PreReviewBench: Benchmarking LLMs for Editorial Desk Screening in Scholarly Publishing

Anonymous ACL submission

Abstract

Journal editorial pre-review determines which manuscripts reach peer review and shapes reviewer allocation in scholarly publishing. Yet this stage has been largely absent from natural language processing (NLP) evaluation. We introduce PREREVIEWBENCH[1], a benchmark for large language model (LLM)-based journal pre-review with 7,292 balanced paper–journal instances spanning four scientific areas, eight subfields, and three tasks: triage decision, branch-specific editorial judgment, and transfer-journal recommendation. Across four current LLM families and 42 runs covering prompting variants, context ablations, robustness tests, and a send-prior intervention, we find a clear capability hierarchy: current LLMs are relatively strong at coarse scope screening, much weaker at within-field caliber judgment, and weaker still at transfer recommendation. Their decisions are also sensitive to noisy keywords, degraded abstracts, and prompt priors. These results suggest that metadata-level journal pre-review is only partially tractable for current LLMs and is better suited to supervised editorial assistance than to autonomous deployment.

1 Introduction

Peer review is costly and unevenly distributed across science, with estimates exceeding 130 million reviewer hours globally in 2020 (Aczel et al., 2021). Before a manuscript reaches external review, editors must decide from limited evidence whether it is suitable, sufficiently strong, and worth allocating reviewer attention to. These desk decisions shape reviewer workload and author waiting time, and desk screening is a routine part of journal practice, with many manuscripts filtered before external review (Primack et al., 2019; Garand and Harman, 2021). Yet natural language processing (NLP) and artificial intelligence (AI) work has focused far more on conference peer review than on journal editorial triage (Staudinger et al., 2024).

This gap matters because journal pre-review differs from conference peer review in three ways. The decision is made under tighter information constraints, often using only manuscript metadata and target-journal information. It must jointly assess topical fit and whether the paper meets the target journal's editorial bar, and when a paper should not be sent to review at the target venue, the editor faces a natural transfer question: which alternative journal is plausible? Existing resources still do not directly cover this setting, as prior work has largely focused on conference submissions, reviews, and post-submission review tasks rather than journal editorial gatekeeping (Kang et al., 2018; Dycke et al., 2023; Staudinger et al., 2024; Du et al., 2024; Ou et al., 2025).

We therefore introduce PREREVIEWBENCH, a benchmark for LLM-based journal editorial pre-review spanning three tasks: pre-review decision, branch-specific editorial judgment conditioned on that decision, and transfer-journal recommendation for rejected papers.

Our contributions are fourfold: (1) a balanced 7,292-instance benchmark spanning four broad fields, eight WoS subfields, and 58 journals; (2) a controlled construction pipeline that separates quality and scope effects through journal stratification, outlier filtering, and synthetic pairing; (3) a 42-run evaluation of four current LLM families spanning direct and rationale-first prompting, journal-context ablations, keyword-noise and synthetic-degradation stress tests, and a send-prior intervention; and (4) evidence that current LLMs remain much less reliable on fine-grained within-field caliber judgment, robustness under perturbation, and transfer recommendation than on coarse scope screening.

2 Related Work

2.1 Peer-review Benchmarks

Research on automated scientific reviewing has primarily targeted conference peer review. Canon-

[1]Code and data are available at: https://anonymous.4open.science/r/PreReviewBench.

ical resources such as PEERREAD and NLPEER provide conference submissions, reviews, and accept/reject labels (Kang et al., 2018; Dycke et al., 2023), while related work studies review structure, aspect prediction, and decision support from review evidence (Hua et al., 2019; Li et al., 2020; Dycke et al., 2021). As summarized by Staudinger et al. (2024), this landscape remains strongly conference-centered. Our setting differs because the decision is made *before* external reviews exist and targets journal desk triage rather than conference acceptance.

2.2 LLM Review Assistance

Another line of work studies LLMs as review generators, critics, or simulators. This includes early review generation systems such as REVIEWROBOT (Wang et al., 2020), later review-assistance and simulation settings such as REVIEWCRITIQUE, AGENTREVIEW, and TREEREVIEW (Du et al., 2024; Jin et al., 2024; Chang et al., 2025), and related work on meta-review support or revision under review feedback (Hossain et al., 2025; D'Arcy et al., 2024; Liu and Shah, 2023; Idahl and Ahmadi, 2025). These settings ask whether LLMs can participate in peer review after submission. Our benchmark instead targets the earlier editorial gate that decides whether reviewer attention is allocated at all.

2.3 Review Quality and Robustness

Recent work has also examined what makes reviews useful or trustworthy, including aspect identification, weak-review detection, review utility, grounded critique, and focus imbalance in LLM reviews (Lu et al., 2025; Purkayastha et al., 2025; Sadallah et al., 2025; Ou et al., 2025; Shin et al., 2025). Related studies emphasize structured critical assessment or expose adversarial brittleness in automated reviewing (Dycke et al., 2025; Lin et al., 2025), and broader evaluations likewise show that fluent outputs do not imply reliable paper assessment (Zhou et al., 2024). This literature motivates our robustness tests on keyword noise, synthetic degradation, and prompt-prior sensitivity.

2.4 Journal Recommendation

Task 3 is related to journal recommendation, but existing systems solve a different problem. Early similarity-based tools such as JANE (Schuemie and Kors, 2008) and later learning-based recommenders such as CNAVER and HASVRec (Pradhan and Pal, 2020; Pradhan et al., 2020) aim to suggest venues from manuscript content, sometimes with explanation or stronger scientific document representations (de Campos et al., 2024; Beltagy et al., 2019; Cohan et al., 2020; Singh et al., 2023). A nearby line studies reviewer recommendation and paper-reviewer matching (Anjum et al., 2019; Thorn Jakobsen and Rogers, 2022; Karan et al., 2025). None of these settings models the coupled editorial problem of first desk rejecting a paper and then recommending a realistic transfer venue.

Across these lines, prior work rarely targets journal editorial pre-review itself. Our benchmark is journal-centered rather than conference-centered, evaluates abstract-level desk-triage decisions rather than full review artifacts, and unifies triage, branch-specific editorial judgment, and transfer recommendation in a single benchmark.

3 PreReviewBench

This section describes how we turn raw 2025 journal and article metadata into controlled paper–journal instances for editorial triage. As shown in Figure 1, the pipeline proceeds in four stages: defining the disciplinary scope and candidate journal pool, stratifying journals within each subfield by SSI, filtering topically isolated venues, and constructing paired instances for the benchmark tasks.

3.1 Data sources and scope

We construct the benchmark from metadata of journal articles published in 2025 and indexed in Web of Science (WoS)[2]. We focus on four broad scientific areas: Biology, Chemistry, Materials Science, and Medicine. Each broad area is represented by two WoS subfields, for a total of eight subfields.

For each broad area, we first retain WoS subfields that have a corresponding Google Scholar Top Publications listing[3]. Within each retained subfield, we keep journals that appear in both systems and use this shared set as the candidate venue pool. Each paper instance contains the manuscript title, abstract, and keywords. Each journal profile contains the journal name, Journal Impact Factor (JIF), Journal Citation Reports (JCR) quartile, h5-index, and official aims & scope text. All benchmark instances are constructed from same-year 2025 papers to avoid mixing historical and contemporary venue conditions.

