How Report Formation Shapes Evidence Use in Multi-Agent Reasoning

Anonymous authors
Paper under double-blind review

Abstract

Giving a final language model every source document may seem to make evidence sharing among upstream agents redundant. We test this assumption by varying source access during report formation while fixing the final sources and the agents' initial answers and confidence values. Across 249 task variants from 128 synthetic decision scenarios, sharing during a report update improves final accuracy by 18.36 percentage points. We exchange naturally generated explanation records, statements relating cited passages, independently of their automatically copied source text. Explanations formed after sharing improve accuracy by 13.48 points under fixed copies. Thus sharing changes useful report content beyond the source passages explicitly repeated for the final model. We next retain the earlier report and supply a correct, reference-derived judgment of whether a rule applies to the current case. This diagnostic revision improves decisions, although the predicted reverse effect of leaving applicability unresolved is not established. Finally, both a correct and an unresolved interpretation remain present at matched positions. Designating the correct interpretation as the report's current judgment rather than an unincorporated external annotation improves accuracy by 9.18 points. Benchmark studies examine relevance under separately specified information controls. Together, these interventions show that complete final access leaves report formation consequential: generated explanations and their assigned report roles affect how available evidence is used.

1 Introduction

A multi-agent reasoning system can divide its sources among several agents and give their reports, together with every original document, to a separate final model. In such a system, does sharing evidence among the agents still matter? If its benefit came only from making missing facts available at the final decision, providing all originals would remove that advantage. But agents may use shared facts to form interpretations that change the final model's decision even when it can read those facts itself. Distinguishing these accounts matters for deciding whether gathering documents at the end can replace communication during report preparation.

Consider a provider barred from a job unless its permit qualifies for an exemption. One agent sees the permit and current job; another sees the exemption rule. Their reports may describe the observations and rule separately. Sharing the rule lets the first agent assess whether the permit qualifies for the current job. We call this case-specific assessment an *applicability judgment*. Even when a final model receives all original sources, it can encounter reports that express different judgments about the same case. The question is whether the way those reports were formed changes its decision.

Evidence exchange can improve collective reasoning (Du et al., 2024; Li et al., 2026), while conversational history can impair the use of information supplied across turns (Laban et al., 2026). We ask a more specific question: what does sharing contribute through a report when a separate final model receives the same complete sources and explicit recommendations? We fix each agent's displayed *initial recommendation*, its first answer and confidence value, and intervene on the report content that remains free to change. This separates the effect of report formation from changes in final factual access or those explicit answers.

We first establish that source access during report formation matters under these controls. We then locate useful variation by exchanging naturally generated report components, before testing how a

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.

supplied interpretation changes the use of an intact report. Figure 1 summarizes what each comparison identifies.

The reports contain *explanation records*, short generated statements that relate cited source passages, and verbatim source copies attached by the system. Exchanging these components independently gives a 13.48-point advantage for explanations formed after sharing under fixed copies. This is the central content result: the benefit is carried by naturally generated records, beyond changes in explicitly copied evidence.

To examine how such content is used, we preserve the earlier report and supply a correct applicability judgment derived from the task reference. This diagnostic revision improves decisions. A further comparison holds both a correct and an unresolved interpretation present, and varies which is labeled current within the report versus external and not incorporated. Its role effect persists at matched text positions. The latter comparison asks how supplied content is used, beyond whether a correct interpretation is available.

Our contributions are three connected findings. We establish a source-sharing effect with complete final sources and fixed initial recommendations. We locate decision-relevant variation in naturally generated explanation records under fixed source copies. We characterize how explicit revision and the assignment of a current-judgment role change the use of report content. HiddenBench, two separately evaluated M3GQA cohorts, and three recorded model configurations provide complementary evidence under their respective information controls.

2 Experimental Setting

Agents, reports, and the final model. Four separately prompted language-model agents prepare reports for a separate model call that selects the final answer. Each agent initially receives sources shared by all agents and one *private source*, a document initially available only to that agent. Its report contains an answer, confidence, selected source references, and explanation records. Each record identifies two source passages, the subject and property being discussed, their relationship, and a short explanation. A *source link* is this explicit pair of cited passages. The system attaches their verbatim text and copies of selected sources; these deterministic quotations are the *copied source text*. They can repeat material already present in the final model's complete originals.

The decision task. Each task asks for the eligible provider with the greatest verified throughput, measured in units per hour, among three candidates. Shared sources give their throughputs and identify two unrestricted providers. The fastest provider has a restriction whose applicability depends on current observations. One private source describes those observations; another gives the applicable rule. The remaining private sources reinforce the use of verified throughput and current observations. Connecting the rule to the observations determines whether the fastest provider is eligible. Appendix A gives the allocation and rule families, which are templates defining different applicability conditions.

Two report updates. An update is a new agent response based on its previous response, original sources, and the other agents' preceding answers and confidence values. Source sharing additionally supplies one other agent's private document. The agents follow a fixed cycle, each receiving its predecessor's document, so the agent holding the current observations receives the restriction rule. We compare sharing during both updates, the second update only, and neither update. Every path reuses the same initial responses; the latter two also reuse the same first update. The comparison therefore changes source access, not the number of report-update calls. All original documents reach the final model in every path.

Identical final sources and recommendations. The final model always receives the full original-source collection. Its displayed agent answers, confidence values, and the accompanying summary of those answers come from the common initial responses. The explanations and copied source text come from the selected report-update arrangement. Its explicit recommendation fields consequently stay fixed. The final model is instructed to use all sources and treat reports as fallible interpretations. It selects an answer directly; the decision rule is language-model synthesis.

Writing E for the original sources, A for the fixed initial recommendations and confidence values, and R_s for the report content formed under sharing arrangement s, the comparison is

$$\hat{y}_s = F(E, A, R_s). \tag{1}$$

Here F is the final-model call and $\hat{y}_s$ its selected answer. The sharing arrangement changes how R_s is produced while E and A stay fixed. An *unmodified report* is the output of its chosen update path before an experimental change to its content or presentation.

Comparison	What changes	What the contrast asks
Formation (§3)	Source access during report updates	Does earlier sharing matter with complete final sources and fixed initial recommendations?
Content (§4)	Explanation origin and copied-text origin, independently	Do natural explanations carry a benefit under identical source copies?
Revision (§5.1)	A note beside an unchanged report	Can a correct applicability judgment change the use of retained content?
Role (§5.2)	Role and position of two fixed interpretation texts	Does the assigned current role matter beyond content availability and order?

Figure 1: Interventions tracing report formation to final evidence use. Complete original sources and initial recommendations are fixed at the final decision.

These interventions address successive parts of the same workflow. Their effects are estimated in separate comparisons; they are not additive components of a numerical decomposition. Fixing the explicit recommendation fields does not fix answer-bearing information expressed within explanations, which is part of the report content under study.

Task sets and evaluation. A synthetic scenario has two task variants that differ in one observation. The rule, provider names, throughputs, and answer order remain fixed. The changed observation makes the restriction apply in one variant and cease to apply in the other. The initial set contains 32 scenarios and 64 variants from eight rule families. The expanded set contains 128 scenarios from sixteen additional families. Seven unavailable reports from the agent-update paths leave 249 variants available under all three sharing arrangements; every scenario remains represented. The content, revision, and role experiments reuse these 249 variants.

The recorded controlled experiments use the DeepSeek Flash model family with temperature zero and explicit thinking disabled. Agents have a 1,400-token output limit and final decisions a 220-token limit. The original sharing comparison, revision study, and role study have two final responses per variant and condition. The explanation and copied-text experiment has one. We score the returned provider choice against the reference, average responses within variants, average variants within scenarios, and give scenarios equal weight. This *scenario-averaged accuracy* is the reported controlled-task metric. Confidence intervals resample scenarios and preserve paired conditions. Differences are computed before rounding. The original common-variant analysis is post hoc and descriptive. Later studies retain their *primary comparisons*, the main tests specified before their final-model evaluations. Appendix B specifies scoring, uncertainty, and batch identities.

