FULL CONTEXT IS NOT ENOUGH: EVIDENCE-GROUNDED META-REVIEW GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As some major research venues begin to evaluate tightly bounded AI assistance while retaining strict limits on delegated reviewing, reliable evaluation of meta-review support is becoming increasingly consequential. Meta-reviewing must reconcile paper evidence, reviews, rebuttals, and unresolved disagreement into a defensible decision rationale; fluent automation can instead introduce unsupported paper claims or fictitious reviewer consensus. Existing benchmarks often expose only partial paper context and rely on reference-based text-similarity metrics, leaving these failures difficult to measure. We introduce MetaBench, a 4,782-instance benchmark pairing full papers with review threads and human meta-reviews, and FACT, a pipeline that separates issue-conditioned planning from a return-to-source evidence gate. On a same-drafter 2×2 comparison over 1,520 cases, Direct+Gate records 84.7% supported paper claims versus 73.8% for Direct, whereas Plan records 76.0%; Plan+Gate reaches 88.9% support and 80.6% major-issue coverage. On a 200-paper expert-gold slice, the same ordering holds, while claim extraction attains .79 precision, .82 recall, and .80 F1. In aggregate judgments from a 120-case blinded study, Plan+Gate receives more wins than losses on all four endpoints against the strongest gated baseline. In the matched factorial, verification is associated with the largest support change, while planning adds most clearly to issue coverage and reductions in conflict errors when paired with the gate. Within the Direct stack, replacing the abstract with PDF text and captions is associated with a smaller support change than explicit verification.

1 INTRODUCTION

As peer-review infrastructure scales, decision-critical artifacts such as meta-reviews must remain complete, consistent, and grounded in the submitted paper. Some venues are evaluating narrowly constrained, human-accountable assistance while their policies remain restrictive: ICLR 2027 allows disclosed limited use in reviews and meta-reviews, NeurIPS 2026 confines sanctioned assistance to an opt-in OpenReview experiment, and ACL permits only secondary assistance such as paraphrasing and proof checking (ICLR Conference, 2026; NeurIPS, 2026; Association for Computational Linguistics, 2024). These developments make reliable meta-review support timely without transferring judgment or accountability away from human reviewers and area chairs. 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.

We study **Meta-Review Auto-Generation (MRAG)**: given extracted submission text and captions, review threads, and explicitly declared article fields, generate a structured meta-review (Fig. 1). 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. We therefore treat evidence support and issue coverage as separate objectives and report decision prediction only in a dedicated diagnostic where the final decision is withheld.

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., 2025;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

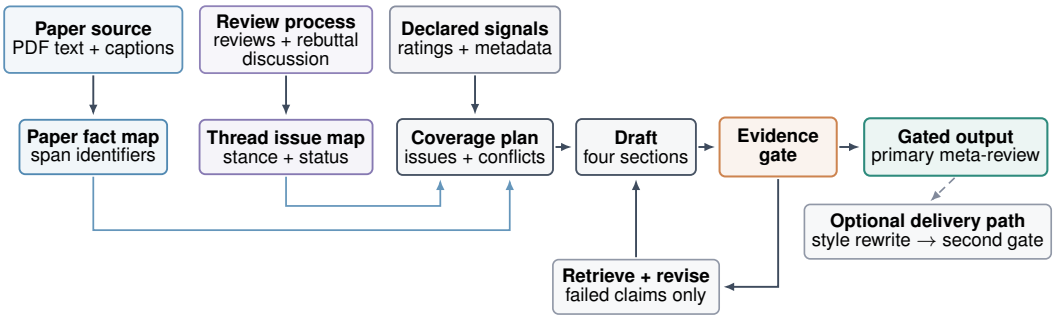


Figure 1: **FACT task and workflow.** The input contract separates paper evidence, the multi-party review process, and declared decision signals. FACT keeps a paper fact map and a thread issue map distinct, plans for coverage, and returns only failed claims to source. Primary experiments evaluate the gated pre-style output; style rewriting is an optional delivery branch whose output can be gated again.

Shen et al., 2022). To our knowledge, existing MRAG resources do not jointly provide full-paper evidence grounding and conflict-aware evaluation.

To address these gaps, we make three 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 dual-source 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.
- **A mechanism-focused evaluation.** A same-drafter Planning $\times$ Gate comparison with a shared output-length target, a 200-paper expert-gold audit, a 120-case blinded comparison, and a decision-blind diagnostic distinguish gains in factual support from gains in issue coverage. The observed cells associate the gate with most of the support improvement, while planning's largest incremental separation appears in issue coverage and reductions in conflict errors; in aggregate blinded judgments, Plan+Gate receives more wins than losses on all four endpoints against the strongest gated baseline.

