FOUND BUT NOT WRITTEN: UNDERSTANDING AND REDUCING EVIDENCE ATTRITION IN DEEP RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep Research agents can retrieve relevant evidence yet omit details readers need. We study this *evidence-to-answer attrition* on 132 bilingual DeepResearch-Bench II (DRB2) tasks, where a selector flags candidate omissions in about 80% of reports. In controlled completion experiments, targeted additions fully express more omitted findings than untargeted additions, without improving mean report quality. We introduce *Evidence-Grounded Residual Completion* (EGRC), which organizes the answer around the request and revisits every original omission to retain content already expressed, revise a passage, or justify deferral. On 36 DRB2 test tasks, requiring a decision for every omission adds 3.9 quality points over freely selecting from the same omission list, with the same organized starting answer and nearly equal final length. The full method improves quality by 4.7 and 5.0 points over untargeted completion and two-pass self-checking, respectively, while recovering more omissions. On 23 scholarly tasks constructed following DeepScholar-Bench, EGRC improves recovery by 33 percentage points and reference-content coverage by 8.8 points over matched autonomous revision at nearly equal length. The content advantage persists under report-only assessment, and blinded readers favor EGRC over completion baselines. Explicit finding-level decisions improve recovery and report quality beyond answer organization and access to an omission list.

1 INTRODUCTION

A research report can fail its reader even when the acquired evidence contains the needed information. Consider a report on firm eligibility for a policy (Figure 1). The evidence states that firms must have fewer than 50 employees and that subsidiaries are excluded, but the answer says only that eligibility depends on firm size. The statement is supported; it still does not tell the reader which firms qualify. Here, the missing step is not another search. The relevant information is already available but has not reached the answer.

We call this loss *evidence-to-answer attrition* and ask: *how can an agent recover relevant information already in its evidence and make it useful in the answer?* Long-context studies show that access does not ensure use (Liu et al., 2024; Xia, 2026). Deep Research agents must also combine many findings into a coherent answer. Recent methods address this challenge through evidence organization, synthesis, and feedback (Li et al., 2026b; Liu et al., 2026; Wang et al., 2026). Yet organizing evidence under the right topic does not establish that the answer conveys a finding's decisive details.

We distinguish **selection**, choosing a finding, from **expression**, conveying its essential content. A report can contain only supported statements while leaving relevant findings absent or incompletely expressed. A selector comparing source-linked findings with reports flags candidate omissions in 106 of 132 DeepResearch-Bench II (DRB2) tasks. We test whether making these omissions explicit changes how the acquired evidence is used.

With the evidence, writing budget, and original report held fixed, we compare targeted and untargeted completion. Targeting raises the proportion of original omissions fully expressed within additions by 9.5 percentage points. Exchanging equally sized target sets redirects which findings are expressed. Yet some source-supported additions do not independently express their referenced findings in full. Targeted completion also does not yield higher mean report quality than untargeted

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.

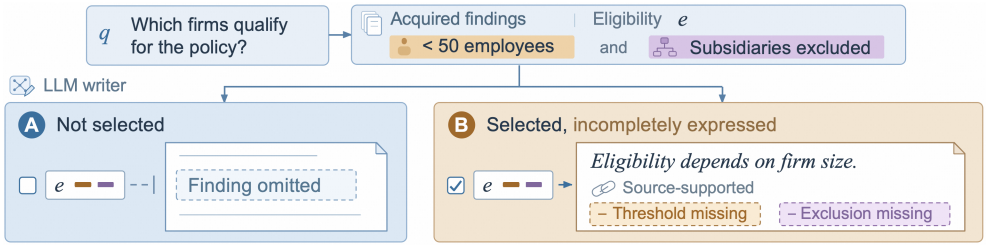

Figure 1: **Two forms of evidence-to-answer attrition.** In this illustrative policy example, (A) the finding is not selected; (B) it is selected but the employee threshold and subsidiary exclusion are omitted. A supported statement need not convey the conditions the reader needs.

completion. Together, these results motivate integrating recovered details into the answer while checking whether the original omissions are fully expressed in the final report.

We propose *Evidence-Grounded Residual Completion* (EGRC) with a two-level design: *the request guides answer organization, while the original omissions remain explicit requirements throughout revision.* After organizing the answer, EGRC revisits every original omission to retain its existing expression, revise a relevant passage, or justify deferral. The full finding context remains available to supply premises and qualifications, while source-reviewed local edits integrate recovered details into the answer.

Once an answer is organized, is supplying the original omission list enough? We compare two revisers given the same organized answer, original omissions, sources, and revision procedure. List-only may choose which findings to address; EGRC must account for every original omission. On 36 DRB2 test tasks, this requirement adds 3.9 quality points and 20.2 percentage points in confirmed recovery, with final lengths differing by only 0.7%.

The full method improves both quality and recovery over untargeted completion and two-pass self-checking. On 23 scholarly synthesis tasks constructed following DeepScholar-Bench, EGRC improves recovery by 33 percentage points and reference-content coverage by 8.8 points over matched autonomous revision at nearly equal length. The content advantage persists under report-only assessment. EGRC also exceeds Self-Refine and DeepWeaver in mean DRB2 quality and scholarly reference-content coverage. Blinded readers favor EGRC over the completion baselines and confirm greater recovery of sampled findings.

Our contributions are:

- **A diagnosis of evidence-to-answer attrition.** We trace acquired findings into reports and identify widespread candidate omissions. Controlled completion experiments distinguish selecting a finding from expressing it completely.
- **A two-level method with persistent finding requirements.** EGRC couples request-level answer organization with explicit decisions for every original omission and source-reviewed local edits.
- **Evidence for joint improvement and transfer.** Same-list comparisons show that explicit per-finding decisions improve recovery and report quality beyond access to the omission list. Evaluations on long reports and concise scholarly synthesis show gains in recovery and independently assessed quality, supported by blinded reader preferences.

2 Related Work

Evidence synthesis and content planning. Deep Research methods organize acquired evidence through dynamic outlines, graphs, and task criteria (Li et al., 2026b; Zhang et al., 2026b; Yan et al., 2026). Long-form story generation uses persistent plans and detailed outlines to guide writing (Yang et al., 2022; 2023), while intent-aware question answering makes paragraph and citation purposes explicit (Zhao et al., 2026). EviReport combines fact extraction and subsection writing with retrieval for missing information (Liu et al., 2026). The concurrent DeepWeaver refines draft-derived thought-block chains using residual evidence not yet covered by the current chain, then generates answers from the revised structure. It also evaluates detail preservation in the final answer (Wang et al., 2026). EGRC instead retains findings identified as omissions in the initial report and requires

an explicit decision for each during final revision. Our same-list comparison tests the added value of these per-finding decisions when both revisers already receive the same organized answer and omission list.

Feedback and grounded revision. Self-Refine, Reflexion, and CRITIC improve outputs through model, environment, or tool feedback (Madaan et al., 2023; Shinn et al., 2023; Gou et al., 2024). For long-form reports, CogGen uses recursive planning and revision to support global restructuring (Tian et al., 2026). CoVe generates verification questions from a draft and answers them independently before revising it (Dhuliawala et al., 2024). RARR retrieves evidence and edits unsupported content, while PAVE extracts question-conditioned premises and revises drafts with low support scores (Gao et al., 2023a; Huang et al., 2026). Our focus complements support-oriented editing: relevant information can remain absent even when the statements an answer makes are supported. EGRC therefore retains original omissions as explicit revision requirements alongside source-support checks.

Measuring evidence use. Available context can remain unused, and answerability does not ensure realized evidence benefit (Liu et al., 2024; Xia, 2026). Factuality and citation measures assess support for generated content (Min et al., 2023; Gao et al., 2023b). Deep Research benchmarks evaluate reference coverage, synthesis, and report quality (Java et al., 2026; Li et al., 2026a; Patel et al., 2025; Jung et al., 2026). DeepSynth-Eval evaluates post-retrieval synthesis using supplied reference-paper contexts and factual and structural checklists derived from reference surveys (Zhang et al., 2026a). Our omission measure instead compares acquired findings with the initial report, without using benchmark reference criteria. We evaluate recovery in the revised report alongside quality judgments made without the omission list. This tests whether omission recovery is accompanied by improvements in overall report quality.

3 Evidence-to-Answer Attrition

Holding the acquired evidence fixed, we examine how widely omissions occur, whether explicit targets change what is written, and whether source-supported additions fully express the findings they reference.

3.1 Measuring answer-relative omissions

For request q_i, let E_i be the acquired source-linked findings and y_i^0 the initial report. A finding records a substantive statement and its qualifications, rather than merely a topic. The answer-relative residual is

$$R_i = \{e \in E_i : e \text{ is judged relevant to } q_i \text{ and not fully expressed in } y_i^0\}. \tag{1}$$

The selector proposes these candidate omissions using the request, findings, and report, without reference answers or benchmark criteria. A separate expression assessment checks what the report conveys. We hold R_i fixed so that revisions address the same original omissions.

Let $M(t, e) \in \{0, 1\}$ indicate whether text t expresses the essential content of finding e, including material quantities, comparisons, qualifications, and scope. Source support does not by itself establish complete expression: stating that an eligibility threshold exists may be supported while still omitting its value. Several passages can jointly express a finding. For a report y and $R_i \neq \varnothing$, whole-report recovery is

$$\text{Recovery}_i(y) = \frac{1}{|R_i|} \sum_{e \in R_i} M(y, e). \tag{2}$$

Supplying R_i makes targets explicit without adding evidence; the complete finding context remains available. The reviser's own declarations do not determine M. We assess report quality separately through reader preferences and benchmark criteria, without showing assessors R_i.

Confirmed recovery counts each finding judged fully expressed once. Findings with unresolved judgments remain in the denominator without credit; supplementary bounds account for them. For full-method comparisons, we average tasks with identified omissions equally. Controlled completion assesses expression within individual additions, whereas full-method evaluation assesses the complete report. Because the selector and expression assessor make separate judgments, an initial report can also receive nonzero recovery.

3.2 Omissions are widespread across reports

A shared research process on all 132 bilingual DRB2 tasks yields 21,030 source-linked findings. The selector flags 1,343 candidate omissions, or 6.4% of findings, across 106 reports. This pattern persists when alternative writers use the same acquired findings.

In an independent 128-finding sample, both annotators confirm source support, relevance, and absence for 40 of 64 selected findings versus four of 64 unselected findings. These checks corroborate that the selector concentrates supported, relevant omissions (Appendix G).

3.3 Targeting changes what is written

To isolate content selection, we keep the original report unchanged and compare two ways of adding acquired information. Untargeted completion (Generic) chooses useful additions freely; targeted completion (TC) receives the original omissions R_i. Both use the full finding context and the same writer, budget, source checks, and insertion rules. Candidate-level recovery credits an original omission only if at least one inserted addition referencing it fully expresses its essential content. For this selection study, we average all 132 tasks, assigning paired zeros to those without omissions.

TC improves candidate-level recovery by 9.5 percentage points over Generic. The advantage persists across two further generation runs and human calibration of expression judgments (Appendices D and G). A separate test exchanges R_i for an equally sized, disjoint same-task finding set, keeping the evidence and list format fixed. Both sets are assessed in each condition, independently of the additions' declared finding references. Complete expression follows the requested set (Figure 2a): target identity, not just the presence of a list, changes which findings are expressed.

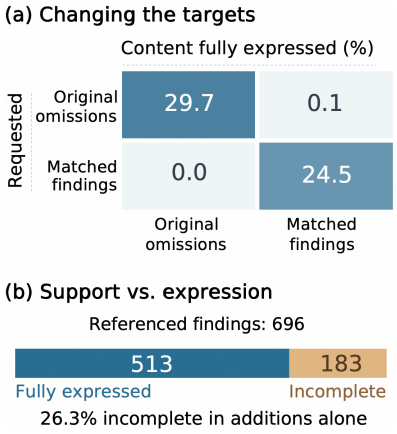

Figure 2: **Selection and expression within additions.** (a) Target exchange (132-task means). (b) Complete expression among findings referenced by source-verified, inserted additions. Neither panel measures whole-report recovery.

3.4 Selection and support do not ensure expression

Targeting tells the writer what to recover; it does not establish that the generated additions express those findings in full. Among 696 original omissions referenced by source-verified, inserted TC additions, 513 are fully expressed within at least one addition (Figure 2b). The other 183 fail this candidate-level criterion. Existing and added passages can jointly express a finding, so this is not a count of omissions remaining in the whole report (Appendix B). This diagnostic motivates checking expression separately from support and assessing the complete revised report in the full-method experiments.

