MetaBench and FACT: Full-Paper Meta-Review Auto-Generation with Evidence Anchoring

Anonymous ACL submission

Abstract

Meta-review Auto-Generation (MRAG) synthesizes structured meta-reviews from submissions and multi-party review discussions. Existing benchmarks often expose only partial paper context, making paper-specific claims hard to verify and reviewer conflicts hard to adjudicate. We introduce MetaBench, a full-paper MRAG benchmark from OpenReview that pairs submission PDFs with review threads, meta-reviews, and decision metadata. We also propose FACT, a fact-check-centered prompting pipeline that converts PDFs into paper facts, distills discussions into issue artifacts, and applies an evidence-anchoring gate with return-to-source retrieval when verification fails. Across multiple backbones, claim-level faithfulness checks, decision F1, and human pairwise judgments on disagreement-heavy cases, FACT's main benefit is grounded decision-rationale generation rather than large surface-overlap gains alone.

1 Introduction

As peer-review infrastructure scales, decision-critical artifacts—especially meta-reviews—must remain complete, consistent, and grounded in the submitted paper. Major venues are also beginning to test and regulate AI assistance around peer review, from NeurIPS' LLM checklist assistant to ACL/ICLR policies governing generative assistance in reviews and meta-reviews (Goldberg et al., 2024; Association for Computational Linguistics, 2025; ICLR Conference, 2025b). This need is amplified by growing submission volumes at top ML/NLP venues and by recent failures in review infrastructure that exposed the fragility of nominally double-blind workflows (ICLR Conference, 2025a). Meta-reviewing is not merely "summarizing reviews": conference guidelines frame it as adjudication and decision justification, requiring area chairs to explain how reviews, rebuttals, and discussions support the final recommendation, especially when reviewers disagree (NeurIPS, 2025a,b). Thus, a faithful meta-review must be grounded in both the review thread and the paper itself.

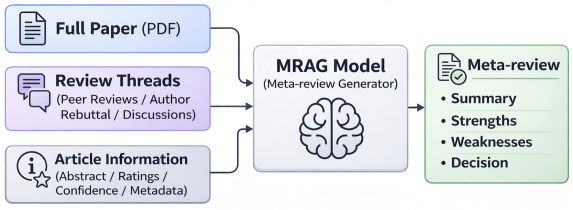

Figure 1: **MRAG task overview.** Inputs include the full paper, review threads, and auxiliary article information; the output is a structured meta-review.

We study **Meta-Review Auto-Generation (MRAG)**: given the full paper, review threads, and auxiliary article information, generate a structured meta-review. The central risk is twofold: automated systems may produce unsupported paper-specific claims, or they may over-weight aggregate ratings while overlooking unresolved reviewer-author conflicts.

Prior MRAG datasets and prompting methods often treat the task as synthesis over reviews and discussions with limited paper context, commonly the abstract (Li et al., 2023; Zeng et al., 2024; Shen et al., 2022). The field still lacks a benchmark and methodology that systematically incorporate the full paper for evidence-grounded verification and conflict-aware decision justification.

To address these gaps, we make the following contributions:

- **MetaBench: a full-paper MRAG benchmark from OpenReview.** We construct standardized instances that include the submission PDF, review threads, meta-reviews, and decision-related metadata across multiple venue-year units, enabling paper-grounded evaluation.

- **FACT: a fact-check-centered prompting pipeline.** We introduce a lightweight workflow

that separates paper facts from thread issues, drafts with role-specific conditioning, and uses paper-grounded evidence anchoring as the only iterative verification gate. FACT is a procedural contribution for full-paper MRAG rather than a new model architecture: it makes paper-specific claims, review-thread issues, and decision-critical assertions separately inspectable.

2 Related Work

2.1 Datasets and Task Formulations for Meta-Review Generation

Recent datasets standardize MRAG with paired review-process inputs and gold meta-reviews. MReD adds sentence-level rhetorical labels for structure-controllable generation (Shen et al., 2022); PeerSum preserves hierarchical review conversations (Li et al., 2023); and ORSUM scales scientific opinion summarization across venues while noting that reference overlap alone is insufficient (Zeng et al., 2024). However, these resources predominantly provide abstracts rather than full paper texts, limiting direct verification of generated factual claims.

2.2 Meta-Review Auto-Generation Methods

Early MRAG systems use supervised, pipeline, or decision-aware generation, including extract-then-refine architectures (Bhatia et al., 2020), decision-aware encoders (Kumar et al., 2021), discourse-label control (Shen et al., 2022), and relationship-aware modeling over review threads (Li et al., 2023). Recent LLM-based work instead emphasizes prompting and assistance: structured prompts improve synthesis across expert perspectives (Hossain et al., 2025), and human-in-the-loop systems such as MetaWriter support draft writing while keeping reviewers in control (Sun et al., 2024).

The closest line of work is iterative or conflict-aware prompting. ORSUM introduces checklist-guided iterative prompting (CGI^2) for completeness and consistency (Zeng et al., 2024), while the Cognitive Alignment Framework models conflicting reviewer opinions through a reasoning pipeline (Chen et al., 2025). We compare against representative prompt variants including Vanilla, an instruction-only prompting baseline, and CGI^2. Unlike these review-thread-centered methods, FACT uses full-paper claim anchoring as an explicit verification gate, drawing on broader agentic revision ideas such as ReAct, Reflexion, and Self-Refine (Yao et al., 2023; Shinn et al., 2023; Madaan et al., 2023).