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., 2025). 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). Adjacent

work on automated paper reviewing increasingly consumes full PDFs: OpenReviewer fine-tunes a specialized model for conference-style reviews, while DeepReview uses multi-stage analysis, retrieval, and evidence-based argumentation (Idahl & Ahmadi, 2025; Zhu et al., 2025). These systems generate a first-order paper review; our task instead synthesizes an existing multi-party review process into a conflict-aware meta-review.

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., 2025), 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). More directly, RARR post-edits unsupported content, Chain-of-Verification separates drafting from verification, and FActScore decomposes long-form text into atomic claims (Gao et al., 2023; Dhuliawala et al., 2024; Min et al., 2023). FACT applies these ideas to dual paper/thread grounding and explicit preservation of disagreement state.

3 METABENCH DATASET CONSTRUCTION

We construct MetaBench from OpenReview[1] using venue/track restriction, instance validation and normalization, quality-aware curation, and balanced benchmark selection; pseudocode is in Appendix B. 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 require a downloadable submission PDF, official reviews, a meta-review, and final-decision metadata; rebuttals, discussions, ratings, and confidences are retained when available. The 16,742-item structured pool spans 31 sources. We retain the main tracks of ICLR, NeurIPS, TMLR, and EMNLP; Fig. 2 summarizes the subsequent filtering and the final source mix.

3.2 PAPER STRUCTURING AND VALIDATION

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

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

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. For generation inputs, each instance is reconstructed at the meta-review timestamp: we retain the then-current submission PDF and thread, exclude camera-ready revisions and post-decision comments, and split by paper-family clusters. Style and in-context exemplars are drawn from the training split and exclude the target family. The final decision d_p is exposed only in the decision-conditioned rationale track; it is hidden in the separate decision-blind diagnostic. 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 $Q_p = 0.8S_{\text{st}}(p) + 0.2S_{\text{cov}}(p)$ only as a curation proxy, not as a generation metric. S_{st} checks structural completeness; S_{cov} measures coverage of diverse salient review sentences. We use Q_p for ranking and percentile binning; implementation defaults are in Appendix E. An equal-size alternative

[1] https://openreview.net/

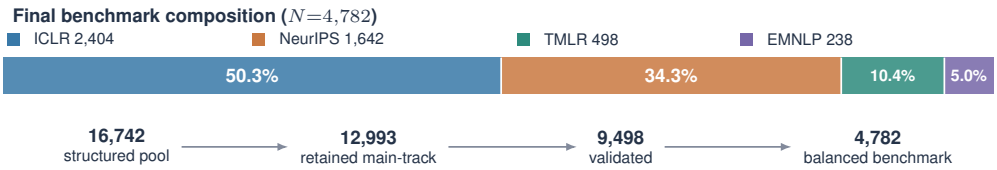


Figure 2: **MetaBench construction and composition.** The funnel shows aggregate item counts after each construction stage. The stacked bar reports the final decision-balanced benchmark by source; accept and reject each contain 2,391 items. Exact venue-year counts are in Appendix C.

curation run shares 4,451 of 4,782 instances with the default subset, yielding 93.08% one-set retention and Jaccard similarity $4{,}451/(4{,}782 + 4{,}782 - 4{,}451) = 0.8705$.

3.4 Balanced Benchmark Subset Construction

To control for acceptance-rate differences, for each venue unit u we freeze a per-class quota q_u and select the top-q_u instances by Q_p from each decision group. The resulting 4,782-item benchmark has 2,391 accept and 2,391 reject decisions. It supports controlled comparison rather than estimation of natural acceptance prevalence; Appendix C gives exact counts.

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

As illustrated in Fig. 1, FACT separates long MRAG inputs into paper facts and review-thread issues, drafts from a structured plan, and verifies claims before producing the primary gated output (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-based text-similarity metrics.

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 ratings, confidences, consensus cues, and, in the decision-conditioned track, the final decision. 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 used in the primary factorial. 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 when all automatically flagged paper claims are judged supported and decision-critical claims are consistent with their issue cards. Residual errors can therefore arise when either extraction or support judgment misses a claim, which we quantify separately below.

Return-to-Source Retrieval and Revision. When anchoring fails, FACT builds an evidence pack $\mathcal{G} = \mathcal{G}_P \cup \mathcal{G}_T$ from retrieved paper spans and the relevant issue-card records. Paper-span retrieval uses hybrid lexical/dense retrieval and re-ranking (Robertson & Zaragoza, 2009; Karpukhin et al., 2020; Nogueira & 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. We freeze the retrieval cutoffs, support threshold τ_{sup}=0.70, and two-pass revision budget across FACT and all ablations; Appendix E lists the values.

Style Alignment. We evaluate style alignment as optional post-processing rather than mixing it into the primary factorial. For venue-year unit u, a surface rewrite $\phi_{\text{style}}(\tilde{m}, \mathcal{E}_u)$ uses training-split exemplars $\mathcal{E}_u$ that exclude the target paper family. Separate diagnostics compare style rewriting without re-verification against a second gate pass. The 88.9% primary result scores the gated pre-style output, whereas the final-style variants are separate delivery configurations.