TC also recovers more findings per added character, yet its benchmark quality score (0–100) is not higher than Generic's (38.1 versus 38.6). Together, these results motivate coupling answer-level organization with explicit finding-level decisions: the request guides how recovered content serves the answer, while each original omission requires an explicit revision decision.

4 Residual-Guided Answer Revision

EGRC first organizes missing content around the request, then revisits every original omission to determine what still needs to be expressed (Figure 3). The answer plan can change how findings are grouped while preserving the requirement to consider each original omission.

For a single task, a selector compares the acquired findings E with the initial report y^0 under request q to identify the original omission set R (Section 3.1). If $R = \varnothing$, EGRC returns y^0 without revision. Otherwise, it uses the acquired source passages S to produce a revised report y' under length limit B, without another search. Finding identities in R remain fixed, while their expression is re-

assessed against the updated report. Each finding receives an explicit decision, including deferral when source, relevance, or space constraints prevent inclusion.

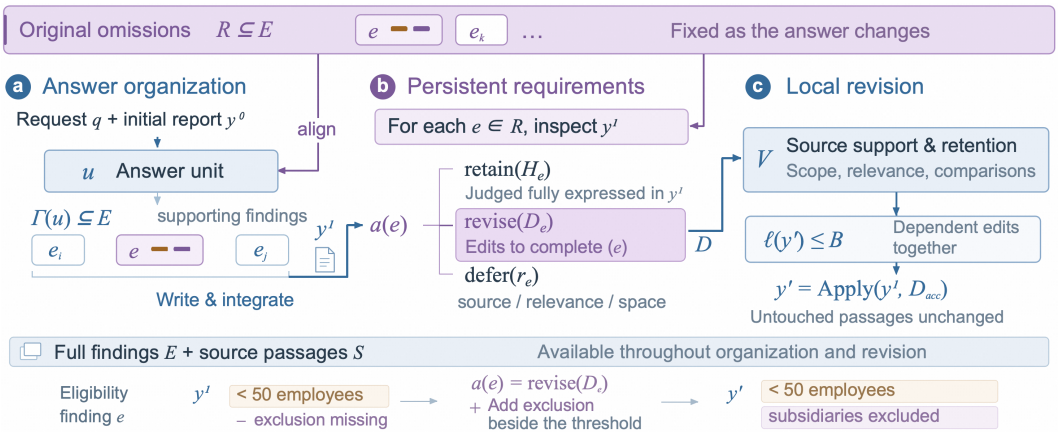

Figure 3: **Residual-guided answer revision.** (a) Answer units organize evidence into y^1. (b) Each original omission receives a retain, revise, or defer decision. (c) Source-reviewed local edits produce y' within budget. The full finding context remains available. The example adds the missing exclusion beside the employee threshold.

4.1 Organizing omissions around the requested answer

We first propose answer units for parts of q that y^0 has not adequately answered (Figure 3a). Each unit u specifies a fact, comparison, explanation, or condition needed to answer the request, with supporting findings $\Gamma(u) \subseteq E$. We align these units with R, linking omissions to the answers they help complete and adding a unit when a finding reveals a further unanswered part of the request. Several findings can support one unit, and one finding can qualify several units.

Selecting targets does not restrict their evidence context: $\Gamma(u)$ need not lie within R, since an already expressed rule may supply the premise for a missing exception. The complete finding context remains available throughout revision. Using associated passages in S, we integrate supported content into relevant parts of the report and use source feedback to correct relations and scope, producing the intermediate report y^1. Recovered content can thus complete an existing explanation, comparison, or condition rather than appear as a detached fact.

4.2 Retaining finding-level requirements

Plan membership does not establish complete expression: $e \in \Gamma(u)$ does not imply $M(y^1, e) = 1$. A draft can follow the plan while omitting a finding's decisive quantity, condition, or comparison. In the eligibility example, y^1 states the employee threshold but omits the subsidiary exclusion. EGRC therefore revisits every finding in R against y^1 (Figure 3b), allowing several passages to jointly express a finding.

For a proposed set of local edits D, let $a(e)$ record the decision for each original omission $e \in R$:

$$a(e) \in \{\text{retain}(H_e),\ \text{revise}(D_e),\ \text{defer}(r_e)\}. \tag{3}$$

For retention, H_e identifies passages that the reviser judges to express the finding. For revision, $\varnothing \neq D_e \subseteq D$ identifies the proposed edits intended to complete it. For deferral, r_e records the source, relevance, or space constraint preventing inclusion.

Each original omission must receive a decision, even if the plan absorbs it into a broader topic. This does not require one sentence per finding or inclusion of every selected finding. A concise table row can express a threshold and its exclusion together; a source conflict or mismatch with the request can justify deferral.

These decisions guide revision, not evaluation. Recovery is assessed independently on the final report using $M(y', e)$ for every $e \in R$, including deferred findings. Credit depends on what the resulting text expresses, not on the reviser's declared action.

4.3 EVIDENCE AND LENGTH CONSTRAINTS ON LOCAL REVISION

A qualification belongs beside the claim it qualifies. Edits replace complete local passages so that recovered details can modify the explanation, comparison, or condition they belong to (Figure 3c). Each edit $d \in D$ specifies a passage of y^1 and its replacement; the intervals do not overlap.

Starting from an intermediate report with $\ell(y^1) \leq B$, the accepted edits $D_{\text{acc}} \subseteq D$ produce

$$y' = \text{Apply}(y^1, D_{\text{acc}}), \qquad \ell(y') \leq B, \qquad V(d; y^1, D, S) = 1 \quad (d \in D_{\text{acc}}). \tag{4}$$

Here Apply performs the accepted replacements, and ℓ measures report length in the same units as B. The predicate V records whether an edit passes source review in the context of y^1, the proposed changes D, and sources S. Review checks support, relevance, scope, preservation of useful content, and the basis for comparisons.

Dependent edits form indivisible groups for review and admission. For example, moving a qualification requires accepting its removal and reinsertion together. Edits can reuse a premise or consolidate repetition to make room. We consider source-accepted non-expanding groups first and expanding groups second, preserving proposal order within each category. A group is applied only if the resulting complete report fits within B; otherwise, it is left unapplied rather than truncated.

Text outside accepted edits remains unchanged. Preservation of meaning within replaced passages depends on the proposal and source review. Appendix A gives the full algorithm and the conditions establishing locality and budget feasibility.

5 EXPERIMENTS

We evaluate whether EGRC improves finding recovery and report quality together, whether explicit per-finding decisions add value beyond an organized answer and the same omission list, and whether these gains extend to short scholarly synthesis.

5.1 TASKS, BASELINES, AND EVALUATION

DRB2 provides bilingual long-form research tasks (Li et al., 2026a). Section 3 analyzes omissions across all 132 tasks; here we evaluate the full method on tasks with identified omissions. We report results on a 44-task validation set and a separate 36-task held-out test set, with the method configuration fixed before test evaluation. A collection constructed following DeepScholar-Bench (Patel et al., 2025) tests related-work synthesis under a 600-word limit (Section 5.4). Complementary analyses examine SciConBench and the complete DRB1 agent (Section 5.6).

On DRB2, EGRC, Generic, and Two-pass share the original report, complete finding summaries, and per-task length ceiling. Generic adds grounded content without a supplied omission list; Two-pass first identifies its own omissions and then completes the report. Their additions are checked against linked source passages. EGRC additionally reads acquired passages and revises existing text. Matched final revisions start from EGRC's intermediate answer and share its source access, writer, revision procedure, and length ceiling. Autonomous revision chooses its own findings. List-only receives the original omissions but chooses which entries to address; EGRC maintains them as persistent finding-level requirements, with an explicit revision decision for each.

On the 36-task test set and scholarly collection, we also reproduce Self-Refine (Madaan et al., 2023) and DeepWeaver (Wang et al., 2026) within our fixed-evidence setting. Both use the same writer, acquired findings and sources, and per-task report-length limits as EGRC. Actual output lengths and inference costs vary across methods.

Qwen3.8-Max is the main writer for DRB2 and scholarly reports. EGRC obtains factual and scope feedback from DeepSeek-V3.2 and source review from GLM-5.2. Separate GLM-5.2 calls assess whole-report expression without method labels or the reviser's completion declarations. Gemini-2.5-Pro assesses DRB2 quality using the released prompts, with three rating rounds per report averaged within each task. Scholarly reference-content assessment uses GPT-4o, with an additional GPT-5.5 report-only check described below. Reference criteria are used for evaluation, not generation. Quality judges receive neither R nor the reviser's finding decisions. Further settings and analyses appear in Appendices C and E.

5.2 Recovery and quality in the same answer

Table 1 reports recovery and quality for each method's final reports. Unlike the selection study, which preserves the original text and only adds content, EGRC can also revise existing passages.

Table 1: Report quality and finding recovery on DRB2. Quality uses the benchmark's 0–100 scale; confirmed recovery is the task-average percentage of original omissions fully expressed. All reports use the same expression criterion.

(a) 44-task validation

Method	Quality	Recovery (%)
Initial report	40.0	1.6
Generic	42.0	35.1
Two-pass	41.7	37.6
EGRC	**46.8**	**52.7**

(b) 36-task test

Method	Quality	Recovery (%)
Initial report	36.7	1.7
Generic	37.9	21.7
Two-pass	37.6	29.6
Self-Refine	39.3	6.2
DeepWeaver	39.5	8.0
EGRC	**42.6**	**57.0**

EGRC improves mean report quality and omission recovery in both collections. With the configuration unchanged, quality gains on the 36-task test set are 4.7 points over Generic and 5.0 over Two-pass. Quality improves on 30 and 32 of the 36 tasks, respectively.

Mean information recall, analysis, and presentation scores all improve over the completion baselines. These benchmark criteria cover the report as a whole, including content already present or outside the original omissions. Recovery also remains higher on omissions identified by both Qwen and DeepSeek selectors and under an independent GPT-5.5 expression assessment. The latter gives a smaller advantage over Two-pass, preserving the direction while changing its magnitude (Appendix E.1). EGRC also exceeds Self-Refine and DeepWeaver in mean quality (+3.2 and +3.1 points) and recovery. Information recall and analysis favor EGRC, while presentation favors the external methods. DeepWeaver uses 2.3 times as many writing-stage calls and roughly twice as many input tokens as EGRC. EGRC produces longer reports: 16.2k tokens versus 9.6k for DeepWeaver and 8.4k for Self-Refine (Appendix F).

These full-method comparisons include additional reading and inference and allow different final lengths; on the 44 validation tasks, EGRC's reports are about 8% longer than Generic's. The next experiment isolates finding-level decisions with shared starting answers and revision procedures, at nearly equal final length.

5.3 Finding-level decisions improve organized answers

On the 44 validation tasks, EGRC's organized intermediate reports y^1 score above Generic and Two-pass, but whole-report assessment confirms full expression of only 111 of 595 original omissions (30 judgments remain unresolved). We test whether finding-level decisions improve the answers.

Three final revisers share y^1, the full finding context, sources, model, report limit, and review procedure. Autonomous revision chooses its own findings from the full context. List-only receives the original omissions but chooses which entries to address. EGRC additionally requires a retain, revise, or defer decision for every entry.

On the 44 tasks, EGRC adds 3.3 quality points over List-only at a 0.3% length difference. It also improves quality and nearly doubles confirmed recovery over autonomous revision (Figure 4).

The same-list comparison extends to all 36 test tasks with unchanged revision settings. EGRC improves quality by 3.9 points (95% interval [2.7, 5.2]), with gains on 32 of 36 tasks. Confirmed recovery increases by 20.2 percentage points, while final lengths differ by only 0.7%. An independent GPT-5.5 assessment also finds a 17.3-percentage-point gain in confirmed recovery. Thus, per-finding decisions improve recovery and independently assessed quality beyond access to the omission list.

Repeating the comparison with GLM-5.2 for answer organization and revision on a balanced 20-task test subset gives EGRC 3.0 more quality points and 13.5 percentage points higher mean recovery than List-only, with only a 0.3% final-length difference (Appendix E).

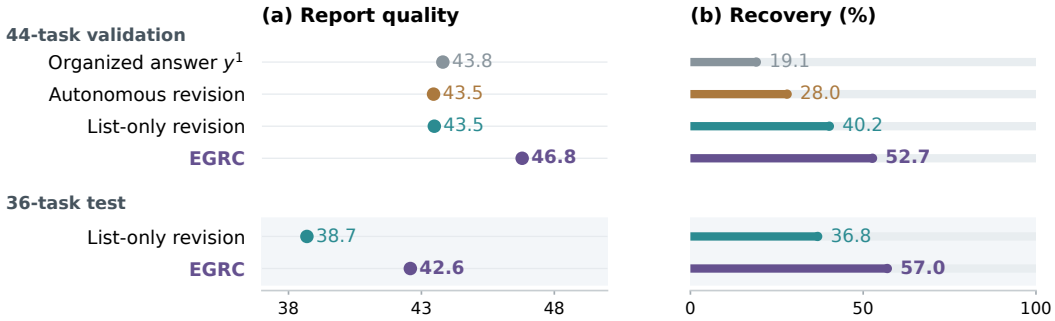

Figure 4: **Finding requirements add value after answer organization.** Within each collection, final revisers share the intermediate answer and revision procedure; List-only also receives the same omissions. Values are task means. Quality uses an enlarged score axis; confirmed recovery uses the 0–100 scale.