[2] https://clarivate.com/webofsciencegroup/solutions/web-of-science/

[3] https://scholar.google.com/citations?view_op=top_venues&hl=en

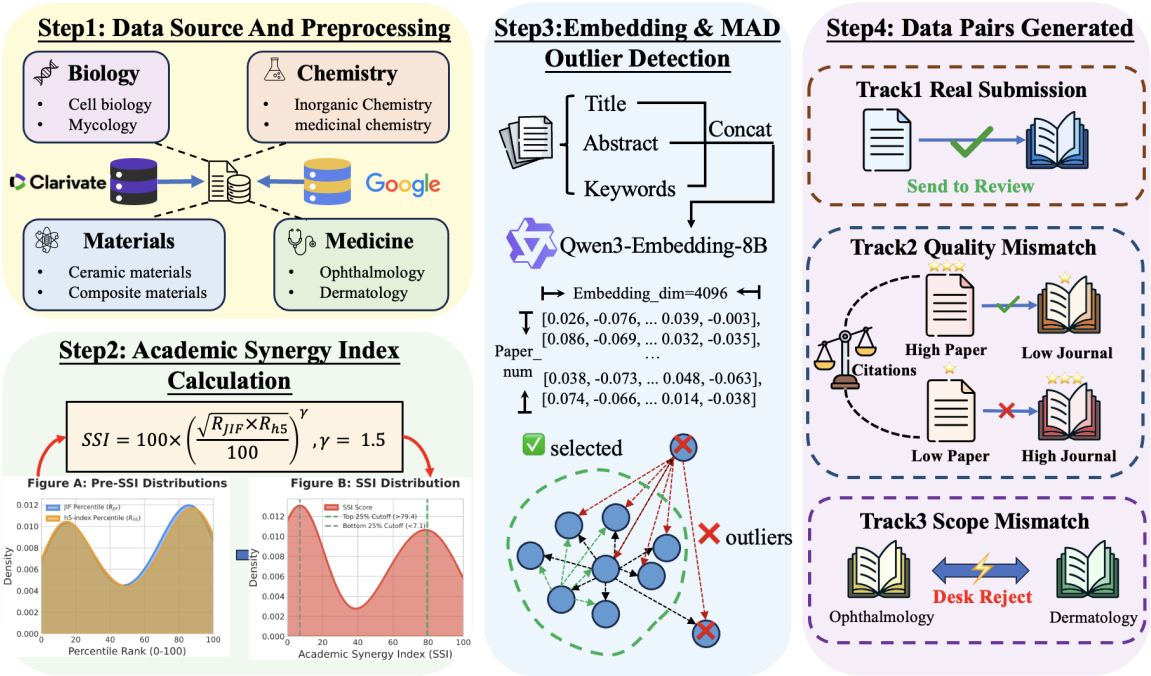

Figure 1: PreReviewBench construction pipeline. Starting from Web of Science and Google Scholar journal pools, we stratify journals by SSI, remove topical outliers with embedding-based MAD filtering, and then generate controlled real, caliber-mismatch, and scope-mismatch paper–journal pairs.

3.2 Journal stratification with the Scholar Synergy Index

To operationalize relative editorial bar within each subfield, we define a Scholar Synergy Index (SSI):

$$SSI = 100 \times \left(\frac{\sqrt{R_{\text{JIF}} \times R_{h5}}}{100}\right)^{\gamma}, \qquad \gamma = 1.5. \tag{1}$$

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

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

We use SSI as the operational proxy for within-subfield journal caliber and retain only the top 25% and bottom 25% of journals in each subfield. The geometric mean penalizes journals that are strong in only one citation signal, while the exponent γ sharpens separation between higher- and lower-selectivity venues for downstream pair construction.

3.3 Embedding-based outlier filtering

Before constructing paper-journal pairs, we filter journals that are topically isolated within their subfield. Using the Qwen embedding model `qwen3-embedding-8b` (**?**), we embed each paper from its title, keywords, and abstract. For each journal j, we compute a centroid vector v_j by averaging its paper embeddings and then L2-normalizing the result. We score a journal by the median cosine distance to other journals in the same subfield:

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

We then compute a median absolute deviation (MAD) z-score inside each subfield and mark a journal as an outlier if its MAD z-score exceeds 3.0, provided that the subfield contains at least four journals. This filtering step removes journals whose topical footprint is too isolated to support stable within-subfield pairing.

3.4 Instance construction

We construct four types of paper–journal instances to capture distinct editorial scenarios: *Published-at-Target*, *Caliber-Surplus*, *Caliber-Deficit*, and *Scope-Mismatch*. The first two are gold send cases, whereas the latter two are gold desk-reject cases arising from within-field caliber mismatch and cross-subfield scope mismatch, respectively.

Published-at-Target instances. These instances pair each paper with the journal where it was actually published. They are labeled `Send for Review` with task-2 label `Matches Journal Caliber`. This serves as a proxy for manuscripts that plausibly passed desk triage at the true venue.

Caliber-Surplus instances. These instances pair a paper from a high-SSI journal with a lower-SSI target journal from the same subfield. To widen the quality gap, we compare highly cited papers from the high journal and weakly cited papers from the low journal. Starting from the top 25% most cited papers in the high journal and the bottom 25% most cited papers in the low journal, we shrink the retained percentile bands in 0.05 steps until the minimum citation count in the retained high set exceeds the maximum citation count in the retained low set. They are labeled `Send for Review` with task-2 label `Exceeds Journal Caliber`.

Caliber-Deficit instances. These instances reverse the previous construction, pairing a paper from the low-SSI journal with the high-SSI target journal from the same pair. They are labeled `Desk Reject` with task-2 label `Insufficient Novelty/Impact`.

Scope-Mismatch instances. These instances model scope error by keeping the broad area fixed while moving across the two constituent WoS subfields. For each source subfield, we select the simulated target journal from the other subfield by closest SSI and use the paper as a cross-subfield submission. They are labeled `Desk Reject` with task-2 label `Out of Scope`. For transfer recommendation, candidate journals come from the union of the two paired subfields.

After constructing the raw pairs, we balance the benchmark by subfield and scenario type, sampling scope-mismatch cases to match the caliber-deficit count within each source subfield. The resulting dataset is exactly balanced across the four instance types and supports a unified editorial evaluation protocol. For each paper–journal pair, the model first makes the primary desk-triage decision (`Send for Review` vs. `Desk Reject`) and reports a 1–10 confidence score. It then makes the corresponding branch-specific follow-up judgment: whether a sent paper `Matches Journal Caliber` or `Exceeds Journal Caliber`, or whether a rejected paper is better explained by `Insufficient Novelty/Impact` or `Out of Scope`. For rejected cases, the benchmark further evaluates transfer recommendation by asking the model to rank candidate journals, using the actual publication venue as the target.

3.5 Dataset statistics

Table 1 summarizes the benchmark. The final dataset contains 7,292 instances, 58 source journals, and eight WoS subfields. It is exactly balanced across the four instance types (1,823 each) and across decision polarity (3,646 send vs. 3,646 reject). Candidate list sizes vary from 5 to 16 depending on the subfield pair. The average abstract length is 203.8 words and the average keyword count is 8.8.

4 Tasks and Experimental Setup

4.1 Model suite and inference setting

Our main inference protocol mirrors an editor-facing single-stage decision process, illustrated in Appendix Figure 4. The model receives (i) manuscript metadata: title, abstract, and keywords; (ii) target-journal metadata: journal name, JIF, JCR quartile, h5-index, and aims & scope; and (iii) a candidate journal list for possible transfer recommendation. The prompt explicitly instructs the model to make the primary desk-triage decision before consulting the candidate list, so that the transfer options act only as post-rejection support rather than label leakage.