Algorithm 1 MetaBench Construction Workflow

Require: OpenReview submissions with PDF, official reviews, meta-review, and decision
Ensure: Standardized corpus $\mathcal{D}_{all}$ and benchmark subset $\mathcal{D}_{bench}$

▷ *venue and track restriction*
$\mathcal{U}_{sel} \leftarrow$ main-track venue-year units from the four dominant venue sources in $\mathcal{P}_{\text{full}}$
▷ *validation and normalization*
$\mathcal{D}_{all} \leftarrow \emptyset$
for $u \in \mathcal{U}_{sel}$ **do**
  for $p \in S_u$ **do**
    $\mathcal{I}_p \leftarrow \phi_{\text{build}}(p)$
    if $\Omega_p(\mathcal{I}_p) = 1$ **then**
      $\hat{\mathcal{I}}_p \leftarrow \phi_{\text{norm}}(\mathcal{I}_p)$
      $\mathcal{D}_{all} \leftarrow \mathcal{D}_{all} \cup \{\hat{\mathcal{I}}_p\}$
    end if
  end for
end for
▷ *curation scoring*
for $\hat{\mathcal{I}}_p \in \mathcal{D}_{all}$ **do**
  $Q_p \leftarrow \Omega_q(\hat{\mathcal{I}}_p)$
end for
▷ *benchmark balancing*
$\mathcal{D}_{bench} \leftarrow \phi_{\text{bal}}(\mathcal{D}_{all}, Q; \rho)$
return $\mathcal{D}_{all}, \mathcal{D}_{bench}$

3 MetaBench Dataset Construction

We construct MetaBench from OpenReview[1] using a three-step curation pipeline (Alg. 1): venue/track restriction, instance validation and normalization, and balanced benchmark selection. We retain both the validated corpus $\mathcal{D}_{all}$ for distributional analysis and the decision-balanced subset $\mathcal{D}_{bench}$ for controlled MRAG evaluation.

3.1 Venue Collection

We first collect submissions that contain the minimum structure required by MRAG: a downloadable submission PDF, official reviews, a meta-review, and final decision metadata. Rebuttals, discussion comments, ratings, and confidence scores are retained when available but are not required for inclusion. This fully structured pool spans 31 conferences, workshops, and journals; Fig. 2 shows its source distribution.

[1] https://openreview.net/

Stage	Instances
Fully structured OpenReview pool	16,742
Selected main-track venue units	12,993
Validated and normalized $\mathcal{D}_{all}$	9,498
Benchmark subset $\mathcal{D}_{bench}$	4,782

Table 1: MetaBench curation flow. The benchmark subset is selected from validated instances for controlled accept/reject evaluation; the larger validated corpus is retained for release and distributional analysis.

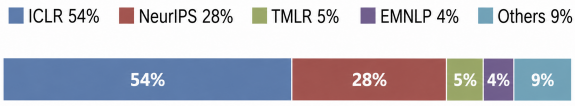

Figure 2: **Source distribution before main-track filtering.** The initial fully structured pool is broad; MetaBench keeps the four dominant sources and their main tracks to obtain a stable, comparable benchmark.

Based on this distribution, we retain the four dominant venue sources (ICLR, NeurIPS, TMLR, and EMNLP) to ensure a stable and sufficiently large benchmark pool. We then keep only their main conference tracks, excluding workshops and other non-main tracks, yielding 16 venue-year units and 12,993 submissions in total.

Table 1 summarizes the main curation flow, with detailed venue-year coverage in Appendix A.

3.2 Paper Structuring and Validation

For each selected venue unit, we convert each submission into a unified instance $\mathcal{I}_p = \phi_{\text{build}}(p)$:

$$\mathcal{I}_p = \left(\text{pdf}_p, a_p, R_p, m_p, B_p, C_p, d_p, \text{rating}_p\right),$$

where pdf_p is the submission PDF (and extracted text when available), a_p is the abstract, R_p are official reviews, m_p is the meta-review, B_p are rebuttals, C_p are discussion comments, d_p is the final decision, and rating_p contains score/confidence metadata when provided. We then apply $\Omega_p(\mathcal{I}_p)$ to remove instances with missing required fields, non-downloadable PDFs, empty meta-reviews, or failed text extraction. For valid PDFs, we preserve section headings and paragraph order, retain table/figure captions as textual proxies, and record lightweight paragraph identifiers for later evidence retrieval. Finally, $\phi_{\text{norm}}(\cdot)$ normalizes formatting artifacts, maps heterogeneous decision labels into accept/reject groups, and organizes the discussion into role-tagged threads. All validated instances are aggregated into $\mathcal{D}_{all}$.

3.3 Curation Score for Meta-Review Quality

We use an automatic score $Q_p = \Omega_q(\hat{\mathcal{I}}_p)$ only as a benchmark-curation proxy, not as an evaluation metric for generated meta-reviews. It prioritizes instances whose human-written meta-reviews are suitable references: structurally complete and covering salient review points. Concretely, $Q_p = \alpha S_{\text{st}}(p) + \beta S_{\text{cov}}(p)$ with fixed weights $\alpha = 0.8, \beta = 0.2$. S_{st} checks whether the meta-review contains summary, strengths, weaknesses, and decision-related content with a minimum-length regularizer; S_{cov} measures whether salient review sentences, selected with sentence embeddings and MMR-style diversity, are covered by the meta-review. We use Q_p for ranking and percentile binning during subset construction. Implementation defaults and weight sensitivity are reported in Appendices E and C.

Human validation of the scoring function. We conduct a blinded human validation study to test whether Q_p correlates with meta-review quality. We sample 1,000 instances from 5 venue units (ICLR 2023–2025; NeurIPS 2024–2025), 200 per unit, shuffle them, and collect binary GOOD/BAD judgments from three annotators who are blind to Q_p. Each annotator labels a disjoint portion plus a shared overlap set used to estimate agreement. We intentionally use coarse binary validation rather than fine-grained scalar ratings, since our goal is to test whether Q_p aligns with broad judgments of benchmark suitability. Annotators mark an instance as GOOD when the meta-review is structurally complete, covers the main review points, and remains well grounded in the review/rebuttal thread; otherwise it is marked as BAD. The full annotation rubric and overlap protocol are provided in Appendix B. Fig. 3 reports the human GOOD rate across Q_p percentile bins, showing an overall increasing trend.

3.4 Balanced Benchmark Subset Construction

To control for varying acceptance rates across venue units, we construct a balanced benchmark subset $\mathcal{D}_{bench} = \phi_{\text{bal}}(\mathcal{D}_{all}, Q; \rho)$. For each venue unit u, let $N_{\min} = \min(\#\text{ACCEPT}, \#\text{REJECT})$ after validation and set the per-class target:

$$N^\star = \lfloor \rho \cdot N_{\min} \rfloor .$$

We then select the top-$N^\star$ instances by Q_p from each decision group and merge them to obtain a balanced subset.