A finding-transition analysis on the 44 validation tasks compares EGRC with autonomous revision. Of the 454 original findings judged incomplete in y^1, EGRC fully expresses 170, compared with 48 under autonomous revision. Both revisers retain most original targets already expressed in y^1 (Appendix E.3).

5.4 Transfer to Short Scholarly Synthesis

We apply the same design to a scholarly synthesis collection constructed following DeepScholar-Bench (Patel et al., 2025). Given a paper abstract, the task is to write its related-work section using earlier arXiv sources, with a 600-word body limit. All 23 tasks with identified omissions in the 40-task collection are included. Generic revises the original report using the full finding context and acquired sources. Autonomous revision instead starts from EGRC's intermediate answer and shares its final revision opportunity.

Table 2: Scholarly synthesis on all 23 tasks with identified omissions. Recovery and reference coverage use a 0–100 scale. The released protocol uses GPT-4o; the report-only check uses GPT-5.5. Ranges bound unresolved judgments; words count the body. EGRC and autonomous revision share y^1.

Method	Recovery (%)	Reference content		Words
		Released	Report-only	
Initial report	0.9	42.4	23.5	372
Generic	0.9	44.1	22.6	416
Autonomous revision	52.9	47.0	26.3	516
Self-Refine	5.1	51.9	31.7	492
DeepWeaver	7.1	51.1	34.5–36.4	530
EGRC	**86.5**	**55.8**	**38.5–39.1**	521

Persistent finding requirements also improve concise synthesis. EGRC's task-average recovery advantage over autonomous revision is at least 33.0 percentage points. Both start from the same intermediate answer and produce nearly equal lengths (Table 2). An independent GPT-5.5 assessment also retains the recovery advantage (Appendix E.1). Under the released protocol, EGRC leads autonomous revision by 8.8 reference-content points. These nuggets come from withheld reference text, not the omission list.

The released assessor also sees the task abstract and credits some nuggets even in empty reports. We therefore reassess the unchanged reports with GPT-5.5, withholding the abstract and requiring sup-

porting quotations from the report. Empty reports receive zero credit. EGRC retains a report-only coverage advantage of at least 12.2 points over matched autonomous revision, supporting content gains in the report itself. The protocols differ in evaluator, abstract visibility, and quotation requirements, so their absolute scores are not interchangeable.

5.5 Reader preference and human-verified recovery

We ask readers to assess both complete reports and recovered findings. Two independent readers compare EGRC with Two-pass on 24 report pairs sampled from the DRB2 validation set, and with Generic on all 23 scholarly pairs. Method names are hidden, and left/right assignment is reversed between readers. Readers assess the complete reports before seeing the sampled findings.

Table 3: Reader preferences and human-confirmed recovery. Preference counts pool both readers over the same report pairs. Recovery gains are task averages on 43 sampled DRB2 findings and 23 scholarly findings, in percentage points.

Setting	Comparator	Pairs	EGRC	Baseline	Tie	Recovery gain
DRB2	Two-pass	24	**41**	6	1	+34.4
Scholarly	Generic	23	**34**	10	2	+93.5

Both readers favor EGRC overall in each setting (Table 3). Completeness and usefulness favor EGRC as well, while scholarly nonredundancy more often favors Generic (Appendix G.1). The readers also confirm greater recovery of sampled findings, with high inter-annotator agreement on complete expression (Cohen's $\kappa = 0.94$). These judgments support both finding recovery and reader preference over the completion baselines.

5.6 Scientific conclusions and a complete research agent

Two complementary analyses examine the role of finding recovery beyond the full-method comparisons. On 200 questions with identified omissions from SciConBench (Jung et al., 2026), residual-targeted completion uses the same acquired evidence as generic completion but raises source-grounded realization by 46.8 points. This extends the content-selection evidence to scientific conclusion writing.

Our complete DeepResearch-Bench (DRB1) agent (Du et al., 2026) scored 55.99 on the official GPT-5.5 evaluation and ranked first in the September 2 leaderboard snapshot (DeepResearch Bench Team, 2026). Within that system, expression of selected findings rises from 14.4% in the intermediate reports to 82.2% after synthesis. The system incorporates finding-guided synthesis alongside further research and rewriting. It demonstrates the relevance of finding recovery within a high-performing agent; the matched revisions above isolate the contribution of EGRC's persistent finding requirements. Appendix I gives both analyses.

6 Conclusion

Finding evidence and using it are separate achievements. Our analysis distinguishes failure to select a finding from incomplete expression. EGRC couples answer organization with persistent finding requirements. The same-list comparisons on validation and test tasks show that per-finding decisions improve an organized answer beyond simply supplying its omissions. Across long reports and concise scholarly synthesis, this design improves recovery together with independently assessed content quality. Blinded readers also favor EGRC over the completion baselines.

Our measurements depend on finding extraction and on relevance and expression judgments. Future work can extend evaluation to broader model families and repeated generations, and reduce the additional inference cost of revision. An important next step is deciding when to revise an answer from existing evidence and when to seek new evidence.

AI USE STATEMENT

We used generative AI tools to help refine the research framework and experimental design, develop and check formal arguments, implement and debug experimental code, analyze and interpret results, search and summarize related work, and draft and edit the manuscript. All AI-assisted material included in this submission was reviewed, verified, and where necessary revised by the authors, who take full responsibility for the correctness and originality of this submission.

REPRODUCIBILITY STATEMENT

We provide an anonymous supplementary package with core revision code, method and expression-assessment prompts, model settings, and numerical records for the full-method comparisons. Offline scripts reproduce task-average quality and recovery from the supplied scores and judgments. Executable examples and tests support inspection of the revision rules; the appendices describe the data, comparisons, and evaluation protocols.

ETHICS STATEMENT

Our study uses public research tasks and source materials. Annotators participated voluntarily with informed consent, with method identities hidden during evaluation. Generated reports may contain errors or biases and require verification before consequential use.

REFERENCES

DeepResearch Bench Team. DeepResearch Bench GPT-5.5 leaderboard. Hugging Face Space, 2026. URL https://huggingface.co/spaces/muset-ai/DeepResearch-Bench-Leaderboard. Snapshot accessed 2026-09-02.

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/.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Licheng Zhang, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench: A comprehensive benchmark for deep research agents. In *The Fourteenth International Conference on Learning Representations*, pp. 42414–42448, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/465f22be10e07b301c6ed58f0472f704-Abstract-Conference.html.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. 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, 2023a. doi: 10.18653/v1/2023.acl-long.910. URL https://aclanthology.org/2023.acl-long.910/.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023b. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/fef126561bbf9d4467dbb8d27334b8fe-Abstract-Conference.html.

Tianyi Huang, Caden Yang, Emily Yin, Eric Wang, and Michael Zhang. PAVE: Premise-grounded answer validation and editing for retrieval-augmented LLMs. *arXiv preprint arXiv:2603.20673*, 2026. URL https://arxiv.org/html/2603.20673v2.

Abhinav Java, Ashmit Khandelwal, Sukruta Midigeshi, Aaron Halfaker, Amit Jayant Deshpande, Navin Goyal, Ankur Gupta, Nagarajan Natarajan, and Amit Sharma. Characterizing deep research: A benchmark and formal definition. In *The Fourteenth International Conference on Learning Representations*, pp. 10315–10349, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/114e1dc345fe31b8b9b0c6f7b55a0644-Abstract-Conference.html.

Hayoung Jung, Pedro Viana Diniz, José Reinaldo Corrêa Roveda, Abner Fernandes da Silva, Haeun Jung, Enoch Tsai, Aleksandra Korolova, and Manoel Horta Ribeiro. Can AI agents synthesize scientific conclusions? *arXiv preprint arXiv:2606.11337*, 2026. URL https://arxiv.org/abs/2606.11337.

Ruizhe Li, Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench II: Diagnosing deep research agents via rubrics from expert reports. *arXiv preprint arXiv:2601.08536*, 2026a. URL https://arxiv.org/abs/2601.08536.

Zijian Li, Xin Guan, Bo Zhang, Shen Huang, Houquan Zhou, Shaopeng Lai, Ming Yan, Yong Jiang, Pengjun Xie, Fei Huang, Jun Zhang, and Jingren Zhou. WebWeaver: Structuring web-scale evidence with dynamic outlines for open-ended deep research. In *The Fourteenth International Conference on Learning Representations*, pp. 4173–4217, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/07fa611a88be1832f4c6a96f044ebede-Abstract-Conference.html.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Zihan Liu, Jianhui Li, Zexin Wang, Fei Sun, Jingjing Li, Zheyuan Li, Ke Xiang, Hang Cui, Houhua Gong, Changhua Pei, and Gaogang Xie. EviReport: From reasoned outlines to evidence tracked long-form reports. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 28024–28048, 2026. doi: 10.18653/v1/2026.findings-acl.1397. URL https://aclanthology.org/2026.findings-acl.1397/.

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://proceedings.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, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Liana Patel, Negar Arabzadeh, Harshit Gupta, Ankita Sundar, Ion Stoica, Matei Zaharia, and Carlos Guestrin. DeepScholar-Bench: A live benchmark and automated evaluation for generative research synthesis. *arXiv preprint arXiv:2508.20033*, 2025. URL https://arxiv.org/abs/2508.20033.

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://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Kuo Tian, Pengfei Sun, Zhen Wu, Junran Ding, and Xinyu Dai. CogGen: A cognitively inspired recursive framework for deep research report generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5961–5988, 2026. doi: 10.18653/v1/2026.findings-acl.296. URL https://aclanthology.org/2026.findings-acl.296/.

Xujia Wang, Yizhe Zhang, Bin Xu, Lei Hou, and Juanzi Li. DeepWeaver: Bridging the evidence synthesis gap in open-ended question answering. *arXiv preprint arXiv:2608.18988*, 2026. URL https://arxiv.org/abs/2608.18988.

Haizhou Xia. Diagnosing evidence utilization in long-context and retrieval-augmented language models under matched evidence conditions. *arXiv preprint arXiv:2606.06758*, 2026. URL https://arxiv.org/abs/2606.06758.

Lingyong Yan, Can Xu, Yukun Zhao, Wenxuan Li, Qingyang Chen, Jiulong Wu, Wenli Song, Xiangnan Li, Weixian Shi, Yiqun Chen, Xuchen Ma, Yuchen Li, Jiashu Zhao, Shuaiqiang Wang, Jianmin Wu, and Dawei Yin. DuMate-DeepResearch: An auditable multi-agent system with recursive search and rubric-grounded reasoning. *arXiv preprint arXiv:2606.07299*, 2026. URL https://arxiv.org/abs/2606.07299.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4393–4479, 2022. doi: 10.18653/v1/2022.emnlp-main.296. URL https://aclanthology.org/2022.emnlp-main.296/.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. DOC: Improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3378–3465, 2023. doi: 10.18653/v1/2023.acl-long.190. URL https://aclanthology.org/2023.acl-long.190/.

Hongzhi Zhang, Yuanze Hu, Tinghai Zhang, Jia Fu, Tao Wang, Junwei Jing, Zhaoxin Fan, Wei Bi, Ruiming Tang, Han Li, Guorui Zhou, and Kun Gai. DeepSynth-Eval: Objectively evaluating information consolidation in deep survey writing. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 33805–33822, 2026a. doi: 10.18653/v1/2026.findings-acl.1688. URL https://aclanthology.org/2026.findings-acl.1688/.

Zhen Zhang, Liangcai Su, Zhuo Chen, Xiang Lin, Haotian Xu, Simon Shaolei Du, Kaiyu Yang, Bo An, Lidong Bing, and Xinyu Wang. Argus: Evidence assembly for scalable deep research agents. *arXiv preprint arXiv:2605.16217*, 2026b. URL https://arxiv.org/abs/2605.16217.

Xinran Zhao, Aakanksha Naik, Jay DeYoung, Joseph Chee Chang, Jena D. Hwang, Tongshuang Wu, and Varsha Kishore. Improving attributed long-form question answering with intent awareness. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=fRCm5c8x0j.