Our completed model suite covers GPT-5.5, Claude-4.7-Opus, Qwen3.6-Plus, and DeepSeek-V4-Pro (**????**). Each model is run in two prompting modes: Direct and rationale-first. In the rationale-first condition, the model is required to produce a brief 2–4 sentence editorial justification before returning the final structured answer. This is a lightweight verbalized-justification setup rather than an unconstrained reasoning trace. Appendix B reports the exact prompt templates, including the send-prior intervention. Counting the full benchmark, journal-context ablations, keyword-noise tests, synthetic degradation tests, and the GPT-only send-prior stress test, the paper reports 42 finalized runs in total.

4.2 Metrics

Task 1. For the binary send/reject decision, we compute accuracy, macro-averaged F1 (Macro-F1), and balanced accuracy. Accuracy summarizes overall correctness on the balanced benchmark, Macro-F1 reduces sensitivity to any residual class asymmetry, and balanced accuracy averages recall over the send and reject classes. In the main paper, we further emphasize the false-positive rate (FPR) and

Broad area	WoS subfield	# journals	# instances	Candidate sizes
Biology	Cell Biology	10	1,872	10 / 15
Biology	Mycology	5	276	5 / 15
Chemistry	Chemistry, Inorganic	7	936	7 / 14
Chemistry	Chemistry, Medicinal	7	948	7 / 14
Materials Science	Materials Science, Ceramics	9	756	9 / 16
Materials Science	Materials Science, Composites	7	1,188	7 / 16
Medicine	Dermatology	7	484	7 / 13
Medicine	Ophthalmology	6	832	6 / 13
Total		58	7,292	5–16

Table 1: Benchmark composition by source subfield. Candidate sizes are shown as "within-subfield / cross-subfield" because scope-mismatch transfer ranking uses the union of both paired subfields.

false-negative rate (FNR), because false desk rejection and false send-to-review have different editorial costs. Appendix A gives the exact formulas used by the released evaluator.

Task 2. Task 2 is evaluated conditionally: on gold send cases, we score *quality assessment* (`Exceeds Journal Caliber` vs. `Matches Journal Caliber`); on gold reject cases, we score *rejection-cause assessment* (`Insufficient Novelty/Impact` vs. `Out of Scope`). Branch-specific Macro-F1 isolates discrimination within each branch, invalid-prediction rate captures schema or label failures, and *joint editorial correctness* measures end-to-end usefulness by requiring both the correct Task 1 decision and the correct branch-specific Task 2 label.

Task 3. We evaluate Task 3 only on gold desk-reject instances. The evaluator uses the realized publication venue as the sole gold target, projects each predicted list onto the candidate set while preserving order, and then computes mean reciprocal rank (MRR), hit rate at k (Hit@k), and normalized discounted cumulative gain at k (NDCG@k) on that candidate-projected list. This keeps the main Task 3 metrics focused on ranking quality rather than full-permutation compliance. Appendix Tables 10 and 9 separately report format-sensitive diagnostics, including ranking-output rate, valid-permutation rate, and strict metrics. We also compute Kendall's τ against a weak text-only baseline that ranks candidate journal *names* by Jaccard overlap with the paper title, abstract, and keywords.

Confidence diagnostics. We also report confidence-related diagnostics, including Brier score, expected calibration error (ECE), confidence-threshold coverage, and confidence-error curves. Because some runs exhibit confidence-scale drift under the evaluator's 1–10 protocol, we treat these quantities as supplementary diagnostics; complete definitions are given in Appendix A.

4.3 Additional experimental variants

Beyond the full-context benchmark, we evaluate four controlled variants that target distinct editorial failure modes.

- **Journal-context ablation.** In `no_metrics`, we remove JIF, quartile, and h5-index but keep the journal name and aims & scope. In `only_name`, we give only the journal name and remove aims & scope plus all numeric indicators.
- **Keyword-noise robustness.** We inject exactly five noise keywords into each published-at-target instance and test whether the model still sends strong in-scope papers to review. The injected set follows a fixed mixed profile, combining 2–3 globally frequent hot terms with the remainder drawn from broad-area cross-domain pools after deterministic duplicate filtering.
- **Synthetic degradation.** We rewrite published-at-target instances so that they remain in scope for the same journal but become weaker in novelty, clarity, logic, or evidential support. In the 0520 release, each paper receives one degraded counterpart generated at temperature 0.4 in the single `innovation_weakening` mode, which preserves topic and journal fit while making the contribution appear more incremental. These degraded instances are labeled `Desk Reject` with `Insufficient Novelty/Impact`, and the transfer task is removed.
- **Send-prior sensitivity.** For GPT-5.5 only, we add a soft prompt hint stating that roughly 90% of papers in the batch are plausible send cases, to test whether an explicit class-prior cue distorts decisions.

5 Results and Analysis

We report the full set of 42 finalized runs: four model families, two prompt modes, three journal-context conditions, two robustness stress tests, and

one GPT-specific send-prior intervention.

5.1 Main benchmark comparison

Table 2 reports the overall comparison across the eight main model-prompt configurations. Appendix C Tables 11–14 provide the corresponding scenario-wise breakdowns for Published-at-Target, Caliber-Surplus, Caliber-Deficit, and Scope-Mismatch, omitting metrics that are undefined by construction.

Overall ranking. Table 2 shows that GPT-5.5 is the strongest family overall. Its direct prompt variant gives the best Task 1 accuracy (0.828), the best joint Task 2 accuracy (0.649), and the strongest send-side quality F1, while rationale-first is effectively tied on decision quality, including Task 3 MRR (0.524 vs. 0.523). DeepSeek-V4-Pro is the next-strongest family overall, with weaker send-side reasoning than GPT-5.5 but stronger reject-side sensitivity and the second-best transfer MRR under candidate-projected scoring.

Stable editorial styles. The model families exhibit consistent editorial styles rather than prompt-specific personalities. Claude-4.7-Opus is the most conservative family, with FPR around 0.21 in both prompting modes, whereas Qwen3.6-Plus is the most permissive, combining the lowest FPR with the highest FNR. DeepSeek-V4-Pro is the most reject-sensitive family, attaining the lowest overall FNR and the strongest caliber-deficit accuracy in Appendix Table 13, while GPT-5.5 is the most balanced overall. These patterns reveal a substantive separation between reject-side detection and overall editorial competence: a model can be strong at catching reject cases without being equally strong at send-side caliber assessment or transfer recommendation.

Transfer remains limited. Task 3 remains substantially harder than coarse desk screening, but the candidate-projected metrics clarify the source of failure. GPT-5.5 still leads, yet DeepSeek-V4-Pro and Qwen3.6-Plus recover markedly relative to their earlier strict scores, indicating that many of their mistakes are incomplete permutations rather than entirely uninformative rankings. Appendix Table 10 confirms this separation: the largest gaps between projected MRR and strict MRR occur precisely in the runs with the weakest valid-permutation rates. Transfer recommendation is therefore limited by both ranking quality and output reliability, but these two failure modes should not be conflated.

5.2 Scenario and domain analyses

Table 3 summarizes two orthogonal analyses that are central to the benchmark: scenario difficulty and disciplinary variation.

Scenario difficulty. The benchmark is not primarily testing coarse topical mismatch. Averaged across the eight main configurations, Task 1 accuracy reaches 0.965 on Scope-Mismatch cases and 0.849 on Published-at-Target cases, but drops to 0.572 on Caliber-Deficit cases. This asymmetry persists downstream: Caliber-Deficit is also the weakest scenario for joint editorial correctness and transfer recommendation, with average MRR only 0.347 versus 0.591 on Scope-Mismatch. Current LLMs are therefore far better at detecting obvious scope problems than at judging whether an in-scope paper falls below the target journal's editorial bar.