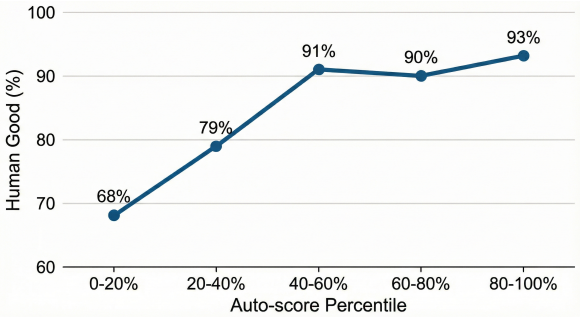

Figure 3: **Human validation of automatic meta-review scoring.** Human GOOD rate (y-axis) versus automatic score percentile bins (x-axis) on 1,000 blinded samples (5 venue units $\times$ 200). Each bin contains 200 samples. The overall increasing relationship supports using Q_p as a quality ranking signal for benchmark curation.

Method	Paper	Issues	Anchor	R2S
CGI[2]	No	Partial	No	No
CAF	No	Yes	No	No
FACT	Yes	Yes	Yes	Yes

Table 2: Procedural comparison. R2S denotes return-to-source revision. CGI[2] is our implemented no-full-paper iterative baseline; CAF is related conflict-aware work.

We set $\rho = 0.5$ in MetaBench. As shown in Fig. 3, quality improves substantially from the top 50% onwards, motivating $\rho = 0.5$ as a practical trade-off between benchmark reliability and dataset size. After balancing, the final benchmark subset contains 4,782 submissions. This balancing is intended for controlled method comparison, not to reproduce natural venue acceptance rates; we therefore retain $\mathcal{D}_{all}$ and report venue-level statistics separately.

4 FACT: Fact-Check-Centered Prompting Pipeline for Meta-Review Generation

As illustrated in Fig. 4, FACT separates long MRAG inputs into paper facts and review-thread issues, drafts from a structured plan, and verifies claims before final style alignment (full prompts in Appendix F.3). Its key difference from checklist-style or cognitive-alignment prompting is where iteration occurs: FACT revises only claims that fail evidence anchoring, rather than repeatedly refining the whole draft. This makes the main failure modes measurable through claim support, false consensus, and false resolution rates, complementing reference-overlap metrics. Table 2 summarizes this procedural distinction.

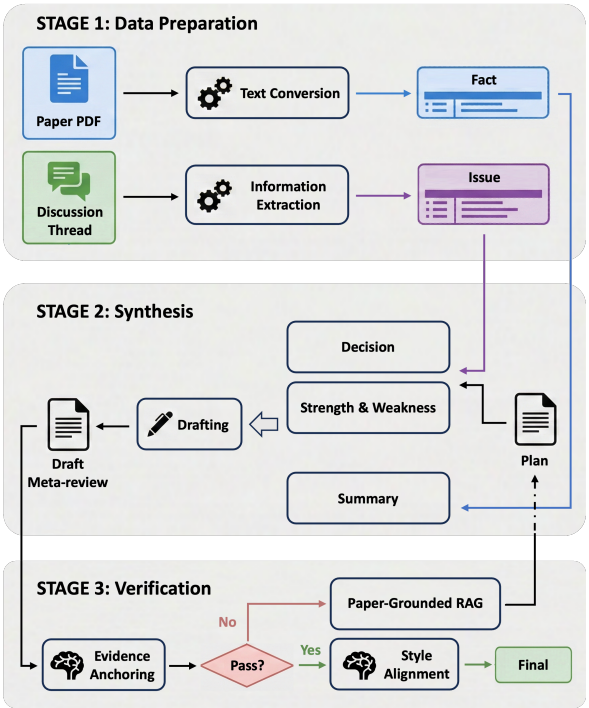

Figure 4: **FACT prompting pipeline.** FACT first creates paper-fact and issue-card artifacts, then drafts from a structured plan. The only iterative step is evidence anchoring: unsupported paper-specific or decision-critical claims trigger return-to-source retrieval and revision. Style alignment is final and non-iterative.

4.1 Data Preparation

This stage converts long inputs into two compact artifacts: paper facts and thread-level issues.

Paper PDF to Fact Representation. We convert the submission PDF into a normalized fact document $\mathcal{F} = \phi_{\text{text}}(P)$ by preserving section headings and paragraph order, replacing tables/figures with captions, and assigning paragraph identifiers for later retrieval. $\mathcal{F}$ grounds both Summary drafting and return-to-source verification.

Discussion Thread to Issue Artifacts. We distill the review/rebuttal/comments thread T into issue cards $\{\mathcal{C}_j\} = \phi_{\text{issue}}(T)$. Each card records the concern, reviewer stance and severity, author response, resolution status, and retrieval keywords. These cards constrain what the meta-review must address without concatenating the full discussion.

4.2 Synthesis

This stage drafts from role-separated inputs rather than from raw concatenation. We first build a plan $\Pi = \phi_{\text{plan}}(\mathcal{F}, \{\mathcal{C}_j\}, \mathbf{s})$, where $\mathbf{s}$ denotes decision-related signals such as ratings, confidences, and

consensus cues. The plan specifies the meta-review template, must-mention issues, conflict map, and decision-rationale sketch. *Summary* draws primarily from paper facts; *Strengths/Weaknesses* from issue cards; and *Decision/Rationale* from unresolved issues and score signals, but not scores alone. The generator $\phi_{\text{write}}(\Pi)$ produces an initial draft $\tilde{m}^{(0)}$ and defers verification to Stage 3.

4.3 Verification

This stage verifies the draft before style alignment. The gate targets two errors: unsupported paper-specific claims and decision rationales that ignore unresolved conflicts or misrepresent reviewer/author positions.

Evidence Anchoring. We apply $\Omega_{\text{anchor}}(\tilde{m}, \mathcal{F}, \{\mathcal{C}_j\})$ to two claim classes. Paper-specific factual claims must be supported by retrieved spans in $\mathcal{F}$. Decision-critical claims must be consistent with issue cards: unresolved issues cannot be marked as resolved, and consensus cannot be claimed when reviewer stances conflict. A claim-identification prompt extracts these statements (Appendix F.3.4); the draft passes only when all flagged claims are supported.

Return-to-Source Retrieval and Revision. When anchoring fails, FACT retrieves an evidence pack $\mathcal{G} = \{(u, \text{id}(u), \text{span})\}$ over $\mathcal{F}$ using hybrid lexical/dense retrieval and re-ranking (Robertson and Zaragoza, 2009; Karpukhin et al., 2020; Nogueira and Cho, 2019). The revision step either anchors each failed claim to supporting evidence or rewrites it conservatively. This return-to-source step is the only iterative loop and is bounded by K_{iter}.