3 Source Sharing Retains Value With Complete Final Sources

The first experiment tests whether complete final sources make sharing during report formation redundant. We vary source access during the two updates while fixing final originals and initial recommendations (Section 2). Table 1 reports paired accuracy differences.

Complete final sources do not reproduce the benefit of letting agents use those sources while preparing their reports. Second-update sharing improves expanded-set accuracy by 18.36 points, with a 95% interval from 13.87 to 23.05. Among 498 paired responses, 99 improve and seven worsen; the initial-set gain is 13.28 points. Final originals, explicit recommendations, and the number of report updates are identical, so their differences cannot account for the improvement.

This finding concerns access during report formation rather than a demonstrated advantage of sharing at every update. Sharing during both updates improves expanded-set accuracy by 19.34 points relative to neither update. Its difference from second-update sharing is 0.98 points, with an interval spanning zero. The relative merits of the two sharing schedules remain unresolved.

Table 1: Source sharing during report updates with complete final sources and fixed initial recommendations. Values are differences in scenario-averaged accuracy (percentage points), with descriptive 95% intervals.

Source-sharing comparison	Initial set	Expanded set
Second update versus neither update	13.28	18.36
95% interval	6.25 to 21.09	13.87 to 23.05
Both updates versus neither update	15.63	19.34
95% interval	8.59 to 23.44	14.65 to 24.22
Both updates versus second update	2.34	0.98
95% interval	−2.34 to 7.81	−2.15 to 4.10

The changed source access reaches the final decision through the resulting report content. Agents may express different relationships, and their source selections may cause different quotations to be copied. The next experiment distinguishes generated explanations from these deterministic source copies.

4 Explanation Records Carry A Benefit Under Fixed Source Copies

Question and intervention. We ask which part of the generated report changes final decisions when all originals are already available. For every variant, we take the reports from sharing during the second update and from neither update. We separate each report into its explanation records and copied source text. The latter includes whole-source copies and quotations from the two passages cited by each explanation record. Every occurrence and its order are retained. The explanation records retain their generated wording, relation labels, source references, subjects, and properties.

We independently choose the explanation records and copied text from the two report origins, producing four conditions. For example, one condition combines explanations formed after sharing with copies from reports formed without sharing. Another changes only those explanations and keeps exactly the same copied text. The task, original sources, initial recommendations, and rendering rule remain fixed across all four conditions. Both sets of explanations are existing model outputs. The intervention performs no new report generation or answer-informed rewriting.

Table 2: Explanation records and copied source text exchanged independently. Sharing occurs during the second update. Values are scenario-averaged accuracy (%) and explanation-origin differences (percentage points), with 95% intervals.

	Origin of copied source text	
Origin of explanation records	Without sharing	With sharing
Without sharing	76.95	80.47
With sharing	91.80	92.58
Explanation-origin difference	14.84	12.11
95% interval	10.16 to 19.92	7.81 to 16.41

Results. Different source-copy selection cannot by itself explain the benefit of explanations formed after sharing. These explanations improve accuracy under either fixed copied-text background (Table 2). The sole primary contrast averages those two improvements: 13.48 points, with a 95% interval from 9.38 to 17.77. Within each contrast, the originals, initial recommendations, and copied passages are identical. Useful variation therefore remains in the generated explanation collection even after the deterministic copies are fixed. The two improvements are 14.84 points with no-sharing copies and 12.11 with sharing copies; Appendix C reports the variant-level transitions.

Copies also affect decisions, but fixing them does not eliminate the explanation effect. Changing copied-text origin gives an average gain of 2.15 points, with an interval from 0.39 to 3.91. The interaction measures how much the explanation effect changes between copied-text backgrounds. Its estimate is −2.73 points, with an interval from −5.86 to 0.39, leaving that difference uncertain. The

manipulated unit is the whole explanation collection: wording, source links, record count, length, and restated facts change together. The benefit thus belongs to this collection, without isolating a particular semantic feature.

An observable change in the report. Figure 2 illustrates a task in which the permit is valid and names the current job. Without source sharing, the reports discuss throughput, the default prohibition, and the exemption rule separately. With source sharing in the second update, the agent holding the observations connects the matching permit and job to the exemption. In this task, replacing the entire collection of explanation records changes the final answer from the slower provider to the eligible fastest provider under both fixed copied-text backgrounds. The example was selected for its clear paired changes. The quantitative estimates use all 249 variants.

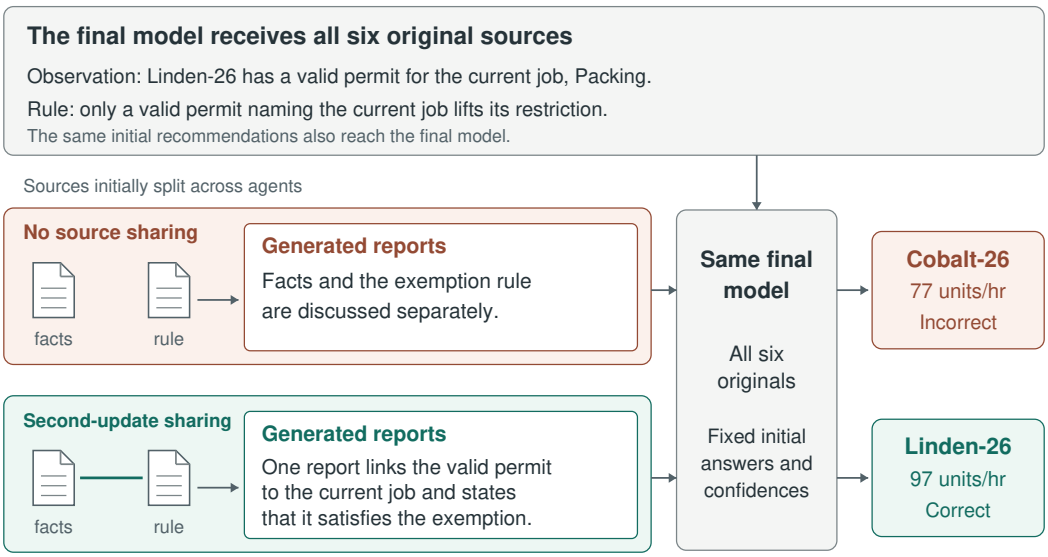


Figure 2: Selected permit example (Appendix A.4). Exchanging the entire explanation collection changes the final choice under either fixed copied-text background, with complete original sources and initial recommendations unchanged.

The example makes one change visible: a generated record connects observations to a rule that the no-sharing reports discuss separately. All 249 sharing-origin collections contain a direct observation-to-rule source link; all 249 no-sharing collections lack one. This is an explicit citation pair, not an independently verified correct interpretation. Its presence documents a structural change accompanying the explanation benefit; the intervention does not separate links from wording or establish each link's semantic correctness.

Connection to the unmodified reports. Separating the components changes their presentation. We therefore evaluate the final-model inputs containing each path's reports in their original, unsplit presentation in the same batch. The difference between these original inputs is 16.02 points. The difference between their matching reconstructed inputs is 15.63 points. Three prespecified checks compare each reconstructed condition with its original input and compare the two differences. Their 90% intervals all lie inside the prespecified tolerance of five percentage points in either direction. Appendix C gives every cell and check. These comparisons connect the content intervention to the original workflow at the aggregate level. The original two-response estimate of 18.36 points and this batch's one-response estimate of 16.02 points retain their separate identities.

This locates useful variation in natural reports, but replacing a whole collection does not show how a final model handles a new interpretation alongside an older report. We next preserve that report and supply an explicit applicability judgment. This moves from identifying useful generated content to testing whether the use of retained content can be revised.