Domain variation. The broad areas are not equally difficult. Biology is easiest on triage and joint editorial judgment, whereas Materials Science is hardest overall and also has the weakest transfer ranking (MRR 0.370). Transfer recommendation is not perfectly aligned with triage difficulty, however: Chemistry slightly surpasses Biology on average MRR (0.536 vs. 0.503). Appendix Tables 8 and 9 report the full per-domain breakdown for every model and both prompting modes.

5.3 Journal-context ablations

Table 4 evaluates how much the models rely on journal metadata beyond the journal name. It reports Task 1 accuracy, joint editorial accuracy, and candidate-projected MRR under the `full`, `no_metrics`, and `only_name` journal-context settings for both prompting modes.

Bibliometric cues help. Removing journal metrics usually weakens the primary editorial decision. GPT-5.5 and DeepSeek-V4-Pro lose 1.0–3.1 Task 1 points and 1.8–2.7 joint-accuracy points under `no_metrics` across the two prompting modes. Claude-4.7-Opus is more mixed: removing metrics can slightly raise direct-prompt Task 1 accuracy, but it still lowers joint correctness. Bibliometric signals therefore remain useful for calibrating the journal's editorial threshold even when topical fit is already inferable from the manuscript and scope text.

Richer context helps transfer. Under the candidate-projected Task 3 metric, the full setting

Model	Prompt	Task 1				Task 2			Task 3	
		Acc.	F1	FPR	FNR	Q-F1	R-F1	Joint	MRR	NDCG@5
GPT-5.5	Direct	0.828	0.828	0.123	0.220	0.588	0.782	0.649	0.523	0.576
	Rationale-first	0.827	0.826	0.129	0.217	0.562	0.777	0.638	0.524	0.576
DeepSeek-V4-Pro	Direct	0.808	0.809	0.191	0.191	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	0.532
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 2: Overall benchmark results across the eight main model-prompt configurations. Task 2 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 Task 1 decision and the correct branch-specific Task 2 label. Task 3 reports candidate-projected MRR and NDCG@5; format-compliance diagnostics appear in Appendix Table 10.

Group	Slice	T1	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
	Chem.	0.745	0.554	0.536
	Mat. Sci.	0.763	0.485	0.370
	Medicine	0.806	0.619	0.464

Table 3: Compact scenario-wise and broad-area summaries averaged across the eight main model-prompt configurations. MRR is undefined for the two all-send scenarios.

is strongest for every model except Claude-4.7-Opus Direct, where removing numeric indicators gives a small MRR gain (0.447 to 0.457) but still lowers joint correctness. The broader pattern is consistent: stripping either bibliometric signals or scope text degrades ranking quality, with the largest name-only losses appearing for GPT-5.5 and DeepSeek-V4-Pro. Journal metadata therefore helps not only the primary send/reject decision, but also the downstream recommendation of plausible transfer venues.

5.4 Robustness to keyword noise and synthetic degradation

Table 5 summarizes the two published-at-target stress tests for both prompting modes. The keyword-noise set preserves the gold label `Send for Review`; the synthetic-degradation set rewrites the same papers so that they remain in scope but should now be rejected for insufficient novelty/impact. Neither stress test includes transfer ranking, because both are defined only on published-at-target instances.

Keyword noise induces false rejects. Keyword noise generally makes models more likely to reject papers that should still be sent to review. GPT-5.5 loses 3.0 points under Direct and 4.2 points under rationale-first; Qwen3.6-Plus loses 3.7 and 2.8 points, respectively. Claude-4.7-Opus is the most fragile under Direct prompting, dropping 15.0 points, but rationale-first substantially stabilizes it, limiting the drop to 1.2 points. DeepSeek-V4-Pro is the exception: it improves under keyword noise in both prompting modes, suggesting that it treats the added topical tokens as informative evidence rather than as distraction.

Synthetic degradation exposes weak within-scope quality judgment. Synthetic degradation is difficult for every model. The best degraded-set reject accuracy is DeepSeek-V4-Pro rationale-first at 0.569, which still means that 43.1% of degraded abstracts are incorrectly sent to review. GPT-5.5 rationale-first is next at 0.532, while Claude-4.7-Opus Direct is particularly vulnerable, passing 74.9% of degraded papers. Because the degraded set remains in scope for the same journal, these failures reflect weak quality judgment rather than scope confusion. Appendix C Figure 9 visualizes the direct-prompt robustness runs.

5.5 Prompt-prior intervention and confidence diagnostics

Table 6 summarizes the compact GPT-5.5 prompt-prior test; Appendix C Table 7 gives the full version with the scenario breakdown.

Prior cues shift the operating point. Even a single soft prior statement changes behavior materially. Under Direct prompting, the 90% send-prior cue raises Published-at-Target accuracy from 0.879 to 0.961, but it also reduces Caliber-Deficit accuracy from 0.603 to 0.351 and lowers MRR from 0.523 to 0.430. The rationale-first variant shows the

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 ↓	0.623 ↓	0.508 ↓	0.808 ↓	0.613 ↓	0.465 ↓
	Rationale-first	**0.827**	**0.638**	**0.524**	0.817 ↓	0.615 ↓	0.503 ↓	0.803 ↓	0.595 ↓	0.460 ↓
DeepSeek-V4-Pro	Direct	**0.808**	**0.566**	**0.465**	0.789 ↓	0.548 ↓	0.437 ↓	0.778 ↓	0.545 ↓	0.409 ↓
	Rationale-first	**0.807**	**0.560**	**0.473**	0.776 ↓	0.534 ↓	0.425 ↓	0.776 ↓	0.536 ↓	0.409 ↓
Claude-4.7-Opus	Direct	0.772	**0.594**	0.447	**0.774** ↑	0.570 ↓	**0.457** ↑	0.770 ↓	0.557 ↓	0.409 ↓
	Rationale-first	**0.779**	**0.578**	**0.460**	0.772 ↓	0.562 ↓	0.450 ↓	0.770 ↓	0.556 ↓	0.422 ↓
Qwen3.6-Plus	Direct	**0.790**	**0.532**	**0.422**	0.771 ↓	0.514 ↓	0.391 ↓	0.770 ↓	0.517 ↓	0.382 ↓
	Rationale-first	**0.789**	**0.535**	**0.437**	0.775 ↓	0.519 ↓	0.397 ↓	0.769 ↓	0.519 ↓	0.389 ↓

Table 4: Journal-context ablations. For each model and prompting mode, columns report Task 1 accuracy, joint editorial accuracy, and Task 3 candidate-projected MRR under the full, no-metrics, and name-only settings. Colored arrow markers indicate whether each score rises or falls relative to the corresponding full-context score, and bold indicates the best value among the three settings for that model, prompt, and metric.

Model	Prompt	Keyword Noise		Synthetic Degradation	
		Clean Send	Noise Send	Reject Acc.	R-F1
GPT-5.5	Direct	0.879	0.849	0.508	0.647
	Rationale-first	0.871	0.828	0.532	0.665
DeepSeek-V4-Pro	Direct	0.811	0.835	0.547	0.650
	Rationale-first	0.796	0.810	0.569	0.661
Claude-4.7-Opus	Direct	0.858	0.708	0.244	0.345
	Rationale-first	0.826	0.814	0.406	0.535
Qwen3.6-Plus	Direct	0.890	0.853	0.431	0.563
	Rationale-first	0.861	0.833	0.419	0.551

Table 5: Robustness stress tests. "Clean Send" and "Noise Send" are Task 1 accuracy on the original and keyword-noised published-at-target sets. "Reject Acc." and "R-F1" are the Task 1 reject accuracy and Task 2 rejection-cause macro-F1 on the synthetic-degradation set, where all instances should now be rejected.