Verification defaults. Unless otherwise stated, we use: BM25 top-k_{lex}=20, dense top-k_{dense}=20, fused top-k_{cand}=30, re-ranked top-k_{sup}=5, support threshold τ_{sup}=0.70, and loop budget K_{iter}=2. These settings are fixed for FACT and all ablations for fair comparison.

Style Alignment. After evidence anchoring passes, we apply a final style alignment step to match venue conventions (tone, hedging, section ordering, and typical decision phrasing). We retrieve a small set of exemplar meta-reviews $\mathcal{E}_u$ for venue-year unit u and perform surface-level rewriting $\phi_{\text{style}}(\tilde{m}, \mathcal{E}_u)$. Crucially, this step does not trigger further verification or iteration; it only adjusts phrasing/organization while preserving the anchored content.

5 Experiments

5.1 Experimental Setup

We evaluate MRAG under four LLM backbones: GPT-5-mini (OpenAI, 2025), Gemini 2.5 Flash (Google DeepMind, 2025), Claude-Sonnet-4.5, and Qwen3.5. We keep the output format constraints identical across models and baselines to avoid confounding effects from formatting differences. All results are averaged over 3 random seeds. For claim-level faithfulness, conflict-heavy analysis, and human evaluation, we use targeted subsets stratified by venue source, decision label, curation-score percentile, and reviewer-disagreement signals.

Evaluation Metrics For compact comparison, we report two representative reference-overlap metrics: ROUGE-L (Lin, 2004) and BERTScore-F1 (Zhang et al., 2020). Because these metrics do not directly measure grounding or decision justification, we add task-specific metrics inspired by fine-grained factuality and summarization evaluation (Min et al., 2023; Song et al., 2024). We decompose each generated meta-review into two claim classes: *paper-specific claims*, which must be supported by retrieved paper-fact spans, and *decision-critical claims*, which must be consistent with issue cards and reviewer stances. We then report: (i) supported and unsupported paper-claim rates; (ii) false consensus rate, where the output claims reviewer agreement despite conflicting stances; (iii) false resolved rate, where an unresolved or partially resolved issue is described as resolved; (iv) major-issue coverage; and (v) decision macro-F1 after extracting the accept/reject polarity from the Decision/Rationale section.

Baselines. We compare against two groups of prompting methods. *W/o full-paper context:* baselines that operate on reviews and discussions without access to full paper facts, implemented as prompt-based approximations on MetaBench: Vanilla (Zeng et al., 2024), an instruction-only prompting baseline adapted from prior prompt-design studies (Hossain et al., 2025), and Checklist-Guided Iterative prompting (CGI^2) (Zeng et al., 2024). *W/ full-paper context:* our methods that additionally incorporate the full paper fact representation (templates in Appendix F.1 and F.2): (1)

Category	Method	GPT-5-mini		Gemini-2.5-Flash		Claude-Sonnet-4.5		Qwen3.5	
		R-L	B-F1	R-L	B-F1	R-L	B-F1	R-L	B-F1
W/o full-paper context	Vanilla	20.14	82.40	20.92	83.23	22.10	84.35	21.86	84.10
	InstructPrompt	21.48	83.48	22.06	84.31	22.74	84.78	22.43	84.55
	CGI^2	21.89	83.97	22.53	84.93	23.18	85.45	22.96	84.92
W/ full-paper context	One-shot	18.09	78.40	19.85	81.13	20.55	82.19	20.25	81.80
	ICL	20.65	79.87	21.31	84.03	22.18	84.92	21.90	84.45
	FACT	22.25	84.01	22.41	85.96	23.54	86.03	23.31	85.48

Table 3: Overlap-metric comparison on MetaBench. *W/o full-paper context* methods use only reviews/discussions without full paper facts. Best results per backbone are in red.

Backbone	Temp.	Max tokens
GPT-5-mini	0.2	520
Gemini-2.5-Flash	0.2	520
Claude-Sonnet-4.5	0.2	520
Qwen3.5	0.2	520

Table 4: Decoding settings used for all backbone comparisons. The same output format constraints are applied to every method and backbone.

One-shot: a single prompt that directly requests a meta-review from MRAG inputs; (2) ICL: one-shot with venue-year exemplars $\mathcal{E}_u$ prepended as in-context demonstrations to induce style alignment, without evidence verification; (3) FACT: our full pipeline with paper facts, issue extraction, and evidence anchoring (Appendix F.3). We do not treat CAF as an implemented baseline in the current revision because its published formulation assumes a different conflict-alignment protocol; instead, Table 2 clarifies the procedural differences and we prioritize CGI^2 as the closest reproducible iterative baseline.

5.2 Results

Table 3 shows that FACT is competitive with the strongest prompt baselines across backbones. Among the methods without full-paper context, stronger prompt-engineering baselines such as CGI^2 approach FACT and surpass it on a small number of recall-oriented metrics, indicating that iterative prompting alone can improve coverage. The margins on ROUGE and BERTScore are modest, so we avoid interpreting them as sufficient evidence of practical superiority. Instead, the remaining analyses test whether paper grounding and evidence anchoring reduce the specific area-chair failure modes that motivated FACT. ICL (venue exemplars) consistently improves over One-shot, primarily in organization and style alignment, while the Gemini-2.5-Flash backbone shows the same overall pattern with slightly smaller gaps between methods.

Stronger-backbone analysis. The Claude-Sonnet-4.5 and Qwen3.5 columns show that the trend is not limited to the two efficient backbones. The pattern is not a large overlap-metric jump, but a stable FACT advantage over CGI^2.

Targeted faithfulness and decision metrics. Table 5 reports paper-grounded and decision-critical failure modes on the targeted evaluation subset. Compared with ICL and CGI^2, FACT reduces unsupported paper claims and prevents false resolution or false consensus statements, especially when reviewers disagree.

5.3 Ablations

Table 6 removes one FACT component at a time. Evidence anchoring is the most important component under overlap metrics: removing it causes the largest drops on both backbones. Removing paper facts also degrades performance, while issue distillation shows smaller overlap-metric effects, consistent with its role as a control and organization mechanism rather than a lexical-overlap booster.