A REVISION DECISIONS AND ADMISSIBLE EDITS

EGRC combines request-level answer organization with explicit decisions for the original omission set R. The plan groups acquired findings into answer units, while final revision revisits every original omission. Algorithm 1 gives the procedure. The same set R is used in alignment and final revision, so a broader answer plan does not replace the specific findings that motivated revision.

Algorithm 1 Evidence-Grounded Residual Completion

Input: request q, report y^0, findings E, acquired sources S, original omission set R, length limit B.
If $R = \varnothing$, **return** y^0.

1. Propose answer units for unanswered parts of q from (q, y^0, E), with supporting findings $\{\Gamma(u)\}$.
2. Align the units with R, keeping already expressed findings when they supply necessary premises.
3. Read associated passages in S; integrate supported answers and use source feedback for bounded local correction, obtaining y^1 with $\ell(y^1) \le B$ before final revision.
4. Jointly propose local edits D and a decision $a(e)$ for every $e \in R$, using (q, y^1, E, R, S). Validate decision coverage, finding-to-edit references, and valid, non-overlapping replacement intervals in y^1 before evaluating the proposed report length.
5. If $\ell(\mathrm{Apply}(y^1, D)) > B$, revise (D, a) once using length feedback against the same report y^1. Recheck the same structural conditions before source review.
6. Review source support, scope, retention, and comparison validity. Keep only groups whose dependent edits all pass. Order these groups $D_1, \ldots, D_k$, non-expanding groups first and proposal order within each category.
7. Starting from $A_0 = \varnothing$, admit each group that fits:

$$A_j = \begin{cases} A_{j-1} \cup D_j, & \ell(\mathrm{Apply}(y^1, A_{j-1} \cup D_j)) \le B, \\ A_{j-1}, & \text{otherwise.} \end{cases}$$

Output: $y' = \mathrm{Apply}(y^1, A_k)$, with $D_{\mathrm{acc}} = A_k$.

Local revisions. Each edit replaces a specified interval of y^1 with a complete passage. The intervals in one proposal do not overlap. A move can require removing text from one location and adding it elsewhere; these changes share a group and must be admitted together. An edit is not accepted merely because it claims to recover a target. Review checks whether its material claims are supported, its scope fits the request, it preserves useful existing information or justifies a removal, and any comparison is valid.

Length constraints. Equation 4 imposes the source and length conditions on accepted edits. Its predicate V records the source reviewer's assessment of each proposed edit. The limit applies to the complete revised report. A group is non-expanding if applying it alone to y^1 does not increase report length. Non-expanding groups precede expanding groups; within each category, groups follow the position of their first edit in the proposal. A group that would exceed the limit is left unapplied rather than truncated. This fixed ordering separates content choices in the proposal from admission under the common length limit.

Locality and budget feasibility. For $R \neq \varnothing$, suppose $\ell(y^1) \le B$, as required before final revision in Algorithm 1, and the proposed replacement intervals do not overlap. The initial choice $A_0 = \varnothing$ returns y^1, so it is feasible. Budget feedback changes the proposal (D, a), keeping both y^1 and B fixed. At iteration j, an accepted group gives a report within B by the admission test; otherwise $A_j = A_{j-1}$ and the previous feasible report is retained. Induction therefore gives $\ell(y') \le B$, including when no edit group is accepted. The test measures the complete report and does not require an additive length function. All replacements refer to the same y^1, so text outside the accepted intervals remains unchanged. These properties establish locality and budget feasibility. Preservation of meaning within a replaced region additionally depends on the proposal and source review.

Finding decisions and measured expression. The proposal contains exactly one decision for every original target. A revise decision references a nonempty subset of proposed edits; a retain

decision identifies passages in y^1 that the reviser judges to express the finding; and a defer decision records its reason. These fields make missing decisions and invalid references detectable. Several findings may reference the same edit, and several edits may jointly express one finding.

Recovery credits each original finding at most once, based on the final report rather than the declared action. An edit can be rejected, a claimed existing expression can be incomplete, or a source conflict can justify deferral. Whole-report assessment checks all original targets independently of these declarations. Deferral changes the proposed revision, not the evaluation set.

Proposal validity. Each proposal undergoes structural checks before length evaluation or source review: exactly one decision per original finding, valid finding-to-edit references, and valid, non-overlapping replacement intervals in the unchanged y^1. Any proposal returned after length feedback undergoes the same checks. If bounded format retries do not yield a valid proposal, revision stops with an error. Dependent edits share a group and are reviewed and admitted together.

Final admission checks $\ell(y^1) \le B$ before applying any edits, independently of model-based source review. All intermediate reports in our DRB2 and scholarly comparisons satisfy this condition. An oversized input raises an error rather than returning a clipped or over-budget report.

B Controlled completion and loss decomposition

The controlled operator isolates content selection from report rewriting. Let C be the ordered candidate additions retained after budget screening: the longest prefix of whole additions within the addition budget b. These candidates are submitted to source review. For each $c \in C$, $F(c)$ denotes its declared finding references and $P(c)$ the linked source passages. Targeted completion requires $\emptyset \neq F(c) \subseteq R$; Generic can choose from all E. Both retain the complete finding-summary context. A common support check receives the candidate and its passages, not the strategy or target labels. Whole accepted additions are inserted without changing the original text.

Expression across passages. In the illustrative policy example, let a selected finding $e \in R$ state both the fewer-than-50-employees requirement and the exclusion of subsidiaries. Suppose y^0 already states the employee threshold but omits the exclusion. A supported addition c states that subsidiaries are excluded. It does not repeat the threshold, giving $M(c, e) = 0$, whereas the two passages together can give $M(y', e) = 1$. The candidate-level measure does not credit e through c alone; the whole-report measure credits its complete expression using both passages. We use the former to diagnose the content of additions and the latter to assess the full revision method.

Text preservation. The resulting operator preserves the initial text in its original order: each insertion adds text without changing an existing span, so induction gives this property for every accepted list. If the source judgments are sound, every inserted material claim is supported by acquired passages. Given identical ordered candidates, source judgments, destinations, and remaining space, insertion is independent of the targeting strategy.

An exact decomposition. Let $U \subseteq C$ denote source-accepted candidates and $A \subseteq U$ the candidates inserted. For a fixed task with $R \neq \varnothing$ and a given strategy define

$$\mathcal{T} = R \cap \bigcup_{c \in C} F(c), \qquad \mathcal{G} = R \cap \bigcup_{c \in U} F(c),$$
$$\mathcal{K} = R \cap \bigcup_{c \in A} F(c), \qquad \mathcal{H} = \{e \in \mathcal{K} : \exists c \in A,\ e \in F(c),\ M(c, e) = 1\}. \tag{5}$$

These sets track declared finding references, counting each finding once even if several candidates reference it. They do not enumerate undeclared semantic matches. The sets are nested, giving

$$\frac{|\mathcal{H}|}{|R|} = \frac{|\mathcal{T}|}{|R|} \frac{|\mathcal{G}|}{|\mathcal{T}|} \frac{|\mathcal{K}|}{|\mathcal{G}|} \frac{|\mathcal{H}|}{|\mathcal{K}|}. \tag{6}$$

The factors describe finding-level selection within the budget-eligible prefix, retention after source review, insertion, and complete expression. They are ratios over distinct findings, not candidate-level pass rates. A ratio with zero denominator is defined as zero; nesting then forces $\mathcal{H}$ to be

empty. Otherwise the factors cancel. The identity requires no independence assumption and does not make the stages orthogonal causes. We compute the product for each task before averaging tasks. Multiplying the four averaged factors would generally give a different quantity. Tasks with $R = \varnothing$ contribute zero to the controlled-study average. Whole-report recovery under revision is evaluated directly and is not inferred from this candidate-level decomposition.

C Experimental settings

C.1 Tasks and acquired evidence

Shared research material. The DRB2 analysis uses all 132 tasks, with 66 tasks per language and 22 themes. Query-based discovery and source reading provide 2,143 source cards. Extraction produces 21,030 findings, each linked to source passages and a requested deliverable and report section. The extractor asks for atomic findings, but a finding can contain a relation and its qualifications. Exact duplicates with identical text, sources, and destination are removed; semantic paraphrases are not globally merged. The measured omission set therefore depends on the extraction and segmentation rule.

The selector compares each initial report with this collection and flags 1,343 candidate omissions: 882 across 56 of the 66 English tasks and 461 across 50 of the 66 Chinese tasks. No reference answer, benchmark criterion, or quality judgment enters this selection task. The same original findings, reports, and omission identities are used when the full revision method is evaluated on their task subsets.

Applying the selector to alternative initial reports written from the same acquired findings flags candidate omissions in 107 GLM-written and 117 DeepSeek-written reports. The selected sets differ across writers. Within-section omission rates rise from 3.5% to 9.7% across evidence-position quartiles and are higher in larger groups. These are observational associations, not an intervention establishing that position or group size causes omission.

DRB2 development, validation, and test tasks. All 132 tasks enter the attrition analysis. We use 24 tasks for initial method design and a disjoint set of 44 tasks for validation. Validation results informed the final design; the reported comparisons use the same final method and settings across all 44 tasks. The validation set comprises all tasks with identified omissions (595 findings) in a language-balanced pool of 60 tasks (30 English and 30 Chinese); the other 16 have empty residual sets.

The test set is disjoint from both the 24 design tasks and the 60-task validation pool and was held out from full-method development. Of the 48 remaining tasks, 38 have identified omissions. Provider-side input checks blocked requests for two of these tasks; all 36 others, containing 431 omissions, form the test set. EGRC, Generic, and Two-pass use these same 36 tasks, with prompts, source-review rules, and length-budget rules fixed before test evaluation.

Scholarly collection. Following DeepScholar-Bench's public construction procedure, we assemble 40 arXiv papers from January–June 2026, excluding papers in the original release. Related-work sections and reference-derived nuggets are unavailable during generation. All 23 tasks with identified omissions (98 findings) in this previously studied collection enter the full-method comparison, with requests, sources, original reports, and target findings held fixed. This comparison tests transfer to scholarly synthesis under a 600-word limit.

C.2 Generation and evaluation settings

Models. For DRB2 and the scholarly tasks, Qwen3.8-Max constructs the shared findings, initial reports, and original omission sets, and serves as the main writer. Table 4 gives the models used for full-method generation and evaluation. Generation uses the public request and acquired evidence; reference answers and scoring criteria are reserved for evaluation. Recovery assessors receive the findings whose expression they judge, whereas quality assessors do not receive the omission list or the reviser's decisions.

Table 4: Models for the main full-revision experiments. Effort levels refer to the requested reasoning setting. Generation and evaluation use separate calls.

Role	Model and effort	Relevant inputs
Planning and alignment	Qwen3.8-Max, medium	Request, report, full finding summaries; omissions added during alignment
Source reading and revision	Qwen3.8-Max, medium	Request, report, selected acquired passages and content requirements
Factual and scope diagnosis	DeepSeek-V3.2, thinking enabled	Proposed content and its acquired passages
EGRC source review	GLM-5.2, high for additions and content correction; max for final finding revision	Proposed content, original context, and acquired passages
Whole-report expression	GLM-5.2, max; independent assessment with GPT-5.5, high	Complete report, canonical findings, and supporting passages; no method labels
DRB2 report quality	Gemini-2.5-Pro	Report and benchmark criteria, three rounds
Scholarly quality	GPT-4o	Task abstract, report, and reference-derived nuggets or target text
Scholarly report-only check	GPT-5.5, high	Report and reference-derived nuggets, without the task abstract

The controlled-completion studies use Qwen3.8-Max (medium) for writing and DeepSeek-V3.2 for source verification. Complete expression is assessed by GLM-5.2 (max) on DRB2 and Qwen3.8-Max (medium) on the scholarly tasks. In these controlled-completion studies, the alternative writer is GLM-5.2 (max), and the alternative selector is DeepSeek-V3.2. The cross-writer omission analysis uses initial reports from GLM-5.2 and DeepSeek-V4-Flash (July 31 version), both at max effort. SciConBench uses Qwen3.8-Max (medium) for writing and DeepSeek-V3.2 for the source-grounding analysis.

The complete DRB1 agent combines agentic deep research using DeepSeek-V4-Pro with report writing using GLM-5.2, followed by revision and rewriting. Its before-and-after finding analysis uses DeepSeek-V3.2; its leaderboard quality score uses the official GPT-5.5 evaluator. External-method settings, including DeepWeaver's embedding model, are given below.

Generation and local comparisons. In the controlled-completion and full-revision experiments, Qwen, DeepSeek-V3.2, and GLM use reasoning-enabled generation and source-checking calls. Sampling uses provider defaults; temperature, top-p, and random seeds are not explicitly set. Qwen writing calls have completion limits of 16,384 tokens for planning and 32,768 for source reading and revision. These include internal reasoning where counted by the provider, while report budgets count only the answer. Planning permits up to 12 answer units, each linked to at most four acquired sources and 12 findings. Revision proposals contain at most 24 non-overlapping edits, with dependent changes grouped together. Full prompts and detailed model settings are provided in the supplementary materials.