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	Base	0.827	0.871	0.605	0.524	0.021
Rationale	+Prior	0.787	0.962	0.357	0.433	0.069

Table 6: Compact GPT-5.5 send-prior intervention results.

same qualitative pattern. The intervention therefore makes the model more permissive, but in precisely the way that harms the benchmark's most editorially nuanced cases.

Confidence remains exploratory. Under the evaluator's clipped normalization, the same prompt-prior cue more than doubles GPT-5.5's ECE in the direct setting (0.028 to 0.071). More broadly, Appendix C Figures 5–7 show that non-monotonic confidence-error curves are concentrated in robustness conditions rather than in the main benchmark runs. We therefore use calibration plots as diagnostic evidence, not as the basis for strong model-ranking claims.

6 Conclusion

We presented PreReviewBench, a benchmark for LLM-based journal editorial pre-review with 7,292 balanced paper-journal instances across four broad areas, eight subfields, and 58 journals. Across 42 finalized runs, the benchmark shows that journal pre-review is not a single capability: current LLMs are useful for coarse desk screening, but remain substantially less reliable on nuanced caliber judgment, robustness under perturbation, and transfer recommendation. Model families also exhibit stable editorial styles, while rationale-first prompting changes these trade-offs only modestly. No tested system is reliable enough for unsupervised editorial use.

Limitations

The benchmark uses publication outcome and citation-stratified pairing as proxies for desk-triage plausibility and journal caliber. These proxies are useful for controlled evaluation, but they do not directly observe real desk-reject decisions and

they may not capture every editor's threshold. The benchmark is also abstract-level and currently covers four broad scientific areas and eight WoS subfields, so its conclusions may not transfer unchanged to other editorial cultures or to settings that rely on full-paper evidence.

Task 3 remains conservative because real transfer often admits multiple reasonable venues, whereas the evaluator uses the realized publication venue as a single gold target. Our main ranking metrics therefore isolate ordering signal on the candidate-projected list, and we relegate strict format-sensitive scores to supplementary diagnostics. Confidence diagnostics are likewise limited by the evaluator's clipped 1–10 normalization, so we treat them as descriptive rather than definitive. Any public release must also separate redistributable derived annotations from source-dependent metadata and reconstruction scripts.

Ethics Statement

This benchmark is intended for decision support, not autonomous editorial replacement. Strong benchmark performance does not rule out topical, disciplinary, or prestige-related bias, so we frame PREREVIEWBENCH as an audit resource rather than a deployment endorsement. Restricting inputs to title, abstract, keywords, and journal metadata also reduces confidentiality risk relative to full-manuscript release.

References

Balazs Aczel, Barnabas Szaszi, and Alex O. Holcombe. 2021. A billion-dollar donation: estimating the cost of researchers' time spent on peer review. *Research Integrity and Peer Review*, 6:14.

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.

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.

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.

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.

Eftekhar 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, Ilia 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.

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.

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.

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.

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[cite: 99, 110]. Let N be the number of evaluated instances, y_i the gold label, and $\hat{y}_i$ the model prediction for instance i[cite: 203].

A.1 Task 1: Decision Metrics

For the binary send/reject decision[cite: 96],

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

For each class $c \in \{\texttt{Send}, \texttt{Reject}\}$, precision, recall, and F1 are

$$\begin{aligned} P_c &= \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}, \\ R_c &= \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}, \\ F1_c &= \frac{2P_cR_c}{P_c + R_c}. \end{aligned} \tag{5}$$

and macro-averaged F1 is

$$\text{Macro-F1} = \frac{1}{2}\sum_{c} F1_c. \tag{6}$$

Balanced accuracy averages recall across the two decision classes[cite: 97]:

$$\text{BalancedAcc} = \frac{R_{\texttt{Send}} + R_{\texttt{Reject}}}{2}. \tag{7}$$

The false-positive rate (FPR) and false-negative rate (FNR) reported in the paper are branch-specific editorial error rates[cite: 98]:

$$\text{FPR} = \frac{\sum_i \mathbf{1}[y_i = \texttt{Send} \wedge \hat{y}_i = \texttt{Reject}]}{\sum_i \mathbf{1}[y_i = \texttt{Send}]}, \tag{8}$$

$$\text{FNR} = \frac{\sum_i \mathbf{1}[y_i = \texttt{Reject} \wedge \hat{y}_i = \texttt{Send}]}{\sum_i \mathbf{1}[y_i = \texttt{Reject}]}. \tag{9}$$

The evaluator additionally reports parsed-decision rate[cite: 99],

$$\text{ParsedRate} = \frac{\sum_i \mathbf{1}[\hat{y}_i \in \{\texttt{Send}, \texttt{Reject}\}]}{N}, \tag{10}$$

which isolates strict decision quality from schema-compliance failures[cite: 108].

A.2 Task 2: Branch-Specific Metrics

Task 2 is evaluated conditionally on the gold Task 1 branch[cite: 101, 103]. On gold send cases, the label space is $\{\texttt{Matches}, \texttt{Exceeds}\}$; on gold reject cases, it is $\{\texttt{Novelty}, \texttt{Scope}\}$[cite: 101, 102]. For either branch b,

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

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

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[cite: 211]. Invalid-prediction rate is [cite: 102]

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

Joint editorial correctness measures end-to-end usefulness[cite: 102]:

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

A.3 Task 3: Transfer Metrics

Task 3 is evaluated only on gold reject cases[cite: 103]. 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[cite: 104]. The primary metrics first project the predicted list onto the candidate set while preserving order[cite: 104]:

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

The evaluator then defines a valid permutation indicator [cite: 106]

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

and a ranking-output indicator [cite: 106]

$$o_i = \mathbf{1}[r_i \neq \emptyset]. \tag{17}$$

The effective rank for the main Task 3 metrics is [cite: 104]

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

Mean reciprocal rank is then [cite: 104]

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

with $1/\infty$ defined as 0[cite: 217]. Hit@k is [cite: 104]

$$\text{Hit@}k = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\rho_i \le k]. \tag{20}$$

With a single relevant target, NDCG@k reduces to [cite: 104]

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

To separate ranking quality from output compliance, the evaluator additionally reports [cite: 106]

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

Strict metrics replace r'_i with [cite: 106]

$$r^\star_i = \begin{cases} r'_i, & \text{if } v_i = 1, \\ \emptyset, & \text{otherwise,} \end{cases} \tag{23}$$

and then apply the same MRR/Hit@k/NDCG@k definitions to $r^\star_i$[cite: 220]. The evaluator also reports valid-only variants of MRR/NDCG that average only over instances with $v_i = 1$[cite: 106]. 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[cite: 107]. Kendall's τ is then averaged over cases with at least two common items[cite: 107]:

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

A.4 Confidence Diagnostics

Let c_i be the raw confidence score. The released evaluator clips it to $[1, 10]$ and rescales it to a probability[cite: 109]:

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

Let $a_i = \mathbf{1}[\hat{y}_i = y_i]$ be decision correctness. The Brier score is [cite: 108]

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

Using ten equal-width confidence bins B_m, expected calibration error is [cite: 108]

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

For a threshold $t \in \{8, 9\}$, the evaluator also reports [cite: 109]

$$C_{\ge t} = \frac{\sum_i \mathbf{1}[c_i \ge t]}{N}, \tag{28}$$

$$A_{\ge t} = \frac{\sum_i \mathbf{1}[c_i \ge t \wedge \hat{y}_i = y_i]}{\sum_i \mathbf{1}[c_i \ge t]}. \tag{29}$$

B Prompt Templates

All experiments share the same manuscript block and the same structured JSON output contract[cite: 111]. The direct and rationale-first settings differ only in whether the system prompt explicitly requires a short editorial justification before the final structured answer[cite: 91, 92]. The `no_metrics` and `only_name` ablations keep the same instructions but remove the corresponding journal metadata fields from the journal and candidate blocks[cite: 111, 112]. Figure 2 gives the full direct template, while Figure 3 isolates the rationale-first addition and the GPT-only prompt-prior intervention[cite: 93, 94]. Figure 4 reproduces the task workflow referenced in Section 4 so that the prompt specification and evaluation pipeline can be read together[cite: 85].