5 Revising And Using Judgments In An Existing Report

5.1 An explicit revision can change the influence of an intact report

The next test concerns an intact report: every earlier statement stays visible while a note is appended. We distinguish stating whether the restriction applies, called a *resolved* judgment, from leaving that question open, called *unresolved*. If an explicit applicability judgment can compensate for less useful report content, supplying it should help reports formed without sharing and reduce the *sharing gap*, the accuracy difference between reports formed after second-update sharing and those formed without it.

We compare four conditions. The unmodified report adds no note. A preservation note states that its meaning is retained and adds no applicability judgment. A resolved revision marks earlier current-status statements as superseded and states whether the restriction applies. An unresolved revision also supersedes those statements but leaves applicability open. The three notes have the same position and word count within a report. All originals, initial recommendations, and earlier report text remain unchanged. We construct the resolved judgment from the task's rule and observation reference; with throughput, it determines the correct provider. This is an *oracle diagnostic*: it supplies a correct interpretation to test its effect, rather than evaluating an automatic method for finding it.

Table 3: Final revisions with the earlier report preserved. Upper panel: scenario-averaged accuracy (%). Lower panel: all three prespecified primary contrasts (percentage points), with Bonferroni-adjusted 98.33% intervals.

	Source sharing during updates	
Added note	Neither update	Second update
No note: unmodified report	74.61	91.41
Preservation note	70.12	91.99
Resolved revision	97.46	97.66
Unresolved revision	69.34	91.02
Primary comparison	**Difference**	**98.33% interval**
Resolved revision versus preservation, no sharing	27.34	21.68 to 33.40
Reduction in sharing gap, preservation to resolved	21.68	15.63 to 27.93
Preservation versus unresolved, second-update sharing	0.98	−1.37 to 3.32

A supplied correct judgment can compensate for less useful reports while their earlier text remains visible (Table 3). Without sharing, resolved revision raises accuracy from 74.61% to 97.46%. The secondary gain over the unmodified report is 22.85 points, with a 95% interval from 18.36 to 27.54. The primary gain over preservation is 27.34 points. Both comparisons matter because preservation itself lowers accuracy by 4.49 points relative to no note, and its semantic neutrality has not been independently established. Improvement over the unmodified report shows that the gain is not solely recovery from preservation's adverse effect.

The sharing gap contracts through improved decisions, rather than by making both report paths similarly inaccurate. It falls from 21.88 points under preservation to 0.20 under resolved revision, a prespecified reduction of 21.68 points. Both resolved conditions have high accuracy (Table 3), with the larger improvement occurring without sharing. Supplying the correct judgment can therefore compensate for this report difference without first deleting the earlier text.

This improvement does not establish the predicted reverse effect of leaving a judgment unresolved. After second-update sharing, the preservation-versus-unresolved interval spans zero. Joint support requires positive lower confidence bounds for all three primary comparisons; two are supported, but the third crosses zero, so the joint criterion fails. Earlier explanations and complete sources remain available, potentially allowing recovery of a judgment, but that process was not measured. The resolved revision also combines correct content with an instruction that it supersedes prior judgments. We next keep interpretation content fixed to distinguish its availability from the role assigned to it.

5.2 THE SAME INTERPRETATION IS USED DIFFERENTLY UNDER DIFFERENT REPORT ROLES

Every condition in this experiment contains the unmodified report formed without source sharing and two added interpretation statements. One states the resolved current applicability judgment; the other leaves that judgment unresolved. Both contain the same observation and rule quotations. Each statement appears exactly once. Input labels designate one statement as the report's current judgment and the other as an external annotation that has not been incorporated into the report.

We independently vary which role receives the resolved statement and whether it appears first or second in the added list. The unresolved statement occupies the other role and position. The unmodified report, complete sources, initial recommendations, and both statement texts remain fixed. Table 4 gives the resulting four conditions.

Table 4: Role and position assignments beside reports formed without sharing; both interpretation texts remain fixed. Columns locate the resolved interpretation; external annotations are labeled unincorporated. Values are scenario-averaged accuracy (%) and role differences (percentage points), with 95% intervals.

Role assigned to the resolved interpretation	First	Second
Report's current judgment	89.84	96.88
External annotation	77.73	90.63
Role-assignment difference	12.11	6.25
95% interval	8.98 to 15.43	3.71 to 9.18

Having the correct interpretation present does not fix its influence on the final choice. Designating it as the report's current judgment improves accuracy at either fixed position (Table 4). The sole primary contrast averages these role differences: 9.18 points, with a 95% interval from 7.13 to 11.43. Each comparison contains the same two statements in the same order. The difference therefore concerns their role assignment, including the external statement's non-incorporation cue, rather than adding correct content or moving it later.

Role and position do not contribute independent, additive gains on this accuracy scale. Placing the resolved interpretation second has an average benefit of 9.96 points, similar in magnitude to the average role effect. The role difference is smaller when that interpretation appears second: the interaction is -5.86 points, with an interval from -9.77 to -1.76. Later placement and current-role assignment thus partly substitute in the observed choices, while each retains a positive effect in these cells. This behavioral pattern does not establish a shared internal mechanism.

Here, interpretation use is measured through the final provider choice, not an internal belief state.

6 EVIDENCE ACROSS TASKS AND MODELS

The controlled studies isolate report content under fixed final sources and initial recommendations. The benchmark studies examine exchange of supporting material, meaning agents' cited facts and reasoning, under three distinct final-input controls: complete original sources in the first HiddenBench comparison, presence or absence of the agents' fixed initial support summaries in M3GQA, and revised reports in the three-model HiddenBench comparison.

HiddenBench with complete final sources. HiddenBench tests collective decisions with relevant information distributed among agents (Li et al., 2026). The same 65 questions are evaluated after agents exchange either supporting material or only answers and confidence values. A separate final model receives the resulting reports and all originals in both conditions. Earlier exchange still helps when final source availability is fixed: supporting-material exchange yields 34 correct answers, compared with 25 after answer-only exchange. The paired difference is 13.85 points, with a 95% interval from 4.62 to 23.08; ten questions improve and one worsens. This extends the formation-path finding to benchmark questions, but updated recommendations can differ, so the effect cannot be assigned specifically to explanations. Appendix D gives all six combinations of exchange and final-source access.

M3GQA with identical initial support summaries. M3GQA asks for sets of entities using information in a graph (Peng et al., 2025). We evaluate the same 800 questions in four conditions. Supporting-material exchange during the agent update is crossed with supplying the agents' fixed initial support summaries to the final model. Updated answers and confidences always reach that model; the added summaries contain initial supporting statements and references. Answer F1 measures overlap between predicted and reference entity sets, averaged using fixed weights for four question types. Final support partly substitutes for earlier exchange: the exchange benefit falls from 6.33 points without the summaries to 2.19 with them. The residual interval is 0.77 to 3.63, and the interaction is -4.14 points. Thus access to the same initial support reduces, but does not remove, the exchange benefit. Updated answers can still carry this residual; the summaries are selected support, not complete graph sources. Appendix D gives the historical weights and a separate 192-question cohort, disjoint from these 800 questions, with a residual effect of 4.10 F1 points.

Supporting-material exchange across three models. A separate HiddenBench study uses four agents followed by a final language-model synthesis. Peer answers and confidence values remain visible in both conditions; the treatment adds their supporting material. We compare distributed sources, where agents initially hold different information, with redundant sources, where each agent has all information. Table 5 reports the strict paired results. Gains under distributed sources occur in all three model configurations. Two GPT questions with inconsistent initial drafts are excluded, leaving a strict paired set of 63.

Table 5: HiddenBench accuracy changes (percentage points) from sharing supporting material, relative to peer answers and confidences alone. Distributed and redundant denote the agents' initial source allocation.