On DRB2, Generic and Two-pass use the same Qwen writer, initial report, full finding summaries, and report ceiling as EGRC. The matched Autonomous and List-only revisions instead share EGRC's intermediate answer, source access, final revision opportunity, and source review. Autonomous chooses its own content; List-only receives the original omissions but does not require an explicit retain, revise, or defer decision for each finding. Its instructions and revision settings are unchanged between the 44-task validation and 36-task test comparisons.

Full-revision budgets. On DRB2, report length is measured in tokens using the `cl100k_base` tokenizer. Before comparison, we fix a common task-specific ceiling approximately 8% above the longest existing controlled Generic, Two-pass, or targeted-completion report. This ceiling is derived from generated reports, not benchmark reference answers. Generic and Two-pass receive an addition budget equal to 92% of the space remaining between the initial report and the ceiling, leaving

headroom for final admission. EGRC can also reorganize existing passages within the same ceiling. Shared evidence and report limits allow different computation and realized lengths, which we report with the comparisons. Scholarly reports have a common 600-word body limit, excluding references.

Controlled-completion budget. The DRB2 controlled-completion study sets the addition budget b to 25% of the initial report's token length, bounded between 48 and 8,192 tokens, using the same tokenizer. At most 32 candidate additions are generated. The longest prefix of whole candidates that fits this budget is submitted to source review. Source-accepted additions are then inserted under a 140,000-character final-report ceiling, which never binds in this study. Candidate order is preserved, and additions are never truncated. The supplementary materials provide the exact budget calculations and rounding rules.

Complete expression and recovery. An expression preserves the finding's material content: the relevant entities, quantities and units, comparison or relation, conditions, qualifications, and scope. Paraphrases and information distributed across report passages can qualify. A topic mention, a more general statement, or one part of a compound finding is insufficient when the omitted detail changes the finding's substance. The evaluator supplies report passages for a positive judgment or identifies the missing content for a negative one. Source support is assessed separately: a supported sentence need not express the complete finding. Whole-report recovery is the fraction of original omissions fully expressed in the final report, counting each finding once. Partially expressed content can still benefit a reader; report-quality assessments measure that broader utility.

DRB2 report quality. The released evaluator judges criteria in three dimensions: information recall, analysis, and presentation. Within each task and rating round, a dimension score is the fraction of its criteria satisfied. A returned score of 1 counts as satisfied; 0 and -1 do not. The total is the fraction satisfied across all criteria, rather than an unweighted mean of the three dimension scores. We average the three rounds within each task, then average tasks equally. Scores are displayed on a 0–100 scale.

Scholarly quality. Reference content measures coverage of the withheld reference-derived nuggets. Full support, partial support, and no support receive 1, 0.5, and 0, respectively. We average nuggets within each task and then tasks equally. The strict variant counts only full support; vital-nugget variants restrict the average to nuggets marked essential in the reference criteria. Organization is evaluated separately by comparing the generated report with the reference section twice, exchanging their presentation order. Its score is the mean preference for the generated report, not a direct EGRC–Generic win rate. Content and organization are reported separately. The GPT-5.5 report-only check uses the same nugget scoring rule and requires supporting quotations from the report; it does not reassess organization.

Aggregation and statistical analysis. Recovery also gives each task equal weight. Confirmed recovery credits only resolved positive judgments while retaining all original targets in the denominator. Scores are generally displayed to one decimal place; ranges whose endpoints round to the same value appear as a single number. Invalid-format judgments are retried with the same model and request; remaining unresolved items are kept unknown. The reported mean ranges assign unknown items their lowest or highest possible values. When a judgment is shared by two reports, it takes the same value on both sides of their comparison.

Confidence intervals summarize task-level sampling uncertainty in the mean paired difference. We resample task pairs 8,000 times to obtain the reported 95% intervals. For full-method quality comparisons, tasks are sampled without language stratification. Each resample averages the lower and upper paired differences separately; the interval spans the lower means' 2.5th percentile and the upper means' 97.5th percentile. Each full-method strategy contributes one final report per task; repeated quality ratings assess that same report. The controlled-completion study additionally repeats generation.

C.3 External Method Description

We reproduce the core procedures of Self-Refine and DeepWeaver within this study's shared-evidence setting, using the same writer and per-task report-length limits as EGRC. All 36 test tasks

and all 23 scholarly tasks are included. Both methods use Qwen3.8-Max with medium reasoning, the public request, full finding summaries, and passages from the shared acquired-source collection. The source context preserves passages supporting all findings, not only selected omissions. Optional surrounding passages are omitted when necessary to meet input limits. Neither external method receives R, reference answers, or evaluation criteria. The original R is used only by the common recovery evaluator. The comparisons evaluate the resulting reports for both recovery and quality.

Self-Refine. Following the feedback–revision procedure of Madaan et al. (2023), we start from the same y^0 as EGRC. In each cycle, the writer critiques missing information, correctness, clarity, organization, and citations, then revises the complete answer using its feedback. Earlier iterations remain available while they fit the context limit. We allow up to four cycles, with early stopping when no justified revision remains. If a cycle cannot be completed within its response limit, the last complete revision is retained. All tasks remain in the comparison.

DeepWeaver. We run the released thought-block stages of Wang et al. (2026) on the acquired passages: drafting, retrieval, subordinate-block generation, merging, filtering, expansion, and summary generation. It synthesizes an answer from sources rather than starting from y^0. The released stage prompts and rules for merging and filtering are retained, with the common Qwen writer and `text-embedding-v4` for passage retrieval. Source chunks contain up to 1,024 tokens; the released one- or two-round schedule depends on the number of source chunks. A final polish and, when needed, one length correction use the same task ceiling. One scholarly output remains at 606 rather than 600 words and is evaluated as returned.

D Controlled Studies of Selection and Expression

These experiments examine how target selection and expression affect the use of acquired evidence. They support the analysis in Section 3 and motivate EGRC's finding-level requirements.

D.1 DRB2 controlled completion

We compare answer-preserving additions under shared requests, initial reports, full finding context, writer, addition budget, source review, and insertion rules. Generic chooses useful additions freely. One-pass self-targeting identifies omissions and writes additions in the same call; Two-pass identifies them in a separate call before writing. Targeted completion (TC) receives the independently selected R_i. These comparisons isolate content selection, without the answer organization and local revision of the full EGRC method.

All 132 tasks enter the controlled comparison. The 26 empty-residual tasks contribute paired zeros to recovery, although Generic can still propose additions. Thus these recovery means equal $106/132$ times the nonempty-task means. This differs from the full revision comparison, which reports its nonempty-task population directly.

Table 5: Controlled candidate-level recovery and report quality. TC is the answer-preserving targeted-completion operator, not the full revised EGRC. Differences are recovery percentage points versus Generic. Recovery averages all 132 tasks, including paired zeros for 26 empty-residual tasks.

Selection	Recovery (%)	Difference	Quality
Generic	19.4	–	38.6
One-pass	23.6	4.3	38.1
Two-pass	26.3	6.9	38.3
TC	28.8	9.5	38.1

TC exceeds Two-pass by 2.6 recovery points in Table 5. Repeating writing and verification twice with the same target sets gives three-run mean gains of 8.5 points over Generic and 2.6 over Two-pass; TC leads Two-pass in each run.

The independent quality comparison gives a different picture: TC scores 0.5 points lower on total quality and 0.6 lower on information recall than Generic. These results motivate evaluating revision on quality as well as recovery, rather than treating the two outcomes as interchangeable.

TC fully expresses 513 original omissions versus Generic's 331, using about 190,000 rather than 266,000 added characters: 2.2 times as many recovered findings per character. This is a gain in writing efficiency, not computational savings; condition-specific calls consume 12.5M versus 11.8M tokens, excluding the shared research foundation.

Within the budget-eligible candidates, Generic names 982 original omissions; source checking retains 480, and 331 are fully expressed. The corresponding TC counts are 1,297, 696, and 513. Every source-accepted addition is inserted in both conditions.

Even giving Generic complete recovery credit for every original omission referenced by candidates excluded by the addition budget raises its task-macro recovery only to 19.6%. TC still leads by 9.3 points, so excluding these generated candidates does not account for the recovery advantage.

Target identity. We exchange every original residual for one disjoint same-task non-residual finding, retaining target count and format. Of the 1,343 matched pairs, 1,267 share deliverable and section, 73 share section, and three use a same-task fallback. A fresh semantic assessment evaluates both target classes under all three completions (Table 6). For this diagnostic, a finding is credited if at least one inserted addition fully expresses it, whether or not the addition declares a reference to that finding. The assessment excludes the original report and does not combine incomplete expressions across additions.

Table 6: Complete expression in additions under exchanged target sets. Each column evaluates a different finding set on all 132 tasks.

Completion	Original residuals (%)	Matched findings (%)
Generic	19.9	2.5
Original residual targets	29.7	0.1
Matched non-residual targets	0.0	24.5

Exchanging the target set redirects complete expression while the evidence, list size, and format remain fixed. The effect therefore depends on which findings are named, not merely on supplying a list. This is a test of content selection, not selector accuracy: the alternative set can contain already represented or irrelevant findings.

Alternative writers and selectors. We test whether the benefit of targeting persists beyond the main writer and selector. Changing only the completion writer to GLM-5.2 gives a 6.1-point declared recovery advantage over Generic and a -0.2-point total-quality difference. Declared recovery is the fraction of original omissions referenced by source-accepted, inserted additions, without a separate complete-expression judgment. A DeepSeek selector flags 317 candidate omissions across all 132 tasks; keeping Qwen writing and GLM complete-expression assessment gives a 7.5-point candidate-level recovery advantage and a -0.5-point quality difference. These comparisons test transfer of the content-selection intervention; the full revision method's task transfer is evaluated in Section 5.4.

D.2 Scholarly selection and expression repair

We test whether improving expression complements target selection on the 40-paper scholarly collection. Recovery is averaged equally over its 23 tasks with identified omissions. Targeted and Generic additions share the same sentence-length budgets and admission criteria: each candidate must pass source support, complete expression of every declared finding, nonredundancy, and citation checks. Source-grounded recovery is the fraction of original omissions referenced by such inserted candidates, counting each finding once. Unlike the full revision method, this operator preserves the original report and evaluates recovery through its additions.

Shared candidate repair was specified before candidate generation and applied symmetrically. A candidate is eligible when it passes source support but fails to express all its linked findings. Repair

rewrites that sentence using the same findings and source passages, without changing its destination or exceeding its original word count. Each request permits one retry for malformed or constraint-violating output; a candidate without a valid rewrite retains its original text. Both conditions are reverified under the same admission criteria before insertion.

Without repair, source-grounded recovery is 10.8% for Generic and 12.2% for targeted completion. Applying the same repair to both raises recovery to 19.0% and 34.0%, respectively. Targeted completion gains more from repair and leads by 15.1 points afterward. Repair produces valid rewrites for 81 of 86 eligible candidates; five keep their original text.

Two fresh verification passes reassess both original and repaired text. After repair, targeted completion retains an average 11.8-point recovery advantage over Generic. On reassessment, the repair gain for targeted completion exceeds that for Generic by 6.7 points, compared with 13.6 points under the original assessment. Thus the magnitude of this interaction depends on the assessment, while the post-repair targeting advantage persists. These results support checking expression alongside selection; the full-method scholarly comparison evaluates the resulting revision design on recovery and quality.

E JOINT RECOVERY AND REPORT-QUALITY RESULTS

Table 7 follows the method from answer organization to final revision. Request-planned revision organizes content around the request without aligning its plan to the original omissions. Residual-aligned revision adds this alignment and produces the common intermediate answer. Both variants use the same source-reading and correction design.

Table 7: Additional 44-task comparisons on the same original reports and 595 omissions. Quality and confirmed recovery use a 0–100 scale, with tasks weighted equally. Unresolved findings are not credited as recovered; their counts are reported separately.

Revision	Quality	Recovery (%)	Confirmed / unresolved
Initial report	40.0	1.6	18 / 24
Generic	42.0	35.1	160 / 22
Two-pass	41.7	37.6	184 / 17
Request-planned revision	43.8	10.6	85 / 10
Residual-aligned revision	43.8	19.1	111 / 30
Autonomous final revision	43.5	28.0	155 / 36
List-only final revision	43.5	40.2	215 / 47
EGRC	46.8	52.7	288 / 22

Three final revisers start from this intermediate answer and share the revision procedure. Autonomous chooses its own findings. List-only receives the original omissions but chooses which entries to address. EGRC instead maintains the same omissions as persistent revision requirements, enforcing an explicit decision for each finding. These comparisons test finding-level requirements within the same revision opportunity. Table 8 reports the corresponding 44-task contrasts; the same-list comparison also includes all 36 test tasks below.