C Extended Results

This appendix consolidates the auxiliary results referenced throughout Section 5. To reduce float-only pages and make the appendix easier to follow, we begin with the most explanatory diagnostics, then present the prompt-prior intervention, followed by the broad-area and compliance tables, and finally the dense scenario-wise breakdowns.

C.1 Confidence and visual diagnostics

Confidence-error curves. We rechecked the 42 finalized confidence-error plots corresponding to the 42 evaluated runs, using the evaluator's own rule of clipping confidence to the 1–10 range and excluding bins with fewer than 40 predictions[cite: 265]. Under that protocol, most curves remain monotone decreasing, and the non-monotonic cases are concentrated in the robustness settings rather than in the main benchmark runs[cite: 266]. We

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

therefore use these plots as exploratory diagnostics rather than as definitive calibration comparisons[cite: 267]. Figure 5 mainly serves as a sanity check that the eight main benchmark runs are mostly well behaved[cite: 268]; the only mild main-set exception is Claude-4.7-Opus rationale-first, whose confidence-7 bin is slightly worse than confidence 6[cite: 269]. Figure 6 then isolates the GPT-5.5 send-prior shift, while Figure 7 focuses on the genuinely counterintuitive stress-test failures[cite: 269]. The strongest examples are Claude Direct under keyword noise, where error

Rationale-First Modification

[[Instruction]]

The rationale-first setting keeps the direct template above 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 3: Prompt differences across conditions. The journal-context ablations keep the same instructions but remove the withheld metadata fields from the journal and candidate blocks.

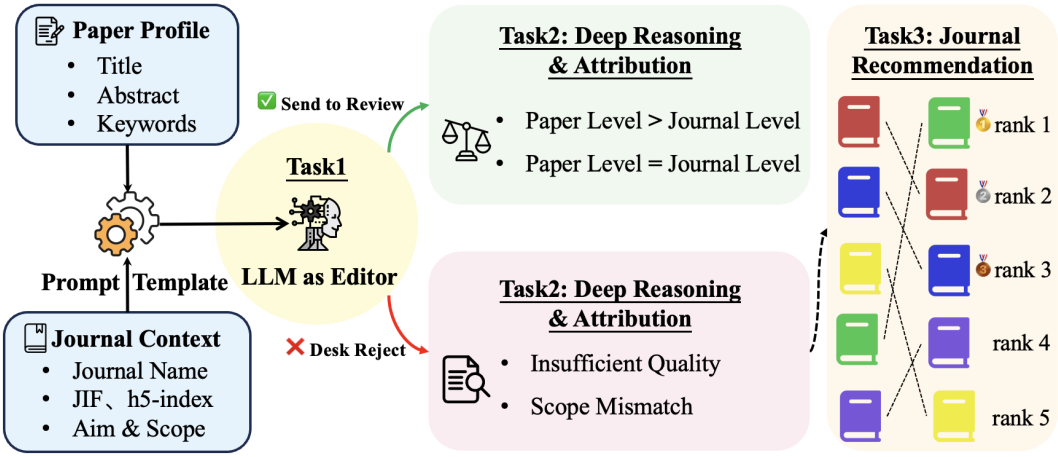

Figure 4: Task framework. The model first makes a desk-triage decision from paper metadata and the target-journal profile. If it sends the paper to review, it predicts whether the paper matches or exceeds the journal's caliber. If it rejects the paper, it predicts whether the main issue is scope mismatch or insufficient novelty/impact, and then ranks candidate transfer journals.

rises from 0.129 at confidence 8 to 0.467 at confidence 9 across 808 predictions, and Claude Direct under synthetic degradation, where error remains above 0.74 from confidence 7 to 9[cite: 270]. GPT-5.5 and DeepSeek-V4-Pro also show 7→8 reversals on synthetic degradation, and Qwen3.6-Plus becomes sharply overconfident at confidence 9 under keyword noise[cite: 270]. One additional protocol issue is not a calibration effect but an output-format drift: in DeepSeek-V4-Pro Direct under `only_name`, 425 predictions use percentage-style confidences from 70 to 100 instead of the requested 1–10 scale[cite: 271]; the evaluator clips them to 10, so we interpret that run as a confidence-format inconsistency rather than as meaningful fine-grained calibration[cite: 272].

Beyond the confidence plots, Figure 8 offers a compact visual summary of scenario-level Task 1 accuracy across the eight main configurations, and Figure 9 complements Table 5 with direct-prompt robustness summaries[cite: 277]. We place these figures here because they mainly elaborate patterns already stated in the main paper: the caliber-deficit condition is the persistent bottleneck, and robustness failures are most visible when the input is perturbed without changing the nominal journal target[cite: 278].

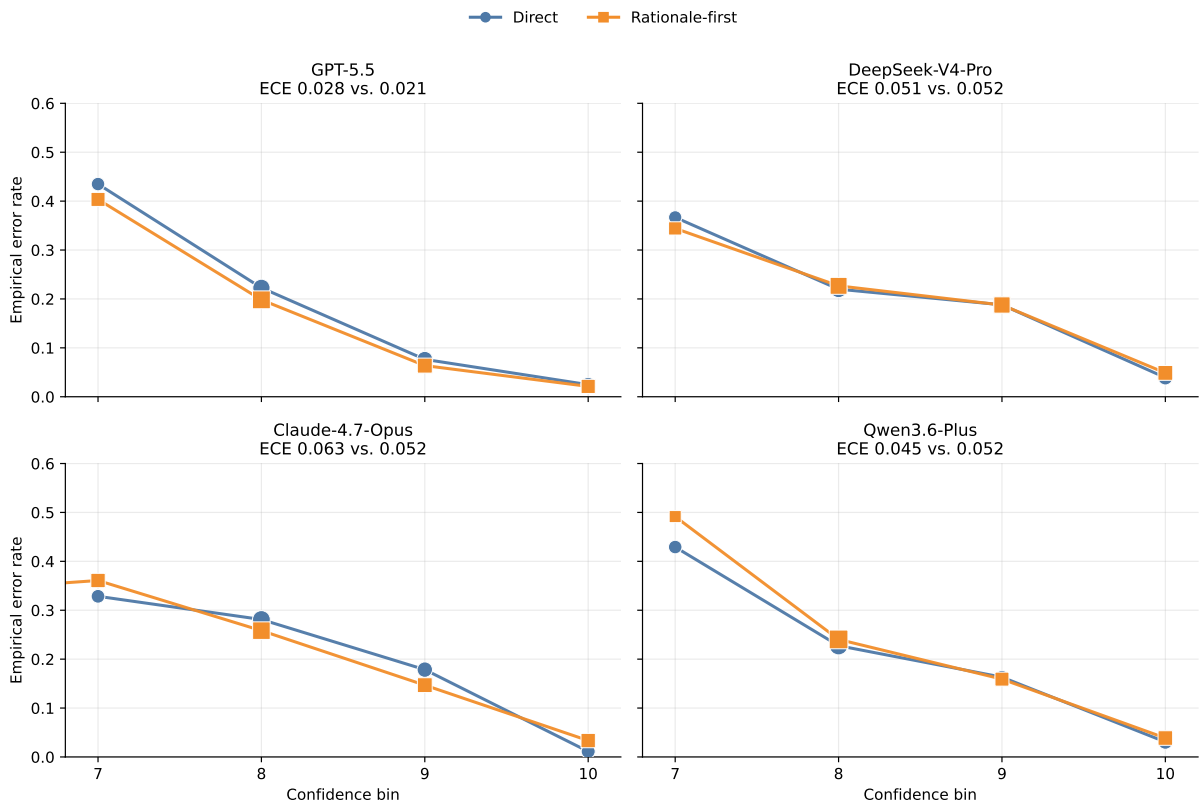

Figure 5: Confidence-error curves for the eight main runs. Each panel compares Direct and rationale-first on the full benchmark for one model family. Marker area is proportional to the number of predictions in that confidence bin. Confidence bins with fewer than 40 predictions are omitted.