Model configuration	Paired questions	Distributed	Redundant
GPT-5.5	63	36.51	-1.59
DeepSeek-V4-Flash	65	20.00	6.15
GLM-5.3-Flash	65	43.08	0.00

Across all three configurations, exchange helps more when agents initially hold different sources than when each already holds all sources. This is consistent with exchange helping agents use complementary information, without isolating how that information changes their reports. Larger gains also do not imply comparable final accuracy. Historical all-question counts in Table 19, separate from the strict paired estimates, show lower absolute accuracy under distributed sources after exchange in every configuration. These benchmarks extend the relevance of the formation question under their respective controls. Only the controlled content experiment exchanges explanation origin with final originals, initial recommendations, and source copies fixed.

7 Related Work

Communication and information history. Multi-agent debate studies how communication improves reasoning (Du et al., 2024); HiddenBench studies disclosure and integration failures under distributed information (Li et al., 2026); and Babu et al. (2026) calibrate communication and speaker selection. Arghal et al. (2026) analyze how information mediation affects sequential social learning. In single-model conversations, Laban et al. (2026) distinguish information-preserving instruction changes from multi-turn degradation and find that later restatement does not fully restore performance. Our comparison instead fixes the complete sources and initial recommendations supplied to a separate final model, then exchanges naturally generated explanations under fixed source copies to locate a consequence of earlier sharing.

Revision, role, and position. Sadhu et al. (2026) study belief revision with formally verified ground truth at each turn. Liu et al. (2024) examine where information appears in context; Liao (2026) manipulates provenance and authority markers while controlling proposition and position. Our revision diagnostic supplies a correct applicability judgment beside an intact report. The role comparison then holds both interpretation texts present while changing which is designated current within the report, at each position. These distinguish the availability of a supplied interpretation from its use under explicit report labels.

Causal explanations of model behavior. Mechanistic studies connect distinct intervention predictions to visual binding stages (Assouel et al., 2026) or use reasoning-graph diagnostics to guide exploratory feature interventions (Zhao et al., 2026). Grant et al. (2026) show why successful interventions can still be unfaithful to natural computation. We study observable reports and final choices. Using natural explanation records and checking reconstructed against original inputs connects our content intervention to the recorded workflow; it does not identify a unique semantic mediator or an internal circuit.

8 Discussion And Conclusion

What explains the formation effect? Complete final access and unchanged initial recommendations leave a sharing benefit, so differences in those inputs cannot explain it. The explanation exchange further excludes deterministic source-copy selection as the sole carrier: natural records formed after sharing help under either copied-text background. Sharing creates direct observation-to-rule links in these records, but links, wording, length, and restated facts vary together. The identified causal unit is therefore the explanation collection. A specific semantic feature's contribution remains unseparated.

What do revision and role add? The revision diagnostic shows that a correct applicability judgment can compensate for less useful reports while their history remains visible. Its uncertain reverse effect prevents a claim that this judgment is necessary for the sharing benefit. The role experiment addresses a different ambiguity: every condition already contains the same correct interpretation, yet its influence changes with its designated role at either position. Together these results explain how formation and later presentation affect evidence use, without treating oracle repair as proof of the natural mechanism.

Implications for report-based systems. An inventory of available documents misses whether agents have related them in reports and how later judgments are presented. These are distinct objects of evaluation in this fixed-budget workflow; greater final-model computation was not tested. Improving reports by sharing also does not establish that reports are preferable to a sources-only workflow. Those comparisons have opposite directions in the synthetic and historical HiddenBench batches (Appendix D.5). The formation results concern systems that use reports; overall report utility requires its own matched workflow comparison.

Conclusion. Complete final sources leave report formation consequential. The controlled comparisons connect source sharing to generated explanations, then show how explicit revision and the role assigned to an interpretation affect the final decision.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting and language editing. They also assisted with code implementation, and analysis and interpretation of experimental outputs. The authors reviewed and edited the AI-assisted text and take responsibility for the final manuscript, claims, and reported results.

REPRODUCIBILITY STATEMENT

Appendices A and B specify the source allocation, task construction, report interventions, scoring units, and paired uncertainty estimates. Appendices C and D give the complete controlled and benchmark comparisons; Appendix E describes execution accounting, retained records, and verification scope. The retained experimental evidence archive is separate from the manuscript's LaTeX package. It contains raw natural-content responses, component provenance, frozen task selections, analysis code, and pairing audits. Its offline checks cover source hashes, joins, rescoring of retained outputs, paired estimates, and stored confidence intervals. The manuscript source package contains editable numerical tables and formatting assets. Each experiment retains its original batch identity and primary comparisons.

ETHICS STATEMENT

This study examines language-model behavior on synthetic decision scenarios and existing benchmarks. Its findings concern the evaluated reporting workflows and do not establish the reliability of autonomous decisions in consequential applications. The distinction between model-assisted judgments and reference-derived task labels is retained in the reporting. Sharing and reuse of benchmark materials and model outputs should respect their applicable licenses and provider terms.

REFERENCES

Raghu Arghal, Kevin He, Shirin Saeedi Bidokhti, and Saswati Sarkar. Steering the Herd: A Framework for LLM-based Control of Social Learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3a807a9e00552b5883e7417ff97fa4e8-Abstract-Conference.html.

Rim Assouel, Declan Campbell, Yoshua Bengio, and Taylor Webb. Visual symbolic mechanisms: Emergent symbol processing in Vision Language Models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ffffd32092164225bb302d438cacbc61-Abstract-Conference.html.

Abhijith Babu, Ramneet Kaur, Vishal Pramanik, Olivera Kotevska, Nathaniel D. Bastian, Susmit Jha, Sunny Raj, Yanzhao Wu, Sumit Kumar Jha, and Anirban Roy. Consilience: Conformally calibrated communication control for hidden-profile multi-agent reasoning. *arXiv preprint arXiv:2608.20564*, 2026. URL https://arxiv.org/abs/2608.20564v1.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/133e588e1429f9f1e25b215da145580e-Abstract-Conference.html.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs Get Lost In Multi-Turn Conversation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/59f6421e64707225fdf5b28840679a07-Abstract-Conference.html.

Yuxuan Li, Aoi Naito, and Hirokazu Shirado. Systematic failures in collective reasoning under distributed information in multi-agent LLMs. *arXiv preprint arXiv:2505.11556*, 2026. URL https://arxiv.org/abs/2505.11556v4. Version 4, revised May 13, 2026; first posted May 15, 2025. Accepted to ICML 2026.

Junchi Liao. Auditing provenance sensitivity in LLM agent action selection. *arXiv preprint arXiv:2607.20827*, 2026. URL https://arxiv.org/abs/2607.20827v1.

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/.

Boci Peng, Yongchao Liu, Xiaohe Bo, Jiaxin Guo, Yun Zhu, Xuanbo Fan, Chuntao Hong, and Yan Zhang. M^{3}GQA: A multi-entity multi-hop multi-setting graph question answering benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30594–30620, 2025. doi: 10.18653/v1/2025.acl-long.1478. URL https://aclanthology.org/2025.acl-long.1478/.

Jayanta Sadhu, Sayem Shahad, and Kenneth Marino. DeReLab: Probing defeasible reasoning and confirmation bias in LLMs with a generative benchmark. *arXiv preprint arXiv:2608.30413*, 2026. URL https://arxiv.org/abs/2608.30413v1. Accepted to EMNLP 2026, as stated in the arXiv record.

Zheng Zhao, Yeskendir Koishekenov, Xianjun Yang, Naila Murray, and Nicola Cancedda. Verifying Chain-of-Thought Reasoning via Its Computational Graph. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c7b5a35ea98b62512a869c19ea7b03cb-Abstract-Conference.html.