Table 8: EGRC minus each comparator on the 44 validation tasks. Quality gains are score points; recovery gains are percentage points. All entries are task-paired mean differences. Ranges reflect unresolved judgments, not confidence intervals.

Comparator	Quality gain	Recovery gain
Generic	4.8	14.2–22.3
Two-pass	5.1	10.3–19.8
Autonomous final revision	3.3	19.5–28.6
List-only final revision	3.3	3.9–16.6

Comparison with the initial answer. The initial and revised reports use the same evaluator, prompts, and three-round averaging. EGRC improves quality over the initial reports by 6.8 points

on the 44 tasks and 5.9 points on the 36-task test set. Independent expression assessment recognizes some findings that the original selector marked as missing, giving nonzero initial recovery. This reflects differences between selection and expression judgments, not a prior revision.

List exposure and per-finding decisions. The list-only control receives the same intermediate answer, original omissions, finding context, sources, model, length ceiling, and review procedure as EGRC. It may choose which findings to revise, without recording a retain, revise, or defer decision for every original entry. The comparison includes all 44 validation and 36 test tasks. The List-only instructions and revision settings are the same in both collections; each comparison reuses the corresponding EGRC reports and assessments.

On the 44 tasks, final lengths differ by 0.3%, and the 3.3-point quality gain has a task-paired 95% interval of [2.2, 4.6]. Information recall leads the improvement (+4.1 points). On the test tasks, the quality gain is 3.9 points, again led by information recall, with a 0.7% final-length difference (Table 9). Both comparisons improve information recall and analysis, while presentation is lower. The benefit is therefore in recovering and explaining content, rather than presentation.

The test reports fully express 207 original omissions under EGRC and 123 under List-only, with 28 and 39 unresolved judgments, respectively. Their confirmed task-average recovery differs by 20.2 percentage points. Allowing the unresolved judgments to vary retains a gain of at least 12.8 points; Table 12 reports the paired quality and recovery intervals. The advantage thus persists beyond the validation set and does not depend on treating unresolved judgments as failed recovery.

Table 9: Same-list final revision on the 36 test tasks. Calls and inference tokens include returned format retries and exclude shared answer organization and outcome evaluation. Output tokens include reasoning.

Measure	List-only	EGRC
Total quality	38.7	42.6
Information recall	32.0	36.8
Analysis	48.4	52.3
Presentation	82.6	81.2
Confirmed recovery (%)	36.8	57.0
Mean final length (tokens)	16,042	16,154
Final-stage calls per task	2.08	2.28
Final-stage input tokens per task (k)	183.1	201.3
Final-stage output tokens per task (k)	11.2	16.4

Both methods share the revision procedure, including opportunities for length feedback and source review. Their proposals can trigger different numbers of calls, so near-equal final lengths need not imply equal inference costs.

A second writing model. We repeat the same-list comparison on 20 tasks sampled from the 36-task test set before generating the new reports: ten English and ten Chinese tasks, containing 230 original omissions. GLM-5.2 (max) replaces Qwen in planning, alignment, source-based writing, correction, and final revision. The acquired evidence, initial reports, selector, original omission sets, and report-length ceilings remain unchanged. Both final revisers start from the same newly generated GLM intermediate answer and share the revision and source-review procedure. GLM writing uses model-specific completion allowances of up to 65,536 tokens, including reasoning; these allowances are the same for both final revisers. The factual-feedback and source-review models are unchanged. Quality uses the same three-round Gemini assessment, while GPT-5.5 independently assesses whole-report recovery using the same complete-expression criterion.

The quality advantage extends to the second writer: EGRC gains 3.0 points (task-paired 95% interval [1.8, 4.4]), improving on 15 of 20 tasks (Table 10). Mean recovery is 13.5 percentage points higher. With final lengths differing by only 0.3%, these results support the value of explicit finding-level decisions with a different writer and comparable answer length.

Table 10: Same-list revision with GLM-5.2 on 20 DRB2 test tasks. All values are task means. Final-stage inference includes returned retries but excludes the shared answer-organization stage and evaluation; output tokens include reasoning.

Measure	List-only	EGRC
Total quality	42.2	**45.2**
Confirmed recovery (%)	41.6	**55.1**
Mean final length (tokens)	16,490	16,540
Final-stage calls per task	2.25	2.50
Final-stage input tokens per task (k)	179.9	197.6
Final-stage output tokens per task (k)	27.7	39.0

Realized length and inference. On the 44 DRB2 validation tasks, mean final lengths are 14,170, 14,357, and 15,283 tokens for Generic, Two-pass, and EGRC, using the shared budget tokenizer. Table 11 reports writing and source-review inference, including returned format retries. It excludes shared research, initial reports, target identification for both EGRC and Two-pass, and outcome evaluation. Output tokens include reasoning tokens.

Table 11: Inference per task on the 44-task DRB2 comparison. Tokens are in thousands. The final-revision rows exclude the common intermediate answer.

Revision	Calls	Input tokens	Output tokens
Generic	2.75	74.6	7.1
Two-pass	2.70	74.9	8.8
EGRC	14.95	799.0	75.6
Autonomous, final revision only	1.98	152.4	15.3
EGRC, final revision only	2.25	188.2	17.4

The lower two rows use the common 44-task y^1 and final revision procedure; counts reflect each method's realized calls and tokens. Final report lengths differ by 0.82%.

36-task test results. Table 12 gives the quality and recovery comparisons on the 36-task DRB2 test set under the fixed configuration. Generic and Two-pass complete the initial report; List-only shares EGRC's intermediate answer and final revision procedure. Three rating rounds assess the same final reports.

Table 12: 36-task DRB2 test results: task-paired EGRC differences. Quality is on a 0–100 scale; recovery differences are percentage points. Intervals combine task resampling and unresolved judgments.

Comparator	Outcome	Mean difference	95% interval
Generic	Quality	4.7–4.8	[3.0, 6.6]
	Recovery	30.2–44.8	[17.2, 56.0]
Two-pass	Quality	5.0	[3.4, 6.7]
	Recovery	22.1–36.9	[7.4, 50.4]
List-only	Quality	3.9	[2.7, 5.2]
	Recovery	12.8–26.9	[6.3, 35.0]

Quality improves on 30 of the 36 tasks against Generic, 32 against Two-pass, and 32 against List-only. Joint quality and recovery gains extend to these test tasks both for the full method and for the same-list comparison.

E.1 Sensitivity of finding recovery

Weighting tasks or findings. The main analysis weights research requests equally. A task with many extracted findings therefore has the same influence as one with few. Table 13 checks the alternative: pool findings within each collection, so that every finding receives equal weight. Statistical

resampling still keeps findings from the same task together. The finding-weighted advantage over autonomous revision remains positive on the 44-task DRB2 validation set and the 23 scholarly tasks. The advantages over Generic and Two-pass also remain positive in both DRB2 collections. These gains persist under either weighting scheme.

Table 13: Finding-weighted recovery: EGRC minus each comparator, in percentage points. Ranges reflect unresolved judgments. All tasks and original omissions in each collection are retained.

Tasks	Comparator	Mean difference
DRB2 (44)	Generic	17.8–25.2
DRB2 (44)	Two-pass	14.6–21.2
DRB2 (44)	Autonomous	18.3–24.0
DRB2 (36)	Generic	24.4–33.9
DRB2 (36)	Two-pass	8.4–17.6
Scholarly (23)	Generic	81.6–82.7
Scholarly (23)	Autonomous	30.6–32.7

Omissions identified by both selectors. On the 36-task test collection, the Qwen and DeepSeek selectors identify 57 common omissions across 23 tasks. We keep the reports and expression judgments unchanged and recalculate recovery on this shared set, giving each of the 23 tasks equal weight. Confirmed recovery is 71.6% for EGRC, 31.2% for Generic, and 49.1% for Two-pass. EGRC therefore retains higher mean recovery when the evaluation is restricted to omissions identified by both selectors.

A different expression assessor. For the comparisons in Table 14, we repeat whole-report expression assessment with GPT-5.5 (high reasoning) on all 44 validation, 36 test, and 23 scholarly tasks. The complete reports, findings, and source passages remain unchanged. Method-blind questions require report quotations for complete expression and distinguish the report from its supporting sources. The assessment also covers List-only on all 36 test tasks, reusing the unchanged EGRC judgments. Across these comparisons, 34 of 4,398 report–finding judgments remain unresolved and enter the bounds. The main recovery results use GLM-5.2; this additional assessment checks whether the recovery advantage persists with a different expression assessor.

Table 14: Recovery under the independent GPT-5.5 assessment. Task-average differences are EGRC minus each comparator, in percentage points. Ranges reflect unresolved judgments.

Tasks	Comparator	Mean difference
DRB2 (44)	Generic	15.9–19.9
	Two-pass	4.7–8.0
	Autonomous	24.6–26.0
DRB2 (36)	Generic	34.5–36.3
	Two-pass	8.3–9.4
	List-only	16.7–18.4
Scholarly (23)	Generic	78.5
	Autonomous	29.5

With the different assessor, EGRC again has higher mean recovery than every comparator in the table. The test-set advantage over Two-pass is 8.3–9.4 points, compared with a 27.4-point difference between the main confirmed-recovery means. The direction persists, but the magnitude depends on the assessor. Agreement on resolved judgments is 94–95% across the three collections; this overall agreement does not imply equal recovery gaps.

The same-list test gives 58.9% confirmed recovery for EGRC and 41.6% for List-only, a 17.3-point difference. With unresolved judgments included, the mean advantage is 16.7–18.4 points and its task-paired 95% interval is [10.0, 26.1]. The independent assessor thus supports the recovery gain from per-finding decisions after both revisers receive the same omissions.

E.2 INDEPENDENT QUALITY IN SCHOLARLY SYNTHESIS

Generic makes one untargeted local revision of the original answer, with the complete findings, acquired abstracts, and the common 600-word ceiling. It is asked to identify useful missing content itself; no residual list is provided. Source and retention checks admit changes in 10 of the 23 reports; the others retain their original text. This comparison evaluates the full method against one-pass revision. To test the continued value of finding requirements, the matched comparison below holds the intermediate answer and final revision opportunity fixed.

Reports for the same 23 tasks are evaluated against reference-derived nuggets and the withheld reference section. Table 15 compares EGRC with Generic. Relative to the initial reports instead, EGRC gains 13.4 reference-content points and 17.4 organization points. The initial reports have one confirmed expression among 98 original targets, giving 0.9% task-average recovery. As on DRB2, this nonzero starting value reflects expression assessment being separate from omission selection.

Table 15: Reference-based scholarly evaluation, on a 0–100 scale. Differences are EGRC minus Generic. Organization scores are based on comparisons of each method's report with the reference section. The outcomes are reported separately, without an aggregate quality score.

Measure	Generic	EGRC	Difference
Reference content	44.1	55.8	11.7
complete matches only	30.8	42.9	12.1
Important reference content	56.4	69.7	13.3
complete matches only	41.9	58.8	16.9
Organization	56.5	71.7	15.2

The reference-content advantage persists when only complete matches receive credit and when evaluation is restricted to important reference content. It therefore extends beyond partial mentions or less important reference content. The following check asks whether the advantage also persists when the assessor cannot draw on the task abstract.

Sensitivity to the evidence shown to the assessor. The reference-content assessor receives the task abstract as well as the report. An empty-report diagnostic still receives full credit for 128 of 323 items and partial credit for 42, indicating that the abstract can supply credited content. This does not identify every error in a complete report's score or determine the direction of a method difference. We assess all 23 tasks and all 323 original reference nuggets with GPT-5.5, using report-only questions that require supporting quotations. The reports are unchanged and include the external comparisons, initial answers, reference sections, and an empty report for each task. The empty reports receive no credit on any of the 323 items. Table 16 retains partial matches and unresolved judgments under the same task-level aggregation.

Table 16: Report-only reference-content sensitivity with GPT-5.5, without the task abstract. Differences are EGRC minus each comparator. Scores and differences use a 0–100 scale. The reference and empty report are controls, not competing methods.

