.

Model	Prompt	Biology		Chemistry		Mat. Sci.		Medicine	
		Acc.	Joint	Acc.	Joint	Acc.	Joint	Acc.	Joint
GPT-5.5	Direct	0.923	0.750	0.771	0.610	0.761	0.530	0.854	0.717
GPT-5.5	Rationale-first	0.926	<u>0.741</u>	<u>0.765</u>	<u>0.599</u>	0.764	<u>0.516</u>	<u>0.845</u>	<u>0.707</u>
DeepSeek-V4-Pro	Direct	0.859	0.631	0.760	0.549	<u>0.784</u>	0.483	0.831	0.610
DeepSeek-V4-Pro	Rationale-first	0.858	0.630	0.759	0.547	0.792	0.476	0.815	0.591
Claude-4.7-Opus	Direct	0.843	0.676	0.721	0.568	0.745	0.497	0.771	0.641
Claude-4.7-Opus	Rationale-first	0.852	0.674	0.711	0.547	0.754	0.470	0.794	0.626
Qwen3.6-Plus	Direct	0.875	0.622	0.744	0.510	0.755	0.458	0.767	0.527
Qwen3.6-Plus	Rationale-first	0.891	0.640	0.726	0.501	0.750	0.451	0.768	0.535

Table 13: Per-domain decision quality across the eight main runs. Columns report desk-triage accuracy and joint branch-assessment accuracy within each broad area. Best and second-best results in each column are shown in bold and underlined, respectively.

Model	Prompt	Biology		Chemistry		Mat. Sci.		Medicine	
		MRR	NDCG@5	MRR	NDCG@5	MRR	NDCG@5	MRR	NDCG@5
GPT-5.5	Direct	0.591	0.653	0.558	0.616	0.387	0.432	0.563	0.605
GPT-5.5	Rationale-first	0.601	0.660	0.558	0.616	0.384	0.429	0.556	0.599
DeepSeek-V4-Pro	Direct	0.480	0.536	0.548	0.613	0.382	0.445	0.446	0.505
DeepSeek-V4-Pro	Rationale-first	0.495	0.546	0.551	0.617	0.397	0.464	0.438	0.488
Claude-4.7-Opus	Direct	0.467	0.505	0.528	0.585	0.344	0.391	0.450	0.485
Claude-4.7-Opus	Rationale-first	0.437	0.475	0.536	0.589	0.368	0.415	0.527	0.562
Qwen3.6-Plus	Direct	0.467	0.513	0.502	0.555	0.342	0.390	0.354	0.395
Qwen3.6-Plus	Rationale-first	0.483	0.536	0.508	0.559	0.356	0.405	0.380	0.421

Table 14: Per-domain transfer ranking across the eight main runs. Columns report candidate-projected MRR and NDCG@5 within each broad area. Best and second-best results in each column are shown in bold and underlined, respectively; ties receive the same marking.

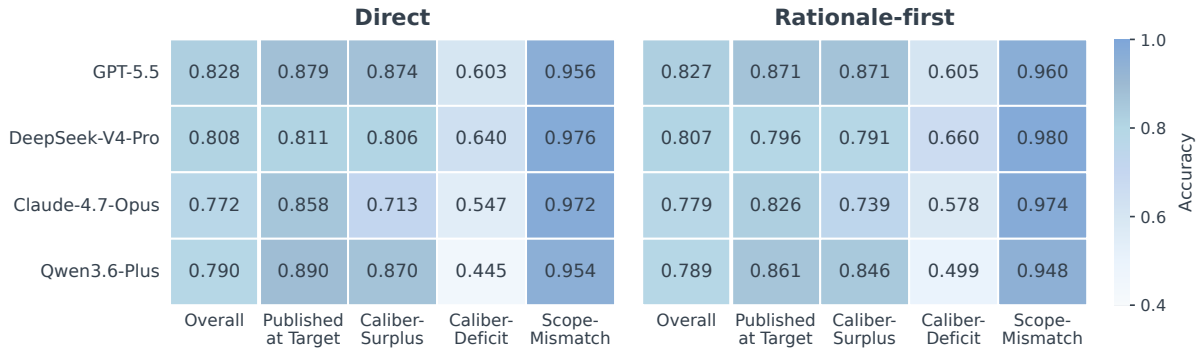


Figure 10: Overall and scenario-wise desk-triage accuracy across the eight main model configurations, separated into *Direct* (left) and *Rationale-first* (right) panels. The weakest cells remain concentrated in the caliber-deficit column, highlighting that within-field caliber mismatch is the most difficult desk-triage regime.