A TASKS, SOURCE ALLOCATION, AND REPORT CONSTRUCTION

This appendix supplies the information needed to reproduce and interpret the paper's experiments. Appendices A and B define inputs, tasks, scoring, and comparisons. Appendix C follows the main experimental sequence: source sharing, explanation exchange, intact-report revision, and role assignment. Appendix D gives the benchmark protocols and complete supporting comparisons; Appendix E identifies the retained execution and verification records.

A.1 WHAT EACH AGENT RECEIVES

The controlled tasks use three candidate providers with distinct verified throughputs. Two have no restriction for the current job. The fastest provider has an explicitly stated restriction and an exemption or triggering condition. A deterministic evaluation reference applies the rule to the current observations and selects the fastest eligible provider. Generation of task text and evaluation of the reference are separate operations.

Table 6 gives the full initial source allocation. "Private" means initially available to one agent. All four agents receive the two shared sources. Agent numbering in this table serves only to describe the experimental allocation. Source sharing follows the cycle in its last column; the observation-holding agent receives the rule. The source scope is one additional private document per agent. Other agents' full reports are absent from the update messages. Their preceding answers and confidence values remain present in every sharing condition.

Table 6: Source allocation in the controlled synthetic workflow. Every final-model input receives all six original sources.

Recipient	Initial private source	Additional source when sharing is enabled
Agent 1	Current registry and job observations	Agent 4's restriction rule
Agent 2	Use verified throughput; infer no unstated restriction	Agent 1's observations
Agent 3	Use current registry entries and job observations	Agent 2's instruction
Agent 4	Restriction and its applicability rule	Agent 3's instruction

Shared by all agents: the three verified throughputs, and the statement that the other two providers are unrestricted and only the explicit restriction applies.

Each update uses the same preceding-round snapshot for all agents. An agent retains its own previous response and original visible sources. Its additional source is available in both rounds, the second round, or neither round according to the chosen arrangement. The two paths without first-round sharing reuse the identical first update. This reuse makes their second-update comparison local to the additional source access at that step.

A.2 REPORT INSTRUCTIONS AND SOURCE COPIES

The agent produces an answer matching one candidate, a confidence value between zero and one, references to at most twenty visible sources or source statements, and at most four explanation records. The six permitted relationships describe support for the same claim, contradiction of the same claim, limits on applicability, updates to status, independent constraints, and unresolved relationships. The instruction asks for a short explanation identifying the subject and the relevant condition or time. It recommends fewer than twenty-four words; the implementation applies a sixty-word cap.

Source statements are mechanically addressed pieces of source text. Agents must consult the complete visible source to preserve negation and context. Selecting a statement selects its full parent source. Selected source references are deduplicated for source selection. The report construction also attaches verbatim quotations to relation endpoints (the two cited passages), so the rendered report can contain repeated text. A relation explanation contains an interpretation of source statements. It is instructed to exclude final recommendations, vote counts, peer endorsements, and confidence values.

The final model receives the question, candidate options, complete source collection, fixed initial recommendations and confidences, a summary of those recommendations, and the selected report content. Its instruction explicitly calls the reports fallible interpretations and asks it to choose using all sources. The recorded response format contains the chosen answer and a confidence value. The exact machine-formatted requests and transformations are retained in the experimental archive; this appendix describes their implemented meaning.

A.3 Matched scenarios and rule families

A scenario fixes provider names, throughputs, the rule, answer order, and all current observations except one registered field. Its two variants change that field so the restriction applies in one variant and does not apply in the other. The initial set has four scenarios in each of eight families. The expanded set has eight scenarios in each of sixteen additional families. Table 7 lists every family. The tasks are generated; the reports are natural model outputs on those tasks.

Table 7: All controlled task families. A variant pair changes only the listed observation.

Set	Rule family	Changed observation
Initial	Class mapped to a division	Class-to-division mapping
Initial	Delegated jurisdiction	Office jurisdiction
Initial	Half-open interval endpoint	Current minute
Initial	Issued permit and revocation	Revocation status
Initial	Issuer authorization	Issuer's authorization
Initial	Offset half-open time window	Local hour
Initial	Threshold and region	Region
Initial	Conjunction of two hazards	Second hazard
Expanded	Overlap of authority domains	Authority domains
Expanded	Conjunction with a negated shield	Shield status
Expanded	Disjoint time windows	Current hour
Expanded	Emergency cancellation of an exemption	Emergency override
Expanded	Exactly one hazard triggers a restriction	Second hazard
Expanded	Interval overlap	Job end
Expanded	Joint waiver exemption	Second waiver
Expanded	Mapped scope and destination intersection	Current destination
Expanded	Mapped scope with exclusions	Excluded document types
Expanded	Nested disjunction and region	Second hazard
Expanded	Overnight time window	Local hour
Expanded	Permit scope for the current job	Job named on permit
Expanded	Reinstatement overrides revocation	Signed reinstatement
Expanded	Relative expiry of a notice	Current day
Expanded	Three-link asset scope	Site zone
Expanded	Two of three hazards	Third hazard

The expanded set originally has 256 variants. Seven stage-specific upstream states are missing: three for sharing during both updates, two for sharing during the second update, and two for neither update. Removing the seven affected variants from the common-stage comparison leaves 249 variants and all 128 scenarios. A missing state is treated as unavailable. It is excluded from the paired comparison without assigning an incorrect answer. Later core interventions use this same intersection.

A.4 The retained permit example

The example in Figure 2 concerns a current Packing job. Linden-26 has a valid permit naming Packing, with verified throughput of 97 units per hour. Cobalt-26 and Aster-26 have throughputs of 77 and 62 and no restriction. The rule prohibits Linden-26 by default and grants its sole exemption to a valid permit naming the current job. The correct answer is Linden-26.

The reports formed without sharing contain eleven explanation records. The records discuss the fastest provider, the default prohibition, and the exemption rule separately. After source sharing in the second update, the reports contain fourteen records. The agent holding the observations

states that the valid permit names the current job and satisfies the exemption. Other records remain present, including the rule-holding agent's statement about the default prohibition. Thus the change is a collection of reports with a newly expressed connection between the observations and the rule.

Table 8 gives all seven final answers from the content experiment. Both fixed copied-text comparisons change from an incorrect answer to the correct answer when the explanation origin changes. The unmodified report inputs have the same direction. These observed answers support the use of this task as an illustration of the complete-record intervention. The illustrated effect applies to the complete exchanged collection.

Table 8: All answers for the permit example in the content-exchange batch. Every condition has one final response.

Explanation origin or reference input	Copied-text origin	Final answer
No sharing	No sharing	Cobalt-26
No sharing	Second-update sharing	Cobalt-26
Second-update sharing	No sharing	Linden-26
Second-update sharing	Second-update sharing	Linden-26
Unmodified report without sharing	Original presentation	Cobalt-26
Unmodified report after sharing	Original presentation	Linden-26
Original sources alone	No report	Cobalt-26

This illustration was selected during manuscript preparation. We first identified variants with an incorrect original no-sharing answer, a correct original sharing answer, and an improvement under both fixed copied-text backgrounds. Twenty-seven variants meet these criteria. We selected the first in ascending task-identifier order. The retained experimental archive records the selected task and complete component records. This outcome-based selection affects only the illustration. The estimates in Section 4 use every retained variant.

B Scoring, Statistical Units, And Experimental Batches

B.1 Scenario-averaged accuracy

Let $Y_{bvr,c}$ indicate a correct answer for scenario b, variant v, response r, and condition c. Let B be the number of scenarios, V_b the retained variants of scenario b, and $m_{v,c}$ the number of final responses to variant v in condition c. The controlled-task accuracy is

$$a_c = \frac{100}{B} \sum_{b=1}^{B} \frac{1}{|V_b|} \sum_{v \in V_b} \frac{1}{m_{v,c}} \sum_{r=1}^{m_{v,c}} Y_{bvr,c}. \tag{2}$$

Paired differences use the same available variants and responses within each comparison. A scenario has one or two retained variants; each condition has one or two final responses per variant. Repeated responses and matched variants remain nested within scenarios, which are the resampling units. The scenario-bootstrap intervals describe variation among the retained scenarios within these generated rule families; multiple scenarios share each rule template.