Paper fact representation: caption-only vs. caption+image. A known limitation of our paper fact representation is that tables and figures are replaced by their captions, discarding visual and numerical content. To quantify this impact, we compare two variants of FACT on GPT-5-mini: (1) *caption-only*, the default setting used throughout; (2) *caption+image*, where we use a VLM-based paper parsing pipeline to extract table/figure images from the PDF and feed them alongside the caption text via the model's vision input. Table 7 reports the results. The two representations yield near-identical scores, suggesting that captions already capture the information most relevant to meta-review generation and that the additional visual content provides

Method	Supported↑	Unsupported↓	False cons.↓	False resol.↓	Issue cov.↑	Decision F1↑
ICL	74.2	25.8	18.6	15.1	69.7	76.9
CGI^2	78.6	21.4	14.8	12.5	72.1	78.4
FACT w/o Evidence	76.0	24.0	16.7	14.3	70.9	77.5
FACT	**88.9**	**11.1**	**6.2**	**5.4**	**80.6**	**82.8**

Table 5: Task-specific evaluation for MRAG on the targeted evaluation subset. Supported/unsupported are paper-specific claim rates; false consensus and false resolution measure decision-critical errors derived from issue cards. All values are percentages except Decision F1.

Setting	R-L↑	B-F1↑
`GPT-5-mini / Gemini-2.5-Flash`		
FACT (full)	22.25 / 22.41	84.01 / 85.96
w/o Issues	+0.2 / −0.1	−0.1 / −0.3
w/o Facts	−0.4 / −0.3	−0.5 / −0.6
w/o Evidence	−0.5 / −0.7	−0.9 / −1.2

Table 6: Ablation study on GPT-5-mini and Gemini-2.5-Flash. The first row reports full FACT; subsequent rows report metric changes after removing one component.

Fact repr.	R-L↑	B-F1↑
`GPT-5-mini`		
caption-only (default)	22.25	84.01
caption+image (VLM)	22.18	84.09

Table 7: Effect of paper fact representation on FACT (GPT-5-mini). Caption-only and caption+image variants produce near-identical overlap scores, suggesting that most information used by current MRAG outputs is already captured by captions.

marginal benefit under the current evaluation metrics.

5.4 Conflict-Heavy Subset Analysis

FACT is designed for cases where the area chair must adjudicate disagreement rather than summarize a clear consensus. We therefore construct a conflict-heavy subset using three signals: (i) high reviewer-score variance, (ii) explicit disagreement markers in the discussion thread, and (iii) borderline decisions where accept/reject recommendations are split or close. Table 8 reports the comparison on this subset. FACT's advantage is larger on false consensus and false resolution than on BERTScore-F1, matching the intended role of evidence anchoring.

5.5 Human Evaluation

We additionally conduct a method-blinded pairwise study with three annotators on borderline split-decision submissions. Annotators compare

Method	B-F1	False cons.↓	False resol.↓	Issue cov.↑
ICL	84.6	24.0	18.5	68.3
CGI^2	85.0	18.2	14.6	71.5
FACT	**85.7**	**7.5**	**6.8**	**81.9**

Table 8: Conflict-heavy subset analysis. FACT's gains are concentrated on disagreement and resolution errors rather than surface overlap.

Pair	Cases	Ann.	Win	Tie	Lose
FACT vs ICL	40	3	61.7	25.8	12.5
FACT vs CGI^2	40	3	49.2	28.3	22.5

Table 9: Method-blinded human pairwise evaluation on borderline split-decision cases. Win/Tie/Lose are percentages from FACT's perspective, aggregated over three annotators.

outputs along four criteria: overall usefulness to an area chair, faithfulness to the paper, handling of reviewer disagreement, and decision-rationale quality. Ties are allowed when neither output is clearly better. Table 9 reports the expanded three-annotator version of the rebuttal study. FACT is preferred substantially over ICL and wins more often than it loses against CGI^2, although the CGI^2 comparison remains close once ties are included.

6 Case Study: Reviewer Conflict and Evidence Anchoring

We compare ICL and FACT on a borderline-accepted paper with split reviewer opinions to illustrate how evidence anchoring changes the generated decision rationale. Fig. 5 shows the human meta-review (top), ICL output (middle), and FACT output (bottom), with key phrases highlighted.

The comparison reveals three key differences between ICL and FACT:

Consensus vs. disagreement representation. ICL produces statements like "*reviewers consistently highlighted the originality*," implying uniform agreement. However, the human meta-review reveals a 2-vs-2 split: two reviewers praised novelty

Ground Truth (Reviewer Analysis Only)

[*Summary omitted*] **Two reviewers praise the novelty and the stronger rebuttal**, and recommend accept or borderline accept. **Two others question practical need versus multimodal models, O(n^2) cost and storage, and guidance on metric choice; one barely engaged**. I recommend accept, but this is borderline from my perspective. The idea is fresh and now better supported by added baselines (linear projection, CCA) and runtime/large-scale results. **Scaling and metric-choice risks remain**, but approximations checks help, and the camera-ready should expand limitations and fix notation. With split reviews and moderate confidence, I still see clear value for the community.

ICL Output

[*Summary omitted*] The work is highly novel, offering a rare combination of philosophical motivation, rigorous formalization, and broad empirical validation. **Reviewers consistently highlighted the originality and clarity**, and the rebuttal convincingly addressed concerns about scalability through landmark-based approximations, added downstream tasks, robustness analyses, and clearer limitation discussions. Remaining issues primarily concern the quadratic complexity of the exact formulation, the lack of theoretical guidance for cost function choice, and limited qualitative interpretability. However, **these are acknowledged and partially mitigated, and they do not outweigh the paper's contributions**. I recommend acceptance.

FACT Output

[*Summary omitted*] **Reviewer opinions are mixed but generally positive**. Two reviewers emphasize the novelty, conceptual clarity, and improved empirical support after rebuttal, recommending acceptance or borderline acceptance. **The more critical reviewers raise concerns about practical necessity relative to existing multimodal models, the quadratic time and storage costs, and the lack of principled guidance for metric selection, with one reviewer engaging only minimally in discussion**. My recommendation is accept, though borderline. The idea is genuinely fresh and better substantiated after the rebuttal, including added baselines (e.g., linear projection, CCA) and clearer runtime and large-scale evidence. **Scalability and metric-choice risks remain, but the proposed approximations and sensitivity analyses mitigate the most serious concerns**. With a strengthened limitations section and minor presentation fixes in the camera-ready, I believe the work offers clear value to the community.