5 EXPERIMENTS

5.1 DESIGN AND METRICS

The primary automatic analysis uses 1,520 submissions stratified by venue source, decision label, curation-score percentile, and reviewer-disagreement signals. The primary generator is GPT-5-mini; metrics are averaged across three independent generations per paper. All factorial outputs use the same four-section schema, a 220–280 word target, the same full-paper representation, and the same decision-conditioned input contract. We separately use 200 papers for expert evidence-span annotation and 120 disagreement-heavy cases for blinded pairwise evaluation with three raters per case. The conventional four-backbone ROUGE-L (Lin, 2004) and BERTScore-F1 (Zhang et al., 2020) comparison is retained as a secondary reference-based text-similarity comparison in Table 2.

The experiments examine three linked questions: how factual support, issue coverage, and conflict errors differ under planning and verification in a matched generator; whether these patterns persist under controls for paper context, inference cost, output stage, and evaluator quality; and whether expert annotations and blinded human judgments provide complementary human checks. Together, these analyses combine mechanism-focused comparisons with human evaluation.

We decompose outputs into *paper-specific claims*, which require support from paper spans, and *decision-critical claims*, which must preserve reviewer stances and issue resolution states. Supported/unsupported percentages use only paper-specific claims; major-issue coverage uses a fixed reference issue-card set shared across systems; false-consensus and false-resolution use eligible decision-critical claims. For support, we compute a claim ratio within each paper and then take the unweighted mean over papers with at least one extracted paper-specific claim; zero-claim outputs remain in coverage and submission-level metrics. Because the final decision is visible in the primary factorial, its macro-F1 measures decision *consistency*, not blind prediction. These diagnostics complement rather than replace expert judgment; the expert audit below measures claim-extraction precision, recall, and F1. The gate and automatic support score share the GPT-4.1 judge family and

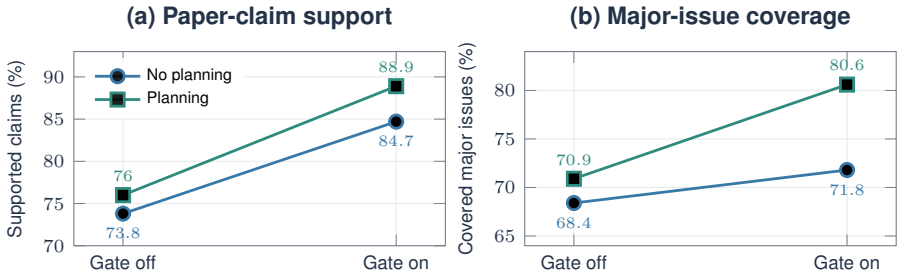


Figure 3: **Matched Planning × Gate results** (N=1,520)**.** Each panel is a factorial interaction plot: the horizontal change turns the gate on within a fixed planning condition, while the vertical separation compares planning conditions at the same gate setting. Numeric labels are the four observed cells. Axes are truncated; the plot shows point estimates without paired intervals.

claim taxonomy, so automatic support is an evaluator-aligned diagnostic rather than an independent validation. The human-adjudicated slice provides an external check for the systems it covers. All automatic values reported in the main text are aggregate point estimates. Factorial contrasts are arithmetic functions of the four displayed cells; we report no paired confidence intervals and do not interpret the contrasts causally.

Systems and controls. DIRECT uses the full paper and thread; PLAN adds issue-conditioned planning; DIRECT+GATE applies the full gate to the same Direct drafter; and PLAN+GATE is FACT. These four cells form the matched 2×2. Non-factorial controls are thread-centered CGI^2, CGI^2 drafts passed through the full gate, compute-matched self-refinement, and shuffled evidence. CGI^2+Full Gate remains a strong cross-drafter control, not a factorial cell.

5.2 AUTOMATIC SYSTEM COMPARISON

The matched cells isolate two descriptive patterns. With planning absent, adding the gate moves DIRECT from 73.8 to 84.7 support ($+10.9$ points) and from 68.4 to 71.8 issue coverage ($+3.4$). With the gate absent, planning moves support by $+2.2$ and coverage by $+2.5$; after gating, PLAN+GATE exceeds DIRECT+GATE by $+4.2$ support and $+8.8$ coverage points. The corresponding interactions computed from the displayed aggregate cells are $+2.0$ for support and $+6.3$ for coverage. Compute-matched self-refinement recovers much of the support gain over DIRECT ($+8.6$ points), while DIRECT+GATE adds a further 2.3 points and shuffled evidence falls below DIRECT. A compute-matched revision control matches much of the observed support difference, while correctly matched evidence is associated with a further separation. CGI^2+Full Gate reaches 86.3 support and 74.8 coverage, versus 88.9 and 80.6 for Plan+Gate. ROUGE-L rises from 20.68 for Direct to 22.25 for Plan+Gate on the matched GPT-5-mini cells, a smaller separation than the 15.1-point support difference; overlap alone does not isolate which mechanism produced the change.