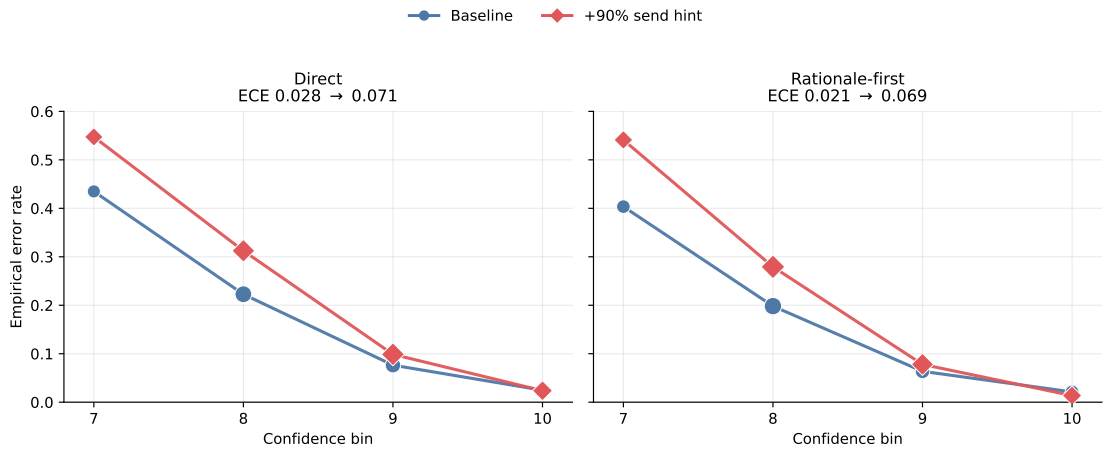

Figure 6: Confidence-error curves for GPT-5.5 with and without the 90% send-prior cue. Under the evaluator's clipped 1–10 normalization, the hint increases empirical error at the same nominal confidence levels in both prompt settings, especially in the 7–9 confidence bins.

C.2 Prompt-prior sensitivity

Table 7 expands the compact GPT-5.5 prior-sensitivity result from the main text[cite: 281]. 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[cite: 282]. This is why Section 5.5 treats the intervention as an operating-point shift rather than as a genuine improvement[cite: 283].

C.3 Broad-area and compliance analyses

Tables 8 and 9 provide the full broad-area breakdowns summarized in the main text, while Table 10 separates ranking-output compliance from ranking quality[cite: 286]. These tables make two patterns

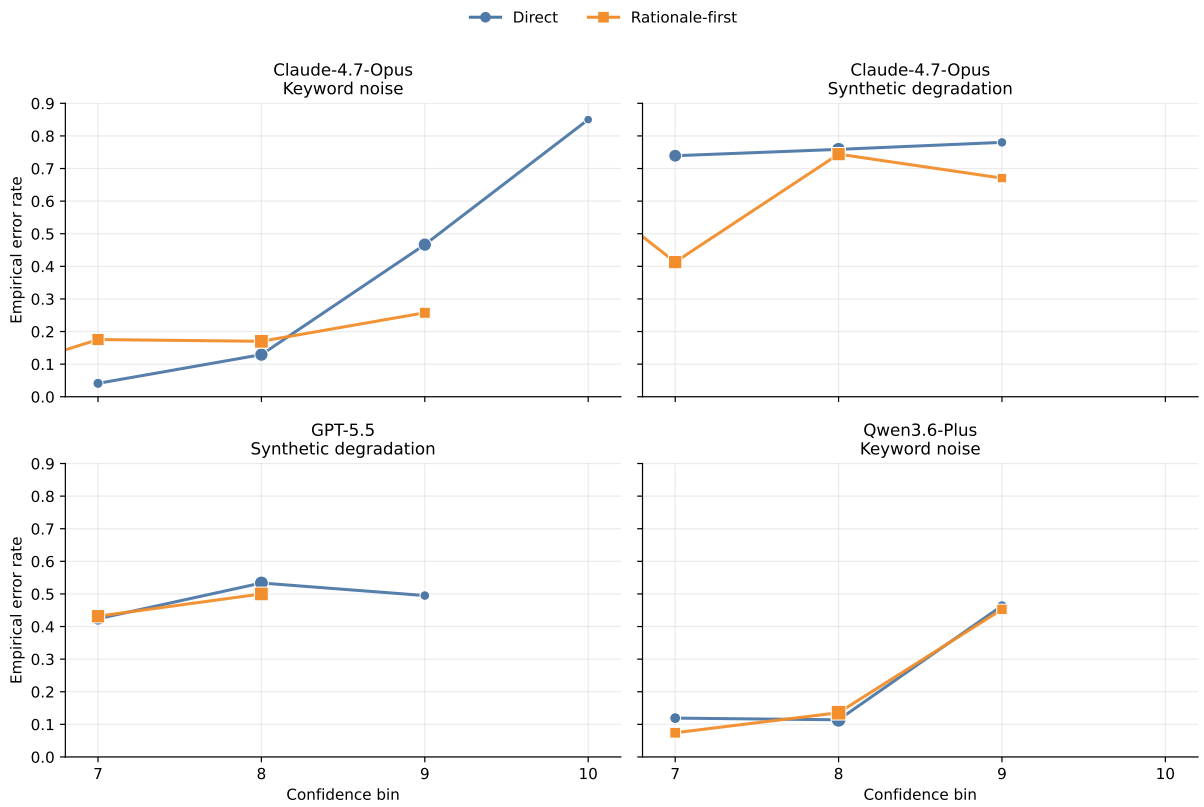

Figure 7: Representative non-monotonic confidence-error curves from the robustness settings. These are the cases where higher stated confidence is paired with higher empirical error in well-populated bins, indicating overconfident failures rather than low-count noise.