Pooled accuracy divides correct responses by all observed responses. In the expanded set, incomplete variant pairs make pooled accuracy differ from scenario-averaged accuracy. For example, the reconstructed sharing-explanation and sharing-copy condition has 230 correct responses out of 249, giving pooled accuracy 92.37% and scenario-averaged accuracy 92.58%. The resolved final revision without sharing has 485 out of 498 correct, giving 97.39% pooled and 97.46% scenario-averaged accuracy. Both are retained under their respective labels.

The earlier controlled parser maps the returned option to the candidate list using its recorded normalization of case and whitespace. The natural-content experiment requires a valid structured response whose answer exactly matches one candidate. All 1,743 responses satisfy that check. The offline verifier rescored the stored core answers using the recorded parsing rules and separately reparsed all natural-content raw responses. Invalid model outputs are scored according to their study's frozen rule; requests with no obtained model output or missing report inputs retain an unavailable status.

B.2 PRIMARY COMPARISONS AND UNCERTAINTY

The original common-variant analysis uses 20,000 percentile bootstrap draws over scenarios with seed 20260920. It was conducted after the stage results were observed and is reported as descriptive. The initial and expanded template sets remain separate.

The explanation and copied-text experiment has one primary contrast: the explanation-origin difference averaged over the two copied-text origins. Its secondary comparisons are the explanation effect within each copied-text origin, the average copied-text effect, and their interaction. The unmodified-input references and reconstruction checks are additional comparisons. Bootstrap analysis uses 20,000 draws with contrast-specific deterministic seeds derived from base seed 20260922. The three reconstruction checks use 90% intervals and a prespecified tolerance of five percentage points in either direction. Each checks aggregate accuracy, with the same variants and scenario weighting.

The final-revision study has three primary predictions: a positive resolved-versus-preservation effect without sharing, a positive reduction of the sharing advantage from preservation to resolved revision, and a positive preservation-versus-unresolved effect after second-update sharing. Each has a Bonferroni-adjusted 98.33% interval. Joint support requires all three lower bounds to be positive; the third interval spans zero. The role-and-position study has one primary average-role contrast. Its within-position effects, average position effect, and interaction are secondary. These studies use 20,000 scenario-bootstrap draws and deterministic contrast-specific seeds derived from base seed 20260921. The archive preserves the exact derivation used for each comparison and interval level.

M3GQA uses paired task-level answer F1 and stratified bootstrap with 10,000 draws. The 800-question study uses its frozen historical stratum weights. The 192-question study averages its four equal-sized strata equally. The general effect routine uses seed 914800. In the 192-question study, separate within-stratum estimates use seed 915192 and the multihop-versus-single comparison uses seed 915193. The implementations and frozen task memberships are included in the archive.

B.3 KEEPING REPEATED EVALUATIONS SEPARATE

The original expanded-set second-update-sharing difference is 18.36 points. The unmodified-report references yield 16.80 points in the final-revision batch and 16.02 in the natural-content batch. Each value uses its own responses on the shared task population. Later readings of a report do not replace its original stage score. Variation between model calls can remain at temperature zero.

The controlled experiments use recorded DeepSeek Flash service configurations. Earlier requests name the V4 Flash configuration; the natural-content run uses the Flash service alias. The later benchmark documentation also names the V4.1 Flash configuration. Each reported effect compares conditions within its recorded batch. The provenance ledger retains the exact request aliases, returned service metadata, and corresponding response sets for each batch.

B.4 WHAT THE CONTROLS IDENTIFY

The stage experiment changes the report-producing process with final sources and initial recommendations fixed. The content experiment exchanges existing explanation collections with explicit source copies controlled. The revision experiment supplies a source-derived applicability judgment with an explicit supersession statement. The role experiment holds two complete interpretations fixed and reassigns their roles at fixed relative positions. Each estimate is the within-batch effect of its specified intervention on the observable reporting workflow. The experiments connect formation, content, revision, and role through their controlled contrasts, rather than through an additive decomposition of the original sharing effect.

C Complete Results For The Main Experiments

C.1 Original source-sharing results

The expanded-set paired second-update versus neither-update comparison contains 99 incorrect-to-correct responses, seven correct-to-incorrect responses, 362 jointly correct responses, and thirty jointly incorrect responses. Inference averages these observations within their 128 scenarios.

Table 9: Original final-response counts on the common-variant sets. The percentage column is pooled accuracy. The paired scenario-averaged differences are in Table 1.

Task set	Source sharing during updates	Correct responses	Pooled accuracy
Initial	Both updates	125/128	97.66
Initial	Second update	122/128	95.31
Initial	Neither update	105/128	82.03
Expanded	Both updates	466/498	93.57
Expanded	Second update	461/498	92.57
Expanded	Neither update	369/498	74.10

C.2 All explanation and copied-text conditions

The deterministic separation traverses the rendered report collection in its existing order. Whole-source quotation records and both quotations attached to each relationship are collected as copied source text, preserving every occurrence. The remaining relationship records retain their type, source endpoints, subject, property, explanation, order, and multiplicity. The reconstructed input presents copied text followed by explanation records under a common rendering rule. Report-origin condition names are absent from the model input. Each explanation collection is reused unchanged under both copied-text backgrounds, and each copied-text collection is reused unchanged under both explanation origins.

The original source references continue to address the same task's complete original-source collection. Facts restated by the model inside an explanation remain in that explanation. The intervention fixes deterministic source copies and retains each natural explanation collection's wording, length, and model-restated facts.

Table 10: All seven conditions in the natural-content batch. There are 128 scenarios, 249 variants, and one final response per condition and variant.

Explanations or reference	Copied-text origin	Correct	Scenario accuracy
No sharing	No sharing	191/249	76.95
No sharing	Second-update sharing	200/249	80.47
Second-update sharing	No sharing	228/249	91.80
Second-update sharing	Second-update sharing	230/249	92.58
Original no-sharing report	Original presentation	187/249	75.39
Original sharing report	Original presentation	227/249	91.41
Sources alone	No report	149/249	60.16

The sources-alone reference removes initial recommendations as well as the reports. It also removes the upstream report-generation computation. Its differences describe complete workflow utility on this set. The four component conditions are the comparisons with initial recommendations held fixed.

The fixed-copy comparisons have 34 improvements, four harms, and 211 unchanged variants under sharing-origin copies; they have forty improvements, three harms, and 206 unchanged variants under no-sharing-origin copies. The same-batch unmodified-input comparison has 45 improvements, five harms, and 199 unchanged variants. These counts describe individual variants; the uncertainty estimates retain scenarios as the resampling units.

Table 11: All natural-content contrasts. Effects are percentage-point differences. The first row is the sole primary contrast.

Comparison	Effect	95% interval
Average explanation-origin effect	13.48	9.38 to 17.77
Explanation effect with sharing-origin copies	12.11	7.81 to 16.41
Explanation effect with no-sharing-origin copies	14.84	10.16 to 19.92
Average copied-text-origin effect	2.15	0.39 to 3.91
Explanation and copied-text interaction	−2.73	−5.86 to 0.39
Matching reconstructed sharing difference	15.63	10.94 to 20.70
Unmodified-input sharing difference in this batch	16.02	11.33 to 20.70
Original sharing report versus sources alone	31.25	26.56 to 35.55
Original no-sharing report versus sources alone	15.23	10.16 to 20.31

Table 12: All prespecified reconstruction checks, in percentage points. Each 90% interval lies within the tolerance from -5 to 5 points.