Factorial decomposition. Averaging the two conditional contrasts, planning changes support by $+3.2$ points and coverage by $+5.7$, whereas gating changes them by $+11.9$ and $+6.6$. For the two error diagnostics, the corresponding average reductions are 2.3/2.1 points for planning and 8.9/7.2 for gating (false consensus/false resolution). The cell-derived interactions are $+2.0$ support, $+6.3$ coverage, -3.2 false consensus, and -3.5 false resolution. These are descriptive contrasts without sampling uncertainty; Fig. 3 shows the two primary outcomes without visually implying a sequential path between cells.

5.3 CONTEXT AND EFFICIENCY CONTROLS

Within the Direct stack, PDF text plus captions records 73.8 support and 68.4 coverage, versus 71.4 and 64.2 with the abstract alone. All five diagnostics in Fig. 4(a) shift favorably by 1.2–4.2 points. This is not a multimodal representation—figures and tables enter only through captions—and the values are descriptive point estimates. More paper text is useful here, but its support difference is smaller than the Direct-to-Direct+Gate contrast.

System	N	Sup.↑	F.Con.↓	F.Res.↓	Cov.↑	Cons. F1↑	R-L↑
Direct	1,520	73.8	17.4	14.6	68.4	75.6	20.68
Plan	1,520	76.0	16.7	14.3	70.9	77.5	21.05
Direct+Gate	1,520	84.7	10.1	9.2	71.8	79.4	21.92
Plan+Gate	1,520	**88.9**	**6.2**	**5.4**	**80.6**	**82.8**	22.25
CGI^2	1,520	78.6	14.8	12.5	72.1	78.4	21.89
Self-refine	1,520	82.4	12.1	10.4	69.8	79.3	21.74
CGI^2+Full Gate	1,520	86.3	9.4	8.6	74.8	80.2	22.12
Shuffled-evidence gate	1,520	70.8	19.6	16.1	66.5	74.8	20.35
Oracle evidence	200	93.2	5.5	4.8	86.5	84.0	22.44

Table 1: **Complete automatic system comparison.** The first four rows form the same-drafter Planning × Gate comparison. Consistency F1 is measured with the final decision supplied. CGI^2+Full Gate changes the drafter and is not a factorial cell; the oracle uses the expert-gold slice.

		`GPT-5-mini`		`Gemini-2.5-Flash`		`Claude-Sonnet-4.5`		`Qwen3.5`	
Context	**Method**	**R-L**	**B-F1**	**R-L**	**B-F1**	**R-L**	**B-F1**	**R-L**	**B-F1**
Thread	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
Full paper	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 2: **Reference-based text-similarity scores across four backbones.** Thread-only and full-paper groups use different inputs and should not be read as a controlled context ablation. On GPT-5-mini, full-paper FACT (22.25 R-L) is the same pipeline as Plan+Gate in Table 1; the One-shot row (18.09) uses a separate unconstrained-length prompt rather than the matched 220–280 word contract. R-L and B-F1 denote ROUGE-L and BERTScore-F1.

Cross-backbone reference-based text-similarity scores. Table 2 preserves the conventional comparison for all four generator families. FACT has the highest ROUGE-L and BERTScore-F1 point estimates among the listed full-paper methods for each backbone, but the absolute separations are modest. Because the thread-only and full-paper blocks use different inputs, this table is not evidence for the controlled context effect; source-grounded diagnostics remain primary. A complementary targeted comparison (Appendix Table 6) uses 1,520 cases per backbone. Relative to Direct+Gate, Plan+Gate is 3.9–4.2 points higher in support and 8.3–8.8 points higher in issue coverage across GPT-5-mini, Gemini-2.5-Flash, Claude-Sonnet-4.5, and Qwen3.5; these are aggregate point estimates.

The resource view in Fig. 4(b) makes the deployment trade-off visible rather than treating extra passes as free. Plan+Gate averages 4.2 model calls and 37.2k input tokens per paper, compared with 3.6/35.8k for Direct+Gate and 1.0/28.4k for Direct. CGI^2+Full Gate is a strong cross-drafter efficiency control. Compute-matched Self-refine recovers much of the Direct-to-gated support difference but remains below both gated systems. Latency and dollar estimates stay in the appendix because they depend on the serving and pricing snapshot.

5.4 Output Stage and Evaluator Diagnostics

The output-stage diagnostic distinguishes which text is being scored. Primary factorial values use the gated pre-style output. Applying the optional style rewrite without another gate lowers support from 88.9 to 87.4 and coverage from 80.6 to 80.1; a second gate partially restores them to 88.5 and 80.4 (Appendix Table 4, panel (a)). This observed aggregate loss indicates that a surface rewrite is not automatically content preserving, so delivered styled outputs should be re-verified.