Figure 5: **Case study on a borderline paper with reviewer disagreement.** The figure compares the human meta-review, ICL output, and FACT output; highlighted phrases show how each output characterizes reviewer opinions and residual risks.

while two others raised concerns about practical necessity and scalability. FACT correctly captures this split by explicitly distinguishing supportive from critical reviewers, including the detail that "*one reviewer engaging only minimally*"—a nuance the area chair noted when weighing opinions.

Specificity in residual risk acknowledgment. Both methods mention unresolved concerns, but with different granularity. ICL uses vague phrasing ("*these are acknowledged and partially mitigated*"), while FACT names specific risks and mitigations ("*scalability and metric-choice risks remain, but the proposed approximations and sensitivity analyses mitigate the most serious concerns*"). FACT's formulation aligns closely with the human meta-review.

Evidence grounding. ICL propagates reviewer claims without verification—e.g., if a reviewer states "*the rebuttal addressed scalability*," ICL accepts this without checking whether the paper contains the cited experiments. FACT's evidence-anchoring gate forces cross-referencing against issue cards, preventing fabrication of resolution status for unresolved issues.

In summary, ICL improves organization over one-shot prompting but can still produce consensus-biased outputs. FACT's issue-card distillation and evidence anchoring improve reviewer-level fidelity, capturing the split opinions and unresolved tensions that area chairs must navigate.

7 Conclusion

We introduced MetaBench, a full-paper MRAG benchmark from OpenReview with 4,782 decision-balanced submissions, and a prompting pipeline that separates paper facts from review-thread issues and verifies decision-critical claims before style alignment. Across multiple backbones, FACT is competitive on overlap metrics, while targeted faithfulness metrics and human pairwise judgments show clearer gains in grounded decision-rationale generation. Ablations identify evidence anchoring as the most important component, and the case study illustrates how FACT preserves reviewer disagreement and residual risks that consensus-biased summaries can obscure. These findings suggest that full-paper context is most useful when it is connected to explicit verification rather than simply appended to a longer prompt. MetaBench provides a testbed for this setting by pairing review-process artifacts with the underlying submissions and decision metadata. This makes MRAG outputs easier to inspect, compare, and correct. Overall, MetaBench and FACT support more transparent, evidence-aware meta-review drafting while keeping human area chairs in control.

8 Limitations

Our work has several limitations. First, MetaBench is constructed from OpenReview-hosted ML venues and may not generalize to other academic fields, non-OpenReview platforms, or single-blind review workflows. Second, although we add claim-level, decision-critical, and human evaluations, these metrics are still approximations of area-chair judgment and depend on the quality of issue-card extraction. Third, although our ablation shows that replacing tables and figures with captions has minimal impact under current metrics, this proxy may lose critical numerical or visual detail in more table-intensive domains; issue distillation may also compress nuanced reviewer discussions. Finally, results may vary across different model versions and decoding settings, and we provide prompts in the appendix to support reproducibility.

9 Potential Risks

We identify several potential risks associated with MRAG systems. Generated meta-reviews may appear authoritative while missing critical reviewer concerns or mischaracterizing disagreements. Therefore, FACT is intended for assistive drafting rather than replacing human area chairs. Additionally, models trained on historical peer-review data may inherit biases present in past reviews. Review threads may also contain identifying information, requiring careful anonymization before any data release or deployment.

10 Use of AI Assistants

We used AI assistants for language polishing, table/figure formatting, and experiment code generation. All outputs were reviewed and edited by the authors, who remain solely responsible for the intellectual contributions.

References

Association for Computational Linguistics. 2025. ACL policy on publication ethics: Generative assistance in peer review. https://www.aclweb.org/adminwiki/index.php?title=ACL_Policy_on_Publication_Ethics.

Chaitanya Bhatia, Tribikram Pradhan, and Sukomal Pal. 2020. Metagen: An academic meta-review generation system. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1653–1656.

Wei Chen, Han Ding, Meng Yuan, Zhao Zhang, Deqing Wang, and Fuzhen Zhuang. 2025. Bridging social psychology and llm reasoning: Conflict-aware meta-review generation via cognitive alignment. *Preprint*, arXiv:2503.13879.

Alexander Goldberg, Ihsan Ullah, Thanh Gia Hieu Khuong, Benedictus Kent Rachmat, Zhen Xu, Isabelle Guyon, and Nihar B. Shah. 2024. Usefulness of LLMs as an author checklist assistant for scientific papers: NeurIPS'24 experiment. *Preprint*, arXiv:2411.03417.

Google DeepMind. 2025. Gemini 2.5 flash: Our best model yet for price and performance. https://deepmind.google/technologies/gemini/flash/.

Eftekhar Hossain, Sanjeev Kumar Sinha, Naman Bansal, Alex Knipper, Souvika Sarkar, John Salvador, Yash Mahajan, Sri 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. *Preprint*, arXiv:2402.15589.

ICLR Conference. 2025a. ICLR 2026 response to security incident. https://blog.iclr.cc/2025/12/03/iclr-2026-response-to-security-incident/.

ICLR Conference. 2025b. Policies on large language model usage at ICLR 2026. https://blog.iclr.cc/2025/08/26/policies-on-large-language-model-usage-at-iclr-2026/.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Asheesh Kumar, Tirthankar Ghosal, and Asif Ekbal. 2021. A deep neural architecture for decision-aware meta-review generation. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 222–225.

Miao Li, Eduard Hovy, and Jey Lau. 2023. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of the Association for Computational Linguistics: EMNLP*

2023, pages 7089–7112. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

NeurIPS. 2025a. NeurIPS 2025 area chair guidelines. https://neurips.cc/Conferences/2025/AC-Guidelines.

NeurIPS. 2025b. NeurIPS 2025 reviewer and AC/SAC guidelines. https://neurips.cc/Conferences/2025/ReviewerGuidelines.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

OpenAI. 2025. GPT-5 mini system card. https://openai.com/index/gpt-5-mini-system-card/.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022. MReD: A meta-review dataset for structure-controllable text generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2521–2535. Association for Computational Linguistics.

Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. *Preprint*, arXiv:2303.11366.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained summarization evaluation using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922, Bangkok, Thailand. Association for Computational Linguistics.