Report	Reference content	Difference
Initial report	23.5	15.0–15.6
Generic	22.6	15.9–16.5
Intermediate answer	24.1–24.9	13.6–15.0
Autonomous revision	26.3	12.2–12.8
Self-Refine	31.7	6.8–7.4
DeepWeaver	34.5–36.4	2.1–4.6
EGRC	38.5–39.1	–
Reference section	80.3–80.7	–
Empty report	0.0	–

Without access to the abstract, EGRC retains the highest mean reference-content coverage among the compared methods, ahead of Generic, autonomous revision, and both external methods. The content advantage thus extends beyond recovery of the selected omissions and remains present under

report-only assessment. The assessments differ in model, abstract visibility, and quotation requirements, so their absolute scores are not interchangeable. Organization is not reassessed here. For the matched comparison with autonomous revision, the report-only gain is 12.2–12.8 points, with a task-paired 95% interval of [7.0, 18.3].

Matched scholarly revision. To test the same design under the scholarly budget, both final revisers receive the same intermediate answer, full finding context, acquired sources, model, and 600-word ceiling. EGRC retains the original omission list and must record a decision for every entry; Autonomous identifies and records useful missing content itself. Source and retention checks are shared. This changes the terminal finding requirements, not the upstream use of the residual. The comparison in Table 17 includes all 23 tasks and the same 98 original omissions. Unchanged full-method reports and identical assessment requests are reused.

Table 17: Matched scholarly revision on the same intermediate answers. Recovery ranges reflect unresolved judgments. Quality outcomes use the reference-based measures above. Differences are EGRC minus autonomous revision; paired intervals describe uncertainty in the mean differences.

Measure	Autonomous	EGRC	Difference	95% interval
Finding recovery	52.9–53.5	86.5–87.1	33.0–34.2	[16.0, 51.3]
Reference content	47.0	55.8	8.8	[3.8, 13.5]
complete matches only	34.8	42.9	8.1	[3.1, 12.9]
Important reference content	60.6	69.7	9.1	[2.3, 15.6]
complete matches only	47.6	58.8	11.2	[2.8, 19.4]
Organization	50.0	71.7	21.7	[8.7, 37.0]

EGRC fully expresses 81 findings and autonomous revision 50, with one unresolved judgment in each. At nearly equal report length, the recovery advantage is accompanied by higher reference-content coverage and organization scores; the report-only check retains the content advantage. The matched starting answer and revision opportunity link this benefit to persistent finding requirements rather than to another revision pass.

E.3 Which findings does revision recover?

We follow each finding in the original omission set R from the common intermediate answer y^1 into the EGRC and autonomous revisions, on the 44 DRB2 validation tasks and the 23 scholarly tasks. This distinguishes completing a finding still incomplete after answer organization from retaining one already expressed there. Table 18 conditions on the expression judgment of the same starting answer, not on either final strategy's result. Counts include every original target with a resolved starting judgment; the 30 unresolved DRB2 starting judgments remain separate. The scholarly starting judgments are all resolved.

Table 18: Finding transitions within the original omission set R, from the same intermediate answer. Entries give confirmed counts over the common starting-state denominator, not task-averaged recovery. A final unresolved judgment is not counted as either successful expression or a lost finding.

Tasks	Revision	Incomplete → complete	Complete → complete
DRB2 (44)	Autonomous	48/454	106/111
DRB2 (44)	EGRC	170/454	109/111
Scholarly (23)	Autonomous	31/79	19/19
Scholarly (23)	EGRC	63/79	18/19

Within R, EGRC completes more than three times as many findings left incomplete in y^1 as autonomous revision on DRB2, and about twice as many on the scholarly tasks. Both revisers retain most targets already expressed in y^1. The principal observed gain is thus completion of remaining omissions after answer organization.

Findings that remain incomplete. Table 19 separates the revision decisions behind resolved incomplete outcomes. A finding can be deferred, considered already expressed by the reviser, or

linked to an edit that is not admitted or remains incomplete. The table describes recorded decisions, not human-validated causes: a deferral may be justified, and a retain decision can disagree with the independent expression assessment. Categories are mutually exclusive. For a finding linked to several edits, the applied category requires all linked edits to be applied. Otherwise, a source-review rejection takes precedence over a length rejection; the length category requires all linked edits to pass source review. Length admission accounts for only a small part of the remaining omissions in these reports.

Table 19: How unrecovered findings were handled. Counts include findings judged incomplete in the final EGRC report. The 22 and 28 unresolved expression judgments in the two collections are kept separate.

Revision decision or outcome	DRB2 (44)	DRB2 (36)
Reviser considered the finding already expressed	81	25
Explicitly deferred	94	103
Proposed edits not admitted by source/retention review	77	42
Source-accepted edits not admitted by the length limit	11	3
Edits applied, but expression still incomplete	22	23
Total resolved incomplete findings	285	196

E.4 Quality across evaluation dimensions

Mean information recall, analysis, and presentation all increase against Generic and Two-pass in the 44-task and 36-task DRB2 collections (Figure 5). Thus, the total-quality gain does not come from one dimension offsetting a decline in another. For scholarly synthesis, gains likewise extend across reference-content coverage and organization, both against Generic and in the matched revision comparison (Tables 15 and 17).

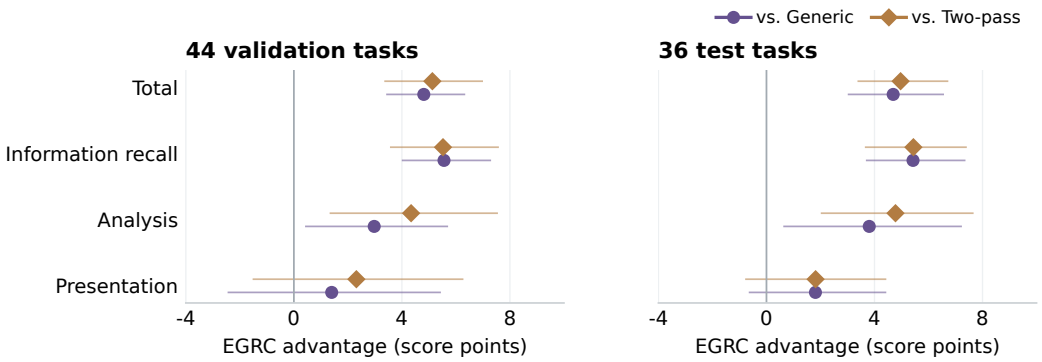

Figure 5: Quality gains over Generic and Two-pass, in score points, on the 44-task and 36-task collections. Symbols mark the lower mean difference allowed by unresolved judgments; thin lines show task-paired 95% intervals including those judgments.

F Comparisons with external methods

We compare EGRC with Self-Refine's iterative self-feedback and DeepWeaver's evidence synthesis on report quality, omission recovery, and computation. The implementation settings are specified in Appendix C.3.

Report quality and recovery. EGRC combines the highest mean overall quality on DRB2 and all-reference coverage on scholarly tasks with substantially higher omission recovery (Tables 20–22). Its content gains extend to information recall and analysis on DRB2, and to complete-match reference coverage on scholarly tasks. The joint improvement therefore includes content measures evaluated independently of the omission list. Presentation scores are higher for the external methods on DRB2, and DeepWeaver scores higher on scholarly organization.

The contrast between broad quality and specific recovery is informative. Benchmark criteria and reference nuggets can credit content already present in the initial report or outside its selected omissions. Recovery instead asks whether the same original omissions are fully expressed by each method, allowing paraphrases and expression across several report passages. A report can therefore cover much of the reference content while leaving those particular findings incomplete. The scholarly results retain this pattern when reference coverage counts only complete matches, rather than giving partial credit for incomplete mentions. EGRC's reference-content advantage over both external methods also persists under report-only assessment (Appendix E.2).

These results motivate evaluating answer organization and finding recovery together: improving broad report criteria does not by itself ensure that specific missing details reach the answer. EGRC couples the two, improving content-related quality scores while recovering more of the original omissions. The matched final-revision experiments complement these full-method comparisons by testing persistent finding requirements under a shared starting answer and revision opportunity. On the same 36 DRB2 tasks, the List-only comparison additionally gives both revisers the original omissions and still finds higher quality with per-finding decisions.

Table 20: External-method comparisons on the 36 DRB2 tasks. All values are on a 0–100 scale. Confirmed recovery uses the same 431 original omissions. Scores are task averages.

Method	Total	Info. recall	Analysis	Presentation	Recovery
Self-Refine	39.3	33.0	46.6	88.1	6.2
DeepWeaver	39.5	32.4	50.5	84.3	8.0
EGRC	42.6	36.8	52.3	81.2	57.0

Table 21: Reference-based scholarly quality for the same 23 tasks. All and Vital credit complete matches as one and partial matches as one half; Strict counts complete matches only. Organization is preference against the withheld reference, not a direct pairwise method comparison. All scores are percentages.

Method	All	Strict all	Vital	Strict vital	Organization
Self-Refine	51.9	39.3	69.1	56.7	67.4
DeepWeaver	51.1	40.4	67.2	56.9	84.8
EGRC	55.8	42.9	69.7	58.8	71.7

Table 22: Task-paired EGRC gains over each external method. Quality denotes DRB2 total or scholarly all-reference coverage. Positive differences favor EGRC; ranges reflect unresolved evaluator judgments.

Tasks	Comparator	Outcome	Mean difference
DRB2 (36)	Self-Refine	Quality	3.2–3.3
		Recovery	49.6–60.3
	DeepWeaver	Quality	3.1
		Recovery	46.1–58.6
Scholarly (23)	Self-Refine	Quality	3.9
		Recovery	81.4–82.0
	DeepWeaver	Quality	4.7
		Recovery	78.8–80.0

All methods use the same complete-expression criterion. On DRB2, confirmed findings number 207 for EGRC, 21 for Self-Refine, and 30 for DeepWeaver, with 28, nine, and 12 unresolved. On scholarly tasks, the counts are 81, seven, and 11, with one, zero, and one unresolved. Unresolved evaluator output is not counted as a negative content judgment.

Computation. Table 23 reports realized answer lengths alongside returned generation and source-review calls, including returned format retries. These writing-stage costs exclude shared retrieval,

finding extraction, initial-report construction, EGRC target identification, and outcome evaluation. DeepWeaver synthesizes a report from sources, whereas EGRC revises an initial report, so the totals describe different writing strategies rather than complete-agent costs. Token counts use returned provider usage, including reasoning where counted. Because model mixtures differ, equal token counts need not imply equal cost or computation.

Table 23: Mean writing-stage inference and final answer length per task. Input and output tokens are in thousands. Final length uses thousands of report tokens for DRB2 and body words for scholarly synthesis. DeepWeaver's additional embedding calls and input tokens are not included.

Tasks	Method	Calls	Input tokens	Output tokens	Final length
DRB2 (36)	Self-Refine	8.0	991.9	45.7	8.4
	DeepWeaver	36.0	1660.9	124.5	9.6
	EGRC	15.4	845.2	75.5	16.2
Scholarly (23)	Self-Refine	7.9	83.8	28.8	492
	DeepWeaver	17.0	83.3	30.1	530
	EGRC	12.3	75.5	36.6	521

The common ceiling permits different realized answer lengths, particularly on DRB2. These comparisons therefore evaluate complete writing strategies. The matched revisions in Sections 5.3 and 5.4 isolate finding requirements with shared starting answers and nearly equal final lengths.

DeepWeaver additionally uses 27.8 embedding batches and 86.4 thousand embedding-input tokens per DRB2 task, and 7.2 batches and 5.3 thousand tokens per scholarly task. On long reports, its sequential expansion and summary stages repeatedly condition on earlier content. EGRC uses fewer generation and source-review calls than DeepWeaver and fewer input tokens than either external method in both collections. Self-Refine uses fewer calls, illustrating a different trade-off between call count and input volume.

G HUMAN VALIDATION

The human studies answer three different questions: whether the target is supported, relevant, and missing; whether a particular text expresses a finding; and which complete report readers prefer. The source, selector, and candidate checks validate the judgments used in the controlled analysis. The full-method assessment then examines the revised reports directly, combining reader preferences with verification of recovered findings.

Source and selector judgments. Two independent annotators assess 192 candidate–source comparisons, balanced by language and completion strategy. They agree on entailment for 177 items (92.2%, AC1 0.912) and on relevance for 158 (82.3%, AC1 0.748). Both readers confirm source support for all 101 candidates accepted by the model in this sample. Some model-rejected candidates also receive positive human source-support judgments.

The selector checks compare selected and nonselected findings. In an initial 48-finding sample, both readers judge 17 of 24 selected findings and one of 24 nonselected findings to be supported, relevant, and missing from the original report. A separate, non-overlapping sample confirms 40 of 64 selections and four of 64 non-selections, a 56.3-point difference. Agreement on grounding, relevance, and absence in this second sample is 84.4%, 87.5%, and 92.2%, respectively. Both samples thus show that the selector concentrates human-confirmed omissions in the selected set.