Comparison	Difference	90% interval
Reconstructed versus original sharing input	1.17	−0.39 to 2.73
Reconstructed versus original no-sharing input	1.56	−1.56 to 4.69
Reconstructed gap versus unmodified-input gap	−0.39	−4.30 to 3.52

C.3 All intact-report revisions

Table 13: All intact-report revision cells. Every condition has two responses for each of 249 variants and 128 scenarios. Accuracy percentages give scenarios equal weight.

Sharing schedule	Report condition	Correct	Accuracy
Both updates	No note: unmodified report	464/498	93.36
Both updates	Preservation note	462/498	92.97
Both updates	Resolved revision	489/498	98.24
Both updates	Unresolved revision	459/498	92.38
Second update	No note: unmodified report	454/498	91.41
Second update	Preservation note	457/498	91.99
Second update	Resolved revision	486/498	97.66
Second update	Unresolved revision	452/498	91.02
Neither update	No note: unmodified report	371/498	74.61
Neither update	Preservation note	349/498	70.12
Neither update	Resolved revision	485/498	97.46
Neither update	Unresolved revision	348/498	69.34

The no-note, preservation, resolved, and unresolved conditions are defined in Section 5. Every original explanation and source copy remains physically present. Resolved and unresolved revisions explicitly supersede previous current-status judgments. Preservation retains their stated meaning. The added notes have matched word count within each report; their character-count spread is below ten percent. Two wording patterns are each assigned to 64 scenarios. Equal word count controls note length at that level; the notes enact their stated different revision operations. For all 747 parent report groups, the construction manifest records `terminal_answers_scores_effects_seen_during_construction=false`.

Table 14: Every intact-report revision contrast. The first three form the prespecified primary family. Other rows retain their secondary status. Effects are percentage points.

Comparison	Effect	95% interval	98.33% interval
Resolved versus preservation, neither update	27.34	22.66 to 32.23	21.68 to 33.40
Sharing-gap reduction from preservation to resolved	21.68	16.60 to 26.95	15.63 to 27.93
Preservation versus unresolved, second update	0.98	−0.98 to 2.93	−1.37 to 3.32
Resolved versus unmodified, neither update	22.85	18.36 to 27.54	17.19 to 28.71
Unresolved versus unmodified, neither update	−5.27	−9.18 to −1.56	−10.16 to −0.78
Preservation versus unmodified, neither update	−4.49	−7.42 to −1.76	−8.20 to −1.17
Resolved versus unmodified, both updates	4.88	2.34 to 7.81	1.95 to 8.40
Unresolved versus unmodified, both updates	−0.98	−2.54 to 0.20	−2.73 to 0.59
Preservation versus unmodified, both updates	−0.39	−1.56 to 0.78	−1.95 to 1.17
Resolved versus unmodified, second update	6.25	3.52 to 9.18	3.13 to 9.96
Unresolved versus unmodified, second update	−0.39	−1.95 to 1.17	−2.34 to 1.37
Preservation versus unmodified, second update	0.59	−0.59 to 1.95	−0.98 to 2.15

Table 15: The complete primary revision family under each wording pattern. Both patterns are evaluated on subsets of the same 128 scenarios.

Wording	Comparison	Effect	98.33% interval
First	Resolved versus preservation, no sharing	27.73	19.53 to 36.33
First	Sharing-gap reduction	22.27	13.67 to 31.25
First	Preservation versus unresolved, sharing	1.95	−1.95 to 6.25
Second	Resolved versus preservation, no sharing	26.95	18.75 to 35.55
Second	Sharing-gap reduction	21.09	12.50 to 29.69
Second	Preservation versus unresolved, sharing	0.00	−2.73 to 1.95

C.4 All role and position comparisons

Both fixed interpretations contain the same supporting quotations, and each appears once. When the resolved interpretation takes the current-judgment role, the unresolved one takes the external role. When the resolved interpretation takes the external role, the unresolved one takes the current-judgment role. The second factor specifies their first or second positions in an ordered list. This controls role at each position by reassigning both interpretations. The common diagnostic instruction in all four inputs states: "The ordered binding-state control is part of this diagnostic input. Role labels and packet order are explicit; both packets are fallible interpretations." Here a packet is an added interpretation statement; the prompt's term 'binding-state' refers to its assigned report role, not an observed internal state. The raw role labels are `integrated_report_current_state` and `external_unintegrated_annotation`. This instruction, the original report, the complete sources, and the initial recommendations do not vary across the four conditions.

Table 16: Full role-and-position results and contrasts. Every condition contains 498 final responses on the same 249 variants. The average role contrast is primary.

Resolved interpretation's assignment	Correct	Scenario accuracy
Current judgment, first	448/498	89.84
Current judgment, second	482/498	96.88
External annotation, first	388/498	77.73
External annotation, second	452/498	90.63
Comparison	**Effect**	**95% interval**
Average role assignment	9.18	7.13 to 11.43
Role assignment at first position	12.11	8.98 to 15.43
Role assignment at second position	6.25	3.71 to 9.18
Average second versus first position	9.96	7.52 to 12.50
Role and position interaction	−5.86	−9.77 to −1.76

The raw labels specify the relationship of each interpretation to the report. The contrast measures how this explicit role assignment changes the final decision when both interpretation texts and their positions remain fixed.

D Benchmark Protocols And Complete External Comparisons

D.1 HiddenBench with matched final-source access

The HiddenBench source-access study uses all 65 questions. Four agents initially answer from their allocated sources. Two update paths retain either supporting facts and rationale in peer messages or only the peer answers and confidence values. The final model sees the resulting reports, whose updated recommendations can differ between paths. Its additional evidence is varied among no additional material, the same frozen initial report support, and all original private sources. Initial report support consists of the initial key facts and rationale. All-original access appends the union of the private originals to the final input.

The resulting six conditions have one final response per question, for 390 responses. The cohort contains eight previously observed small-test questions and 57 subsequently completed questions. Table 17 contains every condition. The no-addition baseline of seventeen correct answers belongs to this study and remains separate from the twenty-correct DeepSeek baseline in the earlier three-model comparison.

Table 17: Complete HiddenBench source-access comparison on the same 65 questions. "Support shared" means peers' supporting material accompanies their answers and confidences during the agent update.

Extra material at final decision	Answers shared	Support shared
No additional material	17/65	33/65
Identical initial report support	18/65	34/65
All original sources	25/65	34/65

Table 18: All retained HiddenBench source-access contrasts. Effects are accuracy percentage points, with the original task-bootstrap intervals.

Comparison	Effect	95% interval
Initial support added after answer-only exchange	1.54	−3.08 to 7.69
All originals added after answer-only exchange	12.31	3.08 to 21.54
Supporting-material exchange with no extra final evidence	24.62	13.85 to 35.38
Supporting-material exchange with identical initial support	24.62	13.85 to 35.38
Supporting-material exchange with all originals	13.85	4.62 to 23.08
Initial support added after supporting-material exchange	1.54	0.00 to 4.62
All originals added after supporting-material exchange	1.54	−3.08 to 7.69
Exchange and initial-support interaction	0.00	−6.15 to 6.15
Exchange and all-originals interaction	−10.77	−20.00 to −3.08

For the comparison with all originals, ten questions improve and one worsens after supporting-material exchange; the retained McNemar probability is 0.0117. Adding all originals after answer-only exchange improves nine questions and harms one, with probability 0.0215. The sixteen improvements and zero harms from supporting-material exchange without extra final evidence give a probability below 0.0001. These are task-level comparisons under the benchmark's updated-recommendation input.

D.2 Three model configurations and strict pairing

The three-model experiment also uses four agents and a final language-model synthesis. In the distributed allocation, each agent initially receives its own information. In the redundant allocation, every agent initially receives all originals. The treatment preserves peer answers and confidences

and adds real supporting material during one report update. The final model consumes the revised reports.