Panel (b) reports separate stratified evaluator audits; sample units differ by target. Claim extraction has the highest audited F1 (.805), while the downstream unsupported and conflict-error labels are noisier. We therefore use support and coverage as the primary descriptive outcomes and treat false-consensus and false-resolution values as diagnostics rather than independently validated error rates.

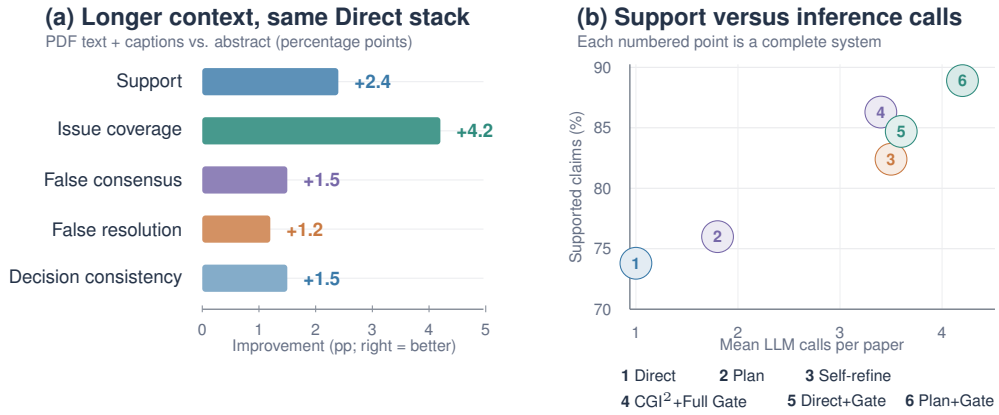


Figure 4: **Context and resource controls** (N=1,520)**.** (a) Same-stack changes from abstract+thread to extracted PDF text+captions+ thread. Error rates are plotted as reductions, so every rightward bar is an improvement. (b) Paper-claim support against mean model calls; numbered labels avoid overlapping system names. Both panels show descriptive aggregate point estimates without paired confidence intervals.

(a) Expert gold (N=200)

System	Sup.↑	Cov.↑
Direct	71.2	65.8
Plan	73.5	68.4
CGI^2+F.G.	83.8	72.3
Plan+Gate	**86.4**	**78.1**

(b) Plan+Gate vs. CGI^2+Full Gate

Endpoint	W	T	L
Paper faithfulness	55.8	27.5	16.7
Disagreement	57.2	26.4	16.4
Grounded rationale	56.9	27.8	15.3
Usefulness	51.4	31.9	16.7

(c) Decision blind

System	F1↑	ECE↓
Mean rating	76.8	.11
Majority vote	78.4	.09
Direct	77.2	.10
Plan+Gate	**80.1**	**.08**

Table 3: **Triangulated evaluation.** (a) Expert-adjudicated claim support and issue coverage. (b) Blinded preferences on 120 cases with three raters; W/T/L are percentages from Plan+Gate's perspective, with integer counts in Appendix Table 9. (c) Decision-blind prediction with the final decision withheld.

5.5 Human Evaluation

Expert-gold audit. Across the four annotated systems, expert adjudication preserves the automatic rank order (Table 3, panel (a)). For Plan+Gate, experts label 12.8 *major* unsupported claims per 100 atomic claims, 7.9% false consensus, and 6.8% false resolution. The first quantity is restricted to major claims and therefore is not the complement of the all-claim support percentage. The automatic claim extractor attains .79 precision, .82 recall, and .80 F1 against expert-enumerated claims. At .82 recall, extraction misses roughly 18% of expert-enumerated claims before support judgment, leaving a measurement blind spot. The pooled result is the unweighted mean of two equal-size strata rather than a population-weighted estimate. The gold slice contains 100 representative and 100 conflict-heavy papers; for Plan+Gate, support is 87.1/85.7 and coverage is 76.2/80.0 across the two strata. The development-calibrated support threshold is 0.69; we freeze the rounded value $\tau_{\text{sup}} = 0.70$ for test evaluation.

Blinded comparison. Three blinded raters compare outputs on 120 disagreement-heavy cases; ties are allowed. Table 3, panel (b), reports the comparison with the strongest gated baseline. In aggregate judgments, Plan+Gate receives more wins than losses on each endpoint, including 51.4/31.9/16.7 win/tie/loss for overall usefulness. Across all four endpoints, win rates range from 51.4–57.2%, compared with 15.3–16.7% losses. W/T/L are judgment-level percentages; because the three judgments within a case are not independent replicates, we use them to summarize preference direction rather than as a significance test. The separate comparison with the matched Direct+Gate cell, including integer vote counts, appears in Appendix A.