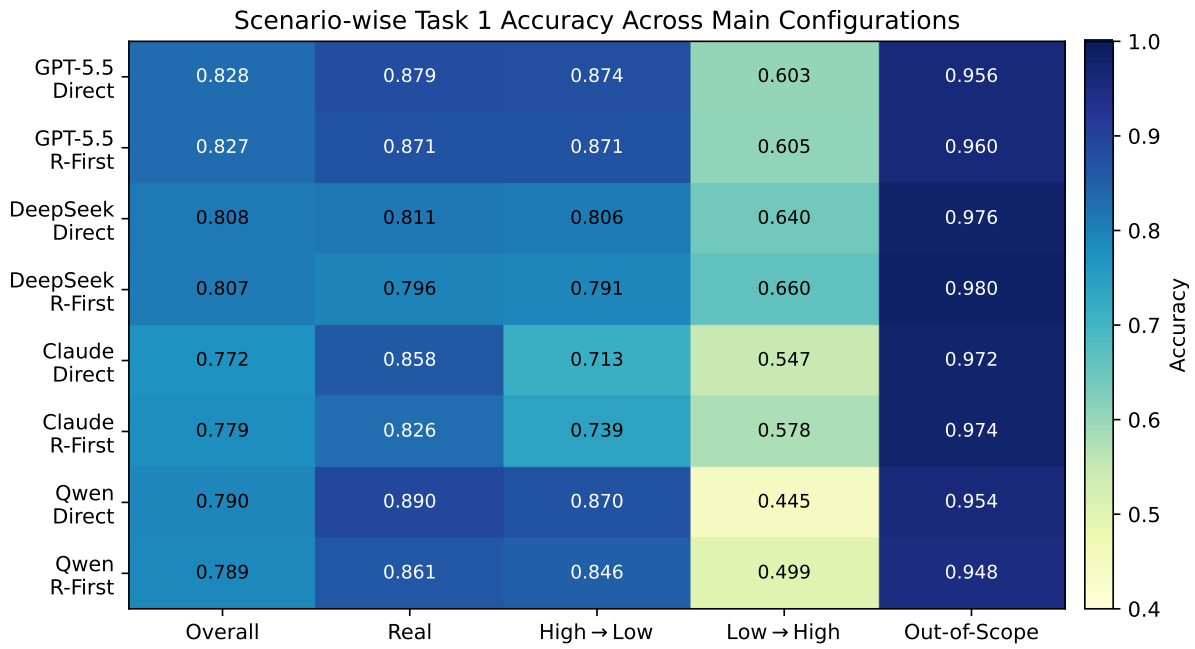

Figure 8: Overall and scenario-wise Task 1 accuracy across the eight main model configurations. The weakest cells remain concentrated in the caliber-deficit column, highlighting that within-field caliber mismatch is the most difficult desk-triage regime.

explicit: disciplinary difficulty is systematic rather than model-specific, and several apparent ranking failures in Task 3 are partly attributable to incomplete permutations rather than wholly uninformative transfer lists[cite: 287].

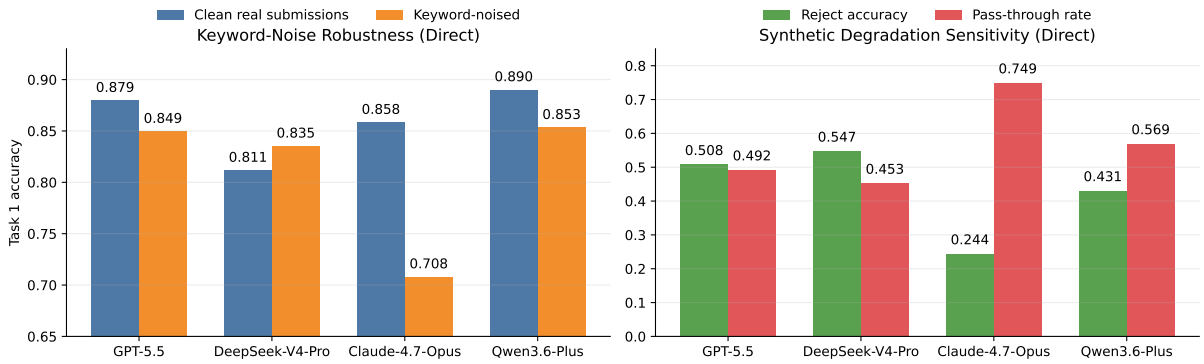

Figure 9: Supplementary robustness summaries derived from the finalized direct-prompt runs. Left: keyword-noise robustness on published-at-target instances. Right: synthetic-degradation sensitivity.

Prompt	Setting	T1 Acc.	Pub.-Target	Cal.-Surplus	Cal.-Deficit	Scope-Mis.	Joint	MRR	FPR	FNR	ECE (clip.)
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 7: GPT-5.5 send-prior sensitivity. The intervention tells the model that roughly 90% of papers in the batch are plausible send cases.

Model	textbfPrompt	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	0.741	0.765	0.599	0.764	0.516	0.845	0.707
DeepSeek-V4-Pro	Direct	0.859	0.631	0.760	0.549	0.784	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 8: Per-domain decision quality across the eight main runs. Columns report Task 1 accuracy and Task 2 joint editorial accuracy within each broad area.

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 9: Per-domain transfer ranking across the eight main runs. Columns report candidate-projected MRR and NDCG@5 within each broad area.

C.4 Scenario-wise breakdowns

Tables 11–14 unpack the main benchmark by editorial scenario[cite: 296]. They support the main-text claim that the benchmark difficulty is highly asymmetric: Published-at-Target and Scope-Mismatch remain comparatively easy, whereas Caliber-Surplus is difficult mainly because send-side caliber discrimination is weak and Caliber-Deficit is difficult because models frequently miss within-field reject cases[cite: 297]. We therefore keep the compact summary in Table 3 of the main paper and place the full scenario tables here, after the more explanatory appendix analyses above[cite: 298].

Model	Prompt	RankOut	ValidPerm	Strict MRR	MRR (Valid Only)
GPT-5.5	Direct	0.780	0.779	0.523	0.671
GPT-5.5	Rationale-first	0.783	0.783	0.524	0.669
DeepSeek-V4-Pro	Direct	0.808	0.648	0.373	0.576
DeepSeek-V4-Pro	Rationale-first	0.820	0.628	0.363	0.579
Claude-4.7-Opus	Direct	0.759	0.759	0.447	0.588
Claude-4.7-Opus	Rationale-first	0.776	0.776	0.460	0.593
Qwen3.6-Plus	Direct	0.699	0.448	0.269	0.601
Qwen3.6-Plus	Rationale-first	0.724	0.435	0.261	0.600

Table 10: Task 3 format-compliance diagnostics for the eight main runs. The main text uses candidate-projected MRR/NDCG@5; this table separates output-format reliability from ranking quality.

Model	Prompt	Task 1			Task 2	
		Acc.	F1	FPR	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

Table 11: Published-at-Target results across the eight main model-prompt configurations. Only the metrics defined for this all-send scenario are shown.

Model	Prompt	Task 1			Task 2	
		Acc.	F1	FPR	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 12: Caliber-Surplus results across the eight main model-prompt configurations. Only the metrics defined for this all-send scenario are shown.

Model	Prompt	Task 1			Task 2		Task 3	
		Acc.	F1	FNR	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	0.657	0.489	0.406	0.448
DeepSeek-V4-Pro	Direct	0.640	0.781	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

Table 13: Caliber-Deficit results across the eight main model-prompt configurations. Only the metrics defined for this reject-side scenario are shown. Task 3 uses candidate-projected MRR and NDCG@5.

Model	Prompt	Task 1			Task 2		Task 3	
		Acc.	F1	FNR	R-F1	Joint	MRR	NDCG@5
GPT-5.5	Direct	0.956	0.978	0.044	0.964	0.930	0.637	0.701
	Rationale-first	0.960	0.980	0.040	0.964	0.930	0.642	0.704
DeepSeek-V4-Pro	Direct	0.976	0.988	0.023	0.980	0.960	0.556	0.631
	Rationale-first	0.980	0.990	0.020	0.982	0.964	0.563	0.633
Claude-4.7-Opus	Direct	0.972	0.986	0.019	0.981	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 14: Scope-Mismatch results across the eight main model-prompt configurations. Only the metrics defined for this reject-side scenario are shown. Task 3 uses candidate-projected MRR and NDCG@5.