Candidate expression. A separate 192-item sample presents candidate–finding pairs with their exact support spans, without strategy or model labels. Agreement is 181/192 (94.3%, Cohen's $\kappa = 0.884$). Each annotator confirms complete expression for 95 of 96 pairs labeled positive by the model.

Calibration-based sensitivity analysis. We use these expression judgments to check how human calibration changes the recovery estimates. Candidate–finding pairs are grouped by strategy, language, and model verdict. Within each group, we estimate the probability of complete expression, giving the two readers equal weight.

For each strategy, let A_i contain the additions actually inserted into report i, and let $F(c)$ denote the findings declared as references for addition c. The calibrated probability that c fully expresses e is $\hat{p}_{ice}$. Assuming conditionally independent expression outcomes across candidates for the same finding, projected recovery for $R_i \neq \varnothing$ is

$$\widehat{r}_i = \frac{1}{|R_i|} \sum_{e \in R_i} \left[1 - \prod_{\substack{c \in A_i \\ e \in F(c)}} (1 - \hat{p}_{ice}) \right].$$

The bracketed term is the probability that at least one inserted addition fully expresses the finding, so multiple references do not double-count it. An empty product equals one, giving zero credit when no inserted addition references a finding. Empty-residual tasks receive zero, and all 132 tasks are averaged equally. The projection assumes that the sampled calibration rates apply to other pairs in the same group.

This gives estimated task-average recovery of 21.2%, 23.0%, 26.9%, and 30.2% for Generic, One-pass, Two-pass, and TC. The ordering is preserved: TC retains a 9.0-point mean advantage over Generic and a 3.3-point advantage over Two-pass. These calibrated sensitivity estimates complement the direct whole-report judgments below.

Reader preferences for controlled completion. Two readers compare the same 48 TC–Generic report pairs, with language and residual-load coverage and reversed left/right presentation. The reports come from the controlled completion study. Across both readers, overall preferences are 51 TC, 35 Generic, and 10 ties. Net coherence preference is 24.0 points, and lower-redundancy preference is 42.7 points. Completeness judgments are close overall, with 38 preferences for TC, 42 for Generic, and 16 ties, and vary with residual load. In this controlled setting, targeting's reader benefits are most apparent in coherence and reduced redundancy. The following study tests the full revision method.

G.1 Full-method report assessment

Two independent annotators, neither involved in this research, conducted the assessment. Both held at least a bachelor's degree in computer science and could independently read the Chinese and English task materials. They participated voluntarily with informed consent.

Two independent readers assess 24 DRB2 pairs against Two-pass, with 12 tasks per language, and all 23 applicable scholarly pairs against Generic. Two-pass tests EGRC against a writer that identifies its own omissions; Generic tests it against untargeted revision under the scholarly 600-word limit. These comparisons assess reader value in the two task settings, and their preferences are reported separately.

The DRB2 tasks are sampled from the 44-task validation set using a fixed pseudorandom order within each language, selecting 12 tasks per language without using quality scores or expression judgments. Both readers see the same 47 task pairs in different orders with opposite left/right assignment; method names, model judgments, and benchmark criteria are hidden. The full reports are provided with navigation through every differing passage. Readers first judge overall preference, completeness, usefulness, coherence, and nonredundancy, and save these judgments before inspecting target findings. This ordering keeps the initial reading task focused on the answer as a whole, before asking whether particular omissions have been recovered. The supplementary materials list the sampled tasks and findings, together with the fixed sampling seeds.

The expression task selects up to two original omissions per sampled DRB2 task and one per scholarly task, again using a fixed pseudorandom order without model-verdict filtering. This yields 43 DRB2 and 23 scholarly findings. Each finding is judged against both complete reports. Full-expression judgments require report quotations; incomplete judgments identify the missing substance. Finding text and source passages remain in their original language, with Chinese instructions. The task permits paraphrases and expression distributed across passages, following the criterion in Appendix C.

Report preferences. Table 24 reports all five dimensions. Each reader favors EGRC overall in both settings. They agree on the overall choice for 19 of 24 DRB2 pairs and 14 of 23 scholarly pairs, with a consistent aggregate direction despite differences on individual reports.

Table 24: Full-method reader judgments by dimension. The comparator is Two-pass for DRB2 and Generic for scholarly synthesis. Each row includes both readers: 48 judgments on 24 DRB2 pairs or 46 on 23 scholarly pairs. No returned judgment uses the available unable-to-judge option.

Tasks	Dimension	EGRC	Comparator	Tie
DRB2	Overall	41	6	1
	Completeness	42	6	0
	Usefulness	41	6	1
	Coherence	33	5	10
	Nonredundancy	33	2	13
Scholarly	Overall	34	10	2
	Completeness	33	10	3
	Usefulness	33	9	4
	Coherence	26	8	12
	Nonredundancy	7	15	24

We assign overall preference $+1$ for EGRC, -1 for the comparator, and zero for a tie. We average the two readers within each task, average tasks equally, and multiply by 100 to report net preference in points. Mean net preference is 72.9 points on DRB2 and 52.2 on scholarly synthesis. The corresponding 95% intervals are [50.0,91.7] and [26.1,76.1]. We resample tasks with both readers kept together, preserving the DRB2 language strata. This measures task-to-task variation with the reader panel held fixed. Completeness and usefulness also favor EGRC in both settings.

Scholarly nonredundancy has a different pattern: more judgments favor Generic, although ties are the most common response. Among the 15 judgments favoring Generic for nonredundancy, the same reader still prefers EGRC overall in 10. Readers thus distinguish repetition from the overall value of an answer: they can prefer EGRC's report even when they find Generic less repetitive.

Human-confirmed recovery. Both readers confirm greater recovery under EGRC (Table 25). The table counts findings judged fully expressed in each report. To obtain task-average recovery, we average finding judgments within each task and reader, average the two readers for each task, and then average tasks equally. Thus a DRB2 task with two sampled findings receives the same weight as one with a single finding. On these sampled omissions, EGRC gains 34.4 points over Two-pass on DRB2 (95% interval [12.5,54.2]) and 93.5 over Generic on scholarly synthesis [82.6,100.0]. Intervals use the same task-level bootstrap as the preference estimates.

We compare human and model assessments on exactly the same sampled findings. For the 43 DRB2 findings, the model-based EGRC–Two-pass gap is 25.0–35.4 points, alongside the human estimate of 34.4. On the scholarly sample, the corresponding estimates are 91.3 and 93.5 points. The recovery advantage is therefore also present when readers, rather than a model, determine whether the report expresses each finding.

Table 25: Complete expression of sampled original findings in the full reports. The comparator is Two-pass on DRB2 and Generic on scholarly tasks. Each reader judges the same findings on both sides; denominators are sampled findings, not independent reports.

Tasks	Reader	EGRC	Comparator
DRB2	A	32/43	16/43
	B	31/43	16/43
Scholarly	A	22/23	0/23
	B	22/23	1/23

Readers agree on 128 of 132 report–finding pairs (97.0%, Cohen's $\kappa = 0.939$): 83 of 86 for DRB2 and 45 of 46 for scholarly synthesis. Readers A and B also agree with the model on 124 and 122 of the 129 pairs with a resolved model judgment (96.1% and 94.6%, respectively). All human

judgments contribute to the recovery estimates, including the three pairs left unresolved by the model. Together, the preference and expression studies support both reader value and recovery of specific findings in the revised reports.

H EXAMPLES OF RECOVERED FINDINGS

One DRB2 validation example and one scholarly example show how a report can discuss a topic and cite its source while leaving a specific finding incomplete. The first restores a conditional payout rule; the second restores a missing comparison within an existing sentence. Table 26 summarizes what changes in the evaluated reports.

Payout conditions in a retirement report. A report on Asian retirement savings draws on an acquired OECD passage about Indonesia's voluntary defined-contribution plans.[1] The shared intermediate answer and autonomous revision describe plan structure and retirement age, but leave out the balance-dependent payout conditions. EGRC adds the three parts of the selected finding: a full lump sum for balances below IDR 500 million, a 20% lump-sum share otherwise, and periodic payment of the remainder for at least ten years. These details are integrated into the report's discussion of contributions and benefits, using the same acquired source. The difference matters to the reader: naming a payout phase does not explain how the account balance changes the permitted form of payment under the cited rule. Recovering the condition, proportion, and duration supplies that missing explanation.

A missing comparison in scholarly synthesis. A related-work section on retrieval cites the *DeeperImpact* abstract.[2] Both the intermediate answer and autonomous revision contrast learned sparse indexes with traditional rankers and dense retrievers. The source also describes an effectiveness gap to sparse methods using more expensive optimizations, including query expansion and query term weighting. EGRC completes the existing comparison, stating that learned sparse indexes "still lag behind the best dense retrievers, or even sparse methods leveraging query expansion and term weighting". The added clause restores a comparison within an already discussed topic, rather than adding another topic to the outline. The revision makes the comparison with enhanced sparse methods explicit while retaining the surrounding discussion.

Table 26: Content changes in two matched-revision examples. Entries summarize the evaluated reports. In each case, the shared intermediate answer and autonomous revision both leave the selected finding incomplete.

Case	Shared answer and autonomous revision	EGRC revision
Retirement report	Describe the pension scheme and retirement age, but omit the balance-dependent payout rule.	Adds the balance threshold, permitted lump-sum share, and minimum duration of the remaining payments.
Scholarly synthesis	Compare learned sparse indexes with traditional rankers and dense retrievers, but omit the comparison with enhanced sparse methods.	Extends the existing comparison to sparse methods using query expansion and term weighting.

Both examples concern evidence that was already available and topics that were already represented in the answer. Their remaining gaps are finer than an outline entry: one is a conditional rule, the other a comparison class. They illustrate why finding-level requirements remain useful after answer organization, by keeping these details explicit during revision.

I SCIENTIFIC CONCLUSIONS AND THE COMPLETE DRB1 SYSTEM

Scientific synthesis. The SciConBench study uses 40 development tasks and 268 subsequent scientific questions. Question-driven retrieval, extraction, and an initial conclusion supply shared evi-

[1] Acquired source: OECD, *Addressing the Challenges for Asset-backed Pensions in Indonesia*, Section 1.1.3.
[2] Acquired source: *DeeperImpact: Optimizing Sparse Learned Index Structures*.

dence for structured and generic completion; reference conclusions are withheld. Structured completion names the original omissions, whereas Generic chooses its own additions. A subsequent source-grounding analysis evaluates the saved additions without modifying the reports. An original omission is credited when the verifier identifies its source passages as supporting at least one fully source-supported, inserted addition. Task-level realization is the fraction of original omissions credited in this way, counting each omission once. No separate assessment of complete expression is applied.

Among the 268 subsequent questions, 200 have identified omissions. On these tasks, task-average source-grounded realization is 56.8% for structured completion and 10.0% for Generic, a 46.8-point difference. With source acquisition held fixed, this comparison extends the evidence for selective use of acquired findings to scientific conclusion writing.

DRB1 system relevance. The complete system was evaluated on all 100 DRB1 tasks, scoring 55.99 on the official GPT-5.5 evaluation and ranking first in the September 2, 2026 leaderboard snapshot (DeepResearch Bench Team, 2026). Its dimensions are 56.37 comprehensiveness, 56.77 insight, 55.92 instruction following, and 53.27 readability. This score evaluates the complete research-and-synthesis system. The following analysis traces how selected findings reach its final reports.

We compare intermediate and final reports on 99 of these 100 tasks; one task lacks the intermediate evidence representation needed for this analysis. Of the 99 tasks, 92 have selected findings, totaling 435. The assessment requires report passages to convey each finding's substantive content, rather than merely mention a related topic or citation. Task-average realization across the 99 tasks rises from 14.37% to 82.21%; the seven tasks without selected findings contribute zero at both stages. Here realization includes all selected findings, whereas the main recovery measure considers only original omissions.

The finding transitions explain this increase. Synthesis expresses 332 of the 375 findings not substantively expressed in the intermediate report, while retaining 56 of the 60 already expressed. The increase therefore combines recovery of previously omitted content with retention of most previously expressed findings.

On 87 of the 92 tasks with selected findings, the analyzed output is identical to the submitted report, covering 408 of 435 findings. The other five tasks contain 27 selected findings. Even assigning all 27 zero expression in the submitted reports leaves a mean of at least 81.28% over the same 99 tasks, with the seven empty-target tasks again counted as zero. Thus the high realization is also supported by findings traced to the submitted reports.

The common principle is to keep acquired findings explicit during synthesis. The complete DRB1 system combines this with further research and rewriting; its outputs are a median 2.37 times as long as the intermediate reports. The before-and-after analysis shows that substantial recovery accompanies high system-level performance. EGRC's matched-revision studies answer the more specific design question: whether persistent finding requirements add value with the starting answer, sources, and revision opportunity shared.