5.6 DECISION-BLIND DIAGNOSTIC

With the final decision hidden, PLAN+GATE reaches 80.1 macro-F1 (+1.7 over majority rating) and ECE .08 (Table 3, panel (c)); without paired uncertainty, both differences remain descriptive.

6 DISCUSSION

Planning and verification address different bottlenecks. The matched factorial does not support the claim that one generic "reasoning" step explains all gains. With planning off, the gate is associated with the largest observed support change; with the gate on, planning contributes a much larger coverage change than it does without verification. One plausible reading is that the issue plan increases the number and diversity of claims worth making, while return-to-source verification makes that richer draft safer. The positive coverage interaction is consistent with this complementarity, but it remains a descriptive aggregate interaction until paired paper-level uncertainty is available.

Longer context is useful but not self-enforcing. Within the Direct stack, the PDF-text-plus-captions condition is 2.4 support points and 4.2 coverage points above the abstract condition. Yet Direct with that same representation remains well below Direct+Gate on support and below Plan+Gate on both primary outcomes. Access and use are therefore distinct: a generator can receive relevant evidence without reliably retrieving the right span, preserving the reviewer stance attached to it, or checking its own final claim. Reference-based scores also change across the pipeline, but they do not identify whether planning, verification, or their interaction is responsible.

The deployment unit is the verified output, not the prompt. The stage ablation shows that even a nominally surface-only rewrite can undo part of the measured gain. A practical MRAG assistant should retain evidence and issue provenance through the final edit, re-run the gate after any style transformation, and expose unresolved flags to the area chair. This safety contract costs additional calls and tokens, and the evaluator audit shows that automatic gates themselves require human calibration. FACT should therefore be understood as an inspectable drafting workflow, not an autonomous decision maker.

7 CONCLUSION

We introduced MetaBench, a 4,782-instance full-paper MRAG benchmark, and FACT, a workflow that separates issue-conditioned planning from evidence verification. In the matched same-drafter cells, the largest support-associated change follows evidence gating, while planning adds most clearly to issue coverage and reductions in conflict errors once the gate is present; the ROUGE-L point estimate also increases but does not identify which mechanism produced the change. The abstract-to-full-text control is positive, but its support change is much smaller than the Direct-to-Direct+Gate contrast; expert-gold aggregates preserve the system ordering while revealing evaluator misses, and a second gate only partially restores the losses introduced by optional style rewriting. In aggregate blinded judgments, Plan+Gate also receives more wins than losses on all four endpoints against the strongest gated baseline. These findings argue for evaluating evidential support and deliberative coverage separately rather than treating longer context or reference-based similarity as sufficient.

8 LIMITATIONS

MetaBench covers OpenReview-hosted machine-learning venues, so transfer to other fields and review workflows remains untested. Automatic evaluation is imperfect—particularly for unsupported claims and conflict errors—and issue-card abstraction may flatten nuanced discussion; we therefore treat support and coverage as the primary descriptive outcomes and the conflict metrics as diagnostics. The shared judge family used for gating and automatic support scoring may favor evaluator-aligned revisions. Finally, PDF text plus captions can omit evidence encoded in figures and tables, and results may vary across model snapshots.

ETHICS STATEMENT

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. Submissions and review threads are confidential; deployment must use venue-approved, privacy-preserving or local infrastructure, follow the applicable venue policy, and avoid retention by unauthorized third parties. Review threads may also contain identifying information, requiring careful anonymization before any data release.

AI USE STATEMENT

We used AI assistants for language polishing and LaTeX table/figure formatting. The authors reviewed and edited all outputs and remain solely responsible for the methods, interpretation, and manuscript.

REFERENCES

Association for Computational Linguistics. ACL policy on publication ethics. https://www.aclweb.org/adminwiki/index.php?title=ACL_Policy_on_Publication_Ethics, 2024. Accessed 2026-09-26.

Chaitanya Bhatia, Tribikram Pradhan, and Sukomal Pal. MetaGen: An academic meta-review generation system. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1653–1656. Association for Computing Machinery, 2020. doi: 10.1145/3397271.3401190. URL https://doi.org/10.1145/3397271.3401190.

Wei Chen, Han Ding, Meng Yuan, Zhao Zhang, Deqing Wang, and Fuzhen Zhuang. Bridging social psychology and LLM reasoning: Conflict-aware meta-review generation via cognitive alignment, 2025. URL https://arxiv.org/abs/2503.13879.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-Verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578, 2024. doi: 10.18653/v1/2024.findings-acl.212. URL https://aclanthology.org/2024.findings-acl.212/.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, 2023. doi: 10.18653/v1/2023.acl-long.910. URL https://aclanthology.org/2023.acl-long.910/.

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. 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)*, pp. 7763–7803, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.395. URL https://aclanthology.org/2025.naacl-long.395/.