Lu Sun, Stone Tao, Junjie Hu, and Steven P. Dow. 2024. Metawriter: Exploring the potential and perils of ai writing support in scientific peer review. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Qi Zeng, Mankeerat Sidhu, Ansel Blume, Hou Pong Chan, Lu Wang, and Heng Ji. 2024. Scientific opinion summarization: Paper meta-review generation dataset, methods, and evaluation. *Preprint*, arXiv:2305.14647.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Venue-Year Coverage

The retained venue-year coverage is as follows.
ICLR: 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025.
NeurIPS: 2021, 2022, 2023, 2024, 2025.
TMLR: retained as a single venue without year-specific partition.
EMNLP: 2023.

Table 10 gives the corpus statistics used in the current benchmark construction.

B Human Validation Rubric

The binary human validation is intended to test whether the automatic score Q_p aligns with coarse human judgments of meta-review quality, rather than to provide fine-grained meta-review scoring. Annotators are shown the reviews, rebuttal/discussion thread when available, and human-written meta-review, but not the automatic score or percentile bin. Annotators assign a single label to each instance using the following rubric: **Good:** the meta-review has the expected structure, covers the major issues raised in the reviews, and stays grounded in the review/rebuttal thread without obvious omissions or distortions. **Bad:** the meta-review is missing key structural elements, omits salient review concerns, or is insufficiently grounded in the review/rebuttal content. Fifteen percent of items are double-labeled to estimate agreement; disagreements are adjudicated only for analysis and do not change the automatic score.

C Curation-Score Sensitivity

Because Q_p is used only to curate the benchmark subset, not to evaluate FACT, we test whether the selected subset is stable under different fixed weights. Table 11 reports a sensitivity analysis over α and β. Moderate changes preserve most selected instances and downstream method ranking, while very low structural weight admits more incomplete meta-reviews.

D Generated-Output Human Evaluation Protocol

The generated-output study uses three annotators who are blind to method names and output order. Each item contains the paper abstract, structured paper facts, review thread, decision metadata, and two generated meta-reviews. Annotators choose A, B, or TIE for four dimensions: (1) overall usefulness to an area chair, (2) faithfulness to paper facts, (3) handling of reviewer disagreement and unresolved concerns, and (4) quality of the final decision rationale. The main table reports the overall preference; dimension-level breakdowns and agreement statistics are used as supporting analysis.

E Implementation Details

Parameter	Default
α, β	$0.8, 0.2$
L_{ref}	100
K_{key}	5
τ (coverage hit)	0.72
$k_{\text{lex}}, k_{\text{dense}}$	$20, 20$
$k_{\text{cand}}, k_{\text{sup}}$	$30, 5$
τ_{sup} (support)	0.70
K_{iter}	2

Table 12: Default settings for scoring and FACT verification.

F Prompt Templates

This appendix provides the prompt templates used in our experiments, including (i) One-shot prompting baselines and (ii) the multi-stage FACT pipeline. All placeholders (e.g., <PAPER_FACTS>) are replaced with instance-specific content.

F.1 One-shot Prompt

```
SYSTEM:
You are an area chair writing a meta-review.
↪ Be concise, balanced, and faithful
to the paper and the review process. Follow
↪ the required section structure.

USER:
You are given three inputs:

[PAPER-FACTS]
<PAPER_FACTS_TEXT>
(Note: tables/figures are represented by their
↪ captions.)

[THREAD]
<PEER_REVIEWS_TEXT>
<AUTHOR_REBUTTAL_TEXT>
<DISCUSSION_TEXT>

[ARTICLE-INFO]
<ABSTRACT_TEXT>
<RATINGS_AND_CONFIDENCE_IF_AVAILABLE>
<OTHER_METADATA_IF_AVAILABLE>

Task: Write a meta-review following the
↪ constraints below.
```

Source	Raw	Validated	Bench	Accept	Reject
ICLR 2017–2025	8,740	6,210	2,404	1,202	1,202
NeurIPS 2021–2025	3,180	2,430	1,642	821	821
TMLR	780	612	498	249	249
EMNLP 2023	293	246	238	119	119
Total	12,993	9,498	4,782	2,391	2,391

Table 10: MetaBench corpus statistics by source. The benchmark subset is decision-balanced within venue units before aggregation.

α, β	Good rate↑	Jaccard↑	FACT B-F1↑
$0.5, 0.5$	78.6	0.82	84.95
$0.7, 0.3$	80.4	0.93	85.01
$0.8, 0.2$	81.2	1.00	85.02
$0.9, 0.1$	80.9	0.96	84.99

Table 11: Sensitivity of MetaBench curation to the fixed quality-score weights. "Good rate" is the human GOOD rate on selected validation samples; "Jaccard" compares selected instances against the default $0.8, 0.2$ subset.

```
Length constraints:
- Target total length: 200-400 words (hard max
↪ 520 words).
- Summary: 2-4 sentences.
- Strengths: 2-4 bullet points.
- Weaknesses: 2-4 bullet points.
- Decision/Rationale: 3-6 sentences.

Formatting constraints (strict):
- Output MUST contain EXACTLY these four
↪ section headings, each on its own line,
  in this exact order:
  Summary:
  Strengths:
  Weaknesses:
  Decision/Rationale:
- Do NOT add numbering, letters, or extra
↪ prefixes to the headings
  (e.g., no "1) Summary", no "(a) Strengths").
- Do NOT add any extra sections (e.g.,
↪ "Overall", "Recommendation", "Confidence").
- Under Strengths/Weaknesses, use plain '-'
↪ bullets only; do not use numbered lists.
- Do NOT repeat the headings elsewhere; do not
↪ add a title line.

Content constraints:
- Summary must describe the paper's
↪ contributions and setup (not reviewer
↪ opinions).
- Strengths/Weaknesses must reflect the review
↪ thread and rebuttal.
- Decision/Rationale must justify the
↪ recommendation and address unresolved major
↪ issues.
- Do not invent experiments, results, or
↪ claims not supported by the inputs.
```