Table 19: Historical all-65 HiddenBench counts from the three-model study. Strict paired differences appear in Table 5.

Model	Allocation	Support shared	Answers shared
GPT-5.5	Distributed	52/65	28/65
GPT-5.5	Redundant	63/65	64/65
DeepSeek-V4-Flash	Distributed	33/65	20/65
DeepSeek-V4-Flash	Redundant	64/65	60/65
GLM-5.3-Flash	Distributed	55/65	27/65
GLM-5.3-Flash	Redundant	65/65	65/65

The retained pairing audit identifies two GPT questions with different initial drafts across the conditions. Excluding those questions leaves 63 matched pairs and the effects 36.51 points for distributed information and -1.59 for redundant information. The raw all-65 differences are 36.92 and -1.54. The strict audit's task membership and effect estimates are preserved; the paper presents the original marginal counts only for the population for which they were recorded. DeepSeek and GLM retain all 65 pairs.

D.3 M3GQA TASK CONSTRUCTION AND SCORING

M3GQA is the multi-entity graph question-answering benchmark of Peng et al. (2025). The 800-question cohort samples 200 validation questions from each of aggregation, multihop, set, and single-question strata. Frozen task selection excludes the earlier 100-question cohort and reserved questions. Of these 800 questions, 549 had appeared in other retained architecture experiments and 251 were unexposed and unreserved within the recorded selection audit.

Four agents receive allocated source units, with complementary units separated by the frozen allocation. Each initially returns entity answers, confidence, source references, and a support summary of at most one hundred words. The subsequent update receives peer answers and confidence values in both branches. One branch includes the peer supporting summaries and references; the other masks them. There is one agent update. The final model receives the updated answers and confidence values. Two final-input conditions either add no further support or add the identical frozen initial support package. Revised support summaries are absent from this final input.

The same frozen initial package appears after either update path. It contains selected supporting summaries and references. Complete original-source access is not the manipulated final factor. This gives four conditions per question. At most four initial agent calls, eight update calls, and four final calls are needed; identical final requests can share a response within a question. The recorded Flash configuration uses temperature zero, disabled thinking, and a 512-token output limit for these roles.

Answers are normalized to unique entity sets and scored by answer F1 against the reference set. A required invalid agent or final output gives its end-to-end condition a zero score according to the frozen evaluation. The 800-question metric uses weights $19/82$, $21/82$, $22/82$, and $20/82$ for aggregation, multihop, set, and single questions. These weights come from an earlier 82-question distribution. The 200-per-stratum sample remains fixed across all four conditions.

Table 20: M3GQA 800-question results. Values are weighted answer F1 multiplied by one hundred. The same initial support package is used in both supported final-input conditions.

Final-model support	Peer answers alone	Peer supporting material
No additional support	72.03	78.35
Identical initial support package	76.79	78.98

Under identical final support, 51 questions improve in F1, 25 decline, and 724 remain unchanged after supporting-material exchange. Adding final support after exchange improves 29 questions,

Table 21: All M3GQA 800-question contrasts. Effects are weighted answer-F1 points; intervals use the frozen stratified analysis.

Comparison	Effect	95% interval
Supporting-material exchange without final support	6.33	4.33 to 8.28
Final support after answer-only exchange	4.76	3.06 to 6.48
Supporting-material exchange with identical final support	2.19	0.77 to 3.63
Final support after supporting-material exchange	0.62	−0.30 to 1.57
Exchange and final-support interaction	−4.14	−6.04 to −2.21

harms fourteen, and leaves 757 unchanged. These are changes in set-valued answer F1. The originally considered two-point equivalence condition remains undecided under its retained analysis.

D.4 The separate 192-question M3GQA cohort

The second cohort contains 48 questions per stratum. Its original questions are disjoint from the 800-question cohort, the earlier 100 questions, and the reserved set. The selection audit records 122 questions shared with an earlier AutoGen experiment. The cohort therefore supplies a separate task sample for this comparison, with its own prior-exposure history. It uses the same agent and final-input design and gives its four strata equal weight.

Table 22: All conditions on the separate 192-question cohort. Values are answer F1 multiplied by one hundred, with equal stratum weights.

Final-model support	Peer answers alone	Peer supporting material
No additional support	65.32	74.34
Identical initial support package	71.96	76.07

Table 23: All principal contrasts on the separate 192 questions. Effects are answer-F1 points with equal stratum weights.

Comparison	Effect	95% interval
Supporting-material exchange without final support	9.03	4.92 to 13.45
Final support after answer-only exchange	6.64	3.56 to 10.06
Supporting-material exchange with identical final support	4.10	1.47 to 7.09
Final support after supporting-material exchange	1.72	0.07 to 3.63
Exchange and final-support interaction	−4.92	−8.52 to −1.64
Multihop versus single residual exchange effect	0.69	−6.25 to 8.33

The prespecified condition-level prediction was a larger residual exchange effect for multihop than for single questions. Its interval spans zero. The positive aggregate residual effect and this unresolved comparison remain distinct findings. The two M3GQA cohorts are not pooled into a new independent sample.

D.5 Source-only and report-input reference comparisons

Two historical HiddenBench batches compare which report fields reach a final model that always receives complete originals. The first uses retained agent reports and two final responses for each of 65 questions. The full-report condition includes answers, confidence, key facts, and rationale. The answers-only condition retains answers and confidence. The text-only condition retains supporting text. A sources-only reference removes the reports.

The second batch regenerates the member reports for two separate trajectories per question, giving 1,040 member calls and 650 final calls. Its two final responses per condition correspond to these separately generated member trajectories. The member output budget is 260 tokens. This is a distinct report-generation and final-decision batch.

Table 24: Historical HiddenBench report-utility controls with complete original sources. Each reported cell has 130 final decisions over 65 questions. Blank entries indicate a condition absent from that batch.

Reports formed after	Final-model report input	Retained reports	Regenerated reports
No report	Sources alone	113/130	114/130
Answer-only exchange	Answers and confidence	94/130	94/130
Answer-only exchange	Supporting text only	65/130	
Answer-only exchange	Full report	49/130	41/130
Supporting-material exchange	Answers and confidence	99/130	96/130
Supporting-material exchange	Supporting text only	73/130	
Supporting-material exchange	Full report	68/130	59/130

In the regenerated-report batch, the sources-only reference reaches 87.69% accuracy and the full report after supporting-material exchange reaches 45.38%. In the synthetic content batch, the corresponding descriptive ordering favors the reports. These results evaluate different task populations and report interfaces. They establish that the comparison between two report-formation paths and the comparison between reports and direct source reading answer different questions.

E Execution Records And Reproducibility

E.1 Final-model evaluations

The natural-content experiment contains 1,743 completed final-model requests: four component combinations, two unmodified-input references, and a sources-only reference for each of 249 variants. Every response is valid, and this run has no retries. The intact-report revision contains 5,976 requests, corresponding to three sharing schedules, four report conditions, 249 variants, and two responses per condition. The crossed role-and-position experiment contains 1,992 requests from four conditions, 249 variants, and two responses per condition. These are separate evaluation batches. Table 10, Table 13, and Table 16 report their complete condition sets.

E.2 Retained data and verification

The retained experimental archive records source texts, task references, agent reports, intervention inputs, final responses, task-level scores, and frozen comparison definitions. It includes the component separation used in the content experiment and the strict initial-draft pairing audit. An offline verifier checks source integrity, rescoring, task matching, scenario-level contrasts, and the stored bootstrap intervals. It also reconstructs M3GQA answer F1 from the saved answers and task snapshots. Each analysis retains its original experiment batch, task membership, scoring rule, and unavailable-state status. Earlier development studies remain in this separate archive; the present appendix provides the methods and complete comparisons for the experiments used in the paper. No experimental outcome was generated during this manuscript revision.