ICLR Conference. ICLR 2027 policy on LLM use for writing reviews or meta-reviews. https://iclr.cc/Conferences/2027/AIPolicyForReviewers, 2026. Accessed 2026-09-26.

Maximilian Idahl and Zahra Ahmadi. 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)*, pp. 550–562, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-demo.44. URL https://aclanthology.org/2025.naacl-demo.44/.

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

Asheesh Kumar, Tirthankar Ghosal, and Asif Ekbal. A deep neural architecture for decision-aware meta-review generation. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 222–225, 2021. doi: 10.1109/JCDL52503.2021.00064.

Miao Li, Eduard Hovy, and Jey Lau. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7089–7112. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.472. URL https://aclanthology.org/2023.findings-emnlp.472/.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023. doi: 10.52202/075280-2019. URL https://papers.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 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*, pp. 12076–12100, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

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

NeurIPS. NeurIPS 2025 reviewer guidelines. https://neurips.cc/Conferences/2025/ReviewerGuidelines, 2025b.

NeurIPS. NeurIPS 2026 AI-assisted reviewing experiment. https://neurips.cc/Conferences/2026/ai-reviewing-experiment, 2026. Accessed 2026-09-26.

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

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

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

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023. doi: 10.52202/075280-0377. URL https://papers.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Lu Sun, Stone Tao, Junjie Hu, and Steven P. Dow. MetaWriter: Exploring the potential and perils of AI writing support in scientific peer review. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1):1–32, April 2024. doi: 10.1145/3637371.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Qi Zeng, Mankeerat Sidhu, Ansel Blume, Hou Pong Chan, Lu Wang, and Heng Ji. Scientific opinion summarization: Paper meta-review generation dataset, methods, and evaluation. In *Artificial Intelligence for Research and Democracy (IJCAI 2024)*, volume 14917 of *Lecture Notes in Computer Science*, pp. 20–38. Springer Nature Singapore, 2025. doi: 10.1007/978-981-97-9536-9_2. URL https://doi.org/10.1007/978-981-97-9536-9_2.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. DeepReview: Improving LLM-based paper review with human-like deep thinking process. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29330–29355, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1420. URL https://aclanthology.org/2025.acl-long.1420/.

A Additional Aggregate Diagnostics

This section records context, cost, gold-slice, and pairwise diagnostics that complement the complete automatic comparison in Table 1.

(a) Plan+Gate output stage (N=1,520)

Scored output	Sup.	Cov.	F.Con.	F.Res.
Gated pre-style	**88.9**	**80.6**	**6.2**	**5.4**
Final style	87.4	80.1	6.8	5.9
Style + re-gate	88.5	80.4	6.4	5.5

(b) Independent pooled evaluator audits

Target	n	P	R	F1	κ
Claim extraction	2,400	.790	.820	.805	–
Unsupported	220	.550	.805	.653	.555
False consensus	110	.476	.714	.571	.494
False resolution	110	.467	.778	.583	.476

Table 4: **Stage sensitivity and evaluator validity.** Panel (a) separates the primary gated output from optional delivery variants. Panel (b) is computed from the confusion counts; sample units differ by row and are not used to establish system ordering.

(a) Same-stack context ablation

Direct input	Sup.	Cov.
Abstract + thread	71.4	64.2
PDF text/captions + thread	73.8	68.4

(b) Compute and cost (N=1,520)

System	Calls	Sec.	USD
Direct	1.0	42	.11
Plan	1.8	58	.15
Direct+Gate	3.6	96	.24
Plan+Gate	4.2	108	.27
CGI^2+Full Gate	3.4	88	.21
Self-refine	3.5	102	.25

Table 5: Controlled context and efficiency diagnostics. Panel (a) changes only the paper representation within the Direct stack; "PDF text/captions" is not full multimodal access. Panel (b) reports aggregate per-paper means for the primary GPT-5-mini runs. Latency and USD values are descriptive because the values are tied to a particular serving and pricing snapshot.

	Direct+Gate		Plan+Gate		Difference	
Backbone	**Sup.**	**Cov.**	**Sup.**	**Cov.**	**ΔSup.**	**ΔCov.**
GPT-5-mini	84.7	71.8	88.9	80.6	+4.2	+8.8
Gemini-2.5-Flash	84.1	71.2	88.2	79.8	+4.1	+8.6
Claude-Sonnet-4.5	85.2	72.4	89.4	81.1	+4.2	+8.7
Qwen3.5	83.9	70.9	87.8	79.2	+3.9	+8.3

Table 6: Targeted grounding metrics across generator backbones (N=1,520 per row). Each difference is Plan+Gate minus Direct+Gate.