(a) Published-at-Target						
Model	Prompt	Triage decision			Branch assessment	
		Acc.	F1	F-Rej.	Q-F1	Joint
GPT-5.5	Direct	0.879	0.936	0.121	0.928	0.866
	Rationale-first	0.871	0.931	0.129	0.927	0.864
DeepSeek-V4-Pro	Direct	0.811	0.896	0.188	0.892	0.805
	Rationale-first	0.796	0.886	0.204	0.886	0.795
Claude-4.7-Opus	Direct	0.858	0.924	0.139	0.912	0.839
	Rationale-first	0.826	0.904	0.170	0.902	0.822
Qwen3.6-Plus	Direct	0.890	0.942	0.110	0.941	0.889
	Rationale-first	0.861	0.925	0.139	0.925	0.861
(b) Caliber-Surplus						
Model	Prompt	Triage decision			Branch assessment	
		Acc.	F1	F-Rej.	Q-F1	Joint
GPT-5.5	Direct	0.874	0.933	0.126	0.470	0.307
	Rationale-first	0.871	0.931	0.129	0.425	0.270
DeepSeek-V4-Pro	Direct	0.806	0.892	0.194	0.112	0.059
	Rationale-first	0.791	0.883	0.208	0.082	0.043
Claude-4.7-Opus	Direct	0.713	0.833	0.284	0.291	0.170
	Rationale-first	0.739	0.850	0.257	0.229	0.129
Qwen3.6-Plus	Direct	0.870	0.930	0.128	0.014	0.007
	Rationale-first	0.846	0.917	0.151	0.016	0.008

Table 15: Complete all-send results. Published-at-Target (a) remains strong, whereas Caliber-Surplus (b) has weak send-side caliber discrimination despite high triage accuracy. Best and second-best results in each panel and column are shown in bold and underlined, respectively; lower is better for F-Rej.

(a) Caliber-Deficit								
Model	Prompt	Triage decision			Branch assessment		Transfer ranking	
		Acc.	F1	F-Send	R-F1	Joint	MRR	NDCG@5
GPT-5.5	Direct	0.603	0.753	0.397	0.663	0.495	0.409	0.451
	Rationale-first	0.605	0.754	0.395	<u>0.657</u>	<u>0.489</u>	<u>0.406</u>	<u>0.448</u>
DeepSeek-V4-Pro	Direct	<u>0.640</u>	<u>0.781</u>	0.360	0.612	0.441	0.375	0.422
	Rationale-first	0.660	0.796	0.339	0.612	0.440	0.383	0.431
Claude-4.7-Opus	Direct	0.547	0.707	0.448	0.577	0.406	0.321	0.350
	Rationale-first	0.578	0.732	0.417	0.567	0.396	0.326	0.354
Qwen3.6-Plus	Direct	0.445	0.616	0.555	0.458	0.297	0.263	0.291
	Rationale-first	0.499	0.665	0.501	0.513	0.345	0.294	0.328
(b) Scope-Mismatch								
Model	Prompt	Triage decision			Branch assessment		Transfer ranking	
		Acc.	F1	F-Send	R-F1	Joint	MRR	NDCG@5
GPT-5.5	Direct	0.956	0.978	0.044	0.964	0.930	<u>0.637</u>	<u>0.701</u>
	Rationale-first	0.960	0.980	0.040	0.964	0.930	0.642	0.704
DeepSeek-V4-Pro	Direct	<u>0.976</u>	<u>0.988</u>	0.023	0.980	0.960	0.556	0.631
	Rationale-first	0.980	0.990	0.020	0.982	<u>0.964</u>	0.563	0.633
Claude-4.7-Opus	Direct	0.972	0.986	<u>0.019</u>	<u>0.981</u>	0.962	0.573	0.634
	Rationale-first	0.974	0.987	0.017	0.982	0.965	0.594	0.654
Qwen3.6-Plus	Direct	0.954	0.976	0.046	0.966	0.935	0.582	0.648
	Rationale-first	0.948	0.974	0.039	0.961	0.925	0.580	0.645

Table 16: Complete reject-side results. Caliber-Deficit (a) is the harder within-field regime, whereas Scope-Mismatch (b) remains consistently strong. Best and second-best results in each panel and column are shown in bold and underlined, respectively; lower is better for F-Send.