F.2 Style Alignment Prompt

This template is used by FACT after evidence anchoring passes.

```
SYSTEM:
You are rewriting a meta-review to match the
↪ venue style. You MUST preserve
the original meaning and claims. Do not add or
↪ remove any factual statements.

USER:
You are given:
(1) A draft meta-review M (content is correct
↪ and should not be changed in meaning).
(2) A small set of exemplar meta-reviews from
↪ the same venue-year unit.

[EXEMPLARS]
<E1>
<E2>
<E3>

[DRAFT META-REVIEW]
<M>

Task: Rewrite M to better match the exemplars
↪ in tone, phrasing, and organization.

Formatting constraints (strict):
- Output MUST contain EXACTLY these four
↪ section headings, each on its own line,
  in this exact order:
  Summary:
  Strengths:
  Weaknesses:
  Decision/Rationale:
- Do NOT add numbering, letters, or extra
↪ prefixes to the headings.
- Do NOT add any extra sections (e.g.,
↪ "Overall", "Recommendation", "Confidence").
- Under Strengths/Weaknesses, use plain '-'
↪ bullets only; do not use numbered lists.
- Do NOT repeat the headings elsewhere; do not
↪ add a title line.

Hard constraints (must follow):
- Preserve ALL substantive claims and
↪ judgments in M. Do NOT add new claims.
- Do NOT delete any existing claims. Do NOT
↪ change the decision.
- Only adjust wording, sentence structure,
↪ hedging level, and local ordering within
↪ sections.
- Keep the output length within +/-10% of M.
```

F.3 FACT Pipeline Prompts

F.3.1 Issue Extraction

```
SYSTEM:
You extract structured issue cards from review
↪ discussions. Be faithful and concise.

USER:
[THREAD]
<PEER_REVIEWS_TEXT>
<AUTHOR_REBUTTAL_TEXT>
<DISCUSSION_TEXT>

Task: Distill the thread into a list of issue
↪ cards. Each issue card must follow:

- Issue statement:
- Reviewer stance (supporting/opposing;
↪ minor/major):
- Evidence requested:
- Author response summary:
- Resolution status (Resolved /
↪ Partially-Resolved / Unresolved):
- Score signals (rating/confidence changes if
↪ present; else N/A):
- Keywords/queries (for paper lookup):

Rules:
- One card per coherent critique thread.
- Do not include long quotations; summarize.
- If reviewers disagree, mark it explicitly in
↪ the issue statement or stance.
Output in JSON list format.
```

F.3.2 Planning

```
SYSTEM:
You produce a writing plan for a meta-review.
↪ The plan must cover key issues and
↪ conflicts.

USER:
[PAPER-FACTS]
<PAPER_FACTS_TEXT>

[ISSUE CARDS]
<ISSUE_CARDS_JSON>

[ARTICLE-INFO]
<ABSTRACT_TEXT>
<RATINGS_AND_CONFIDENCE_IF_AVAILABLE>

Task: Create a plan Pi with:
1) Template: Summary -> Strengths ->
↪ Weaknesses -> Decision/Rationale
2) Must-mention issues: list issue ids
3) Conflict map: disagreements + current
↪ resolution status
4) Decision rationale sketch: justify without
↪ relying on ratings alone

Output Pi in JSON with explicit fields:
- summary_points
- strengths_points (each linked to issue ids
↪ if applicable)
- weaknesses_points (linked to issue ids)
- conflicts
- decision_rationale_outline
```

F.3.3 Drafting

```
SYSTEM:
Write a meta-review from a given plan. Follow
↪ the required section structure.

USER:
[PLAN Pi]
<PLAN_JSON>

Task: Write the meta-review following the
↪ constraints below.

Length constraints:
- Target total length: 180-260 words (hard max
↪ 320 words).
- Summary: 2-4 sentences.
- Strengths: 2-4 bullet points.
- Weaknesses: 2-4 bullet points.
- Decision/Rationale: 3-6 sentences.

Formatting constraints (strict):
- Output MUST contain EXACTLY these four
↪ section headings, each on its own line,
  in this exact order:
  Summary:
  Strengths:
  Weaknesses:
  Decision/Rationale:
- Do NOT add numbering, letters, or extra
↪ prefixes to the headings.
- Do NOT add any extra sections (e.g.,
↪ "Overall", "Recommendation", "Confidence").
- Under Strengths/Weaknesses, use plain '-'
↪ bullets only; do not use numbered lists.
- Do NOT repeat the headings elsewhere; do not
↪ add a title line.

Content rules:
- Use only what is in the plan. Do not invent
↪ new details.
- Ensure each must-mention issue is addressed
↪ in Strengths/Weaknesses and/or
↪ Decision/Rationale.
```

F.3.4 Evidence Anchoring

```
SYSTEM:
You identify paper-specific and
↪ decision-critical statements that require
↪ support.

USER:
[PAPER-FACTS]
<PAPER_FACTS_TEXT>

[ISSUE CARDS]
<ISSUE_CARDS_JSON>

[DRAFT META-REVIEW]
<M>

Task:
1) List paper-specific factual claims in M
↪ that should be verifiable against
↪ PAPER-FACTS.
2) List decision-critical statements that
↪ assert resolution/consensus or downplay
↪ major concerns.
```

```
For each item, output:
- claim_text
- claim_type (paper_fact / decision_critical)
- suggested_query_keywords
Return JSON list. Do not verify yet.
```

F.3.5 Return-to-Source Revision

```
SYSTEM:
Revise a meta-review using evidence spans. Do
↪ not add unsupported claims.

USER:
[DRAFT META-REVIEW]
<M>

[FAILED CLAIMS]
<CLAIMS_JSON>

[EVIDENCE PACK]
<G = {(id, span), ...}>

Task:
- For each failed claim, either (a) revise the
↪ claim to match the evidence span, or
  (b) rewrite conservatively if evidence is
↪ insufficient (e.g., "the paper does not
↪ clearly specify ...").
- Do not introduce new paper-specific facts
↪ beyond the evidence pack.

Formatting constraints (strict):
- Output MUST contain EXACTLY these four
↪ section headings, each on its own line,
  in this exact order:
  Summary:
  Strengths:
  Weaknesses:
  Decision/Rationale:
- Do NOT add numbering, letters, or extra
↪ prefixes to the headings.
- Do NOT add any extra sections.
- Under Strengths/Weaknesses, use plain '-'
↪ bullets only.
- Do NOT repeat the headings elsewhere; do not
↪ add a title line.
- Keep the output length within +/-10% of M.
```

Note. Style alignment uses the template in Appendix F.2 and is applied only after evidence anchoring passes.