(a) Expert-adjudicated systems

System	Support	Coverage
Direct	71.2	65.8
Plan	73.5	68.4
CGI^2+Full Gate	83.8	72.3
Plan+Gate	**86.4**	**78.1**

(b) Evaluator and Plan+Gate diagnostics

Quantity	Value
Claim-extraction recall	.82
Claim-extraction precision	.79
Claim-extraction F1	.80
Calibrated dev threshold τ_{sup}	.69
Frozen test threshold τ_{sup}	.70
Unsupported major / 100 atomic claims	12.8
False consensus (%)	7.9
False resolution (%)	6.8

Table 7: Complete 200-paper expert-gold summary. Panel (b)'s final three rows are for Plan+Gate; the unsupported-major count is not the complement of the all-claim support rate.

	Representative (N=100)		Conflict-heavy (N=100)	
System	**Sup.**	**Cov.**	**Sup.**	**Cov.**
Direct	72.0	66.9	70.4	64.7
Plan	74.8	69.7	72.2	67.1
CGI^2+Full Gate	84.6	73.8	83.0	70.8
Plan+Gate	**87.1**	**76.2**	**85.7**	**80.0**

Table 8: Expert-gold results by prespecified stratum. The pooled table averages the two equal-sized strata and is not a population-weighted estimate.

Comparator	Endpoint	W/T/L counts
Direct+Gate	Paper faithfulness	220/95/45
	Disagreement handling	225/90/45
	Grounded rationale	211/101/48
	Overall usefulness	195/107/58
CGI^2+Full Gate	Paper faithfulness	201/99/60
	Disagreement handling	206/95/59
	Grounded rationale	205/100/55
	Overall usefulness	185/115/60

Table 9: Pairwise counts from Plan+Gate's perspective (120 cases, three blinded raters). W/T/L are aggregate integer judgments. The Direct+Gate and CGI^2+Full Gate comparators are distinct. We report no inferential statistics for these aggregate counts.

Reproducibility assets. The appendix documents the prompts, aggregate tables, validation schema, and implementation defaults. Validation scripts check the reported arithmetic; paper-level intervals and case-level tests are not reported.

B Benchmark Construction Details

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 $\mathcal{D}_{bench}$

$\mathcal{U}_{sel} \leftarrow$ retained main-track venue-year units
$\mathcal{D}_{all} \leftarrow \emptyset$
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
for $\hat{\mathcal{I}}_p \in \mathcal{D}_{all}$ **do**
  $Q_p \leftarrow \Omega_q(\hat{\mathcal{I}}_p)$
end for
$\mathcal{D}_{bench} \leftarrow \phi_{\text{bal}}(\mathcal{D}_{all}, Q; \{q_u\})$
return $\mathcal{D}_{all}, \mathcal{D}_{bench}$

Procedural scope. CGI^2 is checklist-based and thread-centered; CAF explicitly structures reviewer conflicts; FACT additionally represents paper evidence, anchors claims, and performs bounded return-to-source revision (Fig. 1).

C VENUE-YEAR COVERAGE

The retained venue-year coverage is **ICLR** 2017–2025, **NeurIPS** 2021–2025, **TMLR** as a single venue without a year-specific partition, and **EMNLP** 2023.

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

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.

D GENERATED-OUTPUT HUMAN EVALUATION PROTOCOL

The generated-output study uses 120 disagreement-heavy cases and 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 two comparisons are Plan+Gate versus the matched Direct+Gate cell and Plan+Gate versus CGI^2+Full Gate. Outputs follow the same 220–280 word target. We report aggregate integer W/T/L counts over 360 judgments. Cases, rather than the three within-case ratings, are the intended inferential unit; no case-level test is reported. The main text reports all endpoints available for the strongest gated comparison, including the 185/115/60 usefulness counts. Table 9 reports counts without inferential statistics.

E IMPLEMENTATION DETAILS

Parameter	Default	Parameter	Default
α, β	$0.8, 0.2$	$k_{\text{lex}}, k_{\text{dense}}$	$20, 20$
L_{ref}	100	$k_{\text{cand}}, k_{\text{sup}}$	$30, 5$
K_{key}	5	τ_{sup} (support)	0.70
τ (coverage hit)	0.72	K_{iter}	2
Primary generator	GPT-5-mini	Samples per paper	3
Claim/support judge	GPT-4.1	Decision extractor	GPT-4.1-mini
Output target	220–280 words	Hard maximum	320 words

Table 11: Default settings for the primary factorial, scoring, and FACT verification. Dated API snapshot identifiers were not recorded, which limits exact replication across model updates.

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. The final-decision field is populated in the decision-conditioned experiments and omitted in the decision-blind diagnostic.

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>
<FINAL_DECISION_IF_CONDITIONED>
<OTHER_METADATA_IF_AVAILABLE>

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

Length constraints:
- Target total length: 220-280 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
  (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 only in the post-gate style diagnostics; its output is not the frozen primary factorial output.

```
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>
<FINAL_DECISION_IF_CONDITIONED>

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: 220-280 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. The primary factorial terminates at the gated output; post-gate style variants are reported separately as final-style and final-style-plus-reverification outputs.