BEYOND ANSWER PRESERVATION: SCIENTIFIC EVIDENCE MINIMIZATION FOR PRIVATE LLM REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models can support scientific reasoning, but using external models often requires disclosing private scientific context. Reducing this disclosure while retaining useful answers is not sufficient for grounded reasoning: response preservation and downstream utility do not by themselves guarantee that the released context contains sufficient evidence to support the answer. We formulate **Scientific Evidence Minimization**, which seeks a policy-compliant, low-disclosure context that remains sufficient for grounded scientific reasoning. We introduce **CEDAR**, a verification-guided framework that iteratively evaluates an external answer, diagnoses missing evidence, and selectively expands the released context within the disclosure policy. We evaluate CEDAR across multiple scientific domains, reasoning tasks, and disclosure policies, explicitly separating answer correctness from evidence sufficiency to characterize how much scientific information is actually necessary for grounded reasoning.

1 INTRODUCTION

Large language models are increasingly used to interpret experiments, compare mechanisms, evaluate controls, and propose follow-up studies from scientific context. Such systems are often given as much relevant context as possible. Yet the evidence that is available is not necessarily the evidence that is necessary for a particular question. The same experimental record may support a mechanistic interpretation using a small set of relational observations, while a quantitative question may require an exact numerical state. This raises a basic question: *how much scientific evidence must actually be disclosed for a model to reason correctly and with adequate support?*

Our controlled study shows that scientific evidence sufficiency has three important properties. First, it is query-conditioned: for Q1–Q3 reasoning tasks, relational structure is substantially more useful than exact state without relations ($+0.625$ GTS), whereas the direction reverses for Q4–Q5 quantitative and exact-state tasks (-0.375). Second, sufficient evidence can be sparse but is not arbitrary. A support-complete representation retains only 21.6% of available relations while achieving 0.750 GTS, close to the 0.764 obtained with full relational context; a random subset with the same budget reaches only 0.333. Third, answer correctness alone can conceal insufficient grounding: support-complete and random-matched contexts obtain similar answer accuracy despite sharply different grounded success.

These observations motivate a problem different from ordinary context compression or fixed sanitization. The goal is not simply to shorten the context or remove predefined sensitive fields, but to identify a policy-compliant representation that reveals little scientific information while retaining enough evidence for grounded reasoning. We call this *Scientific Evidence Minimization*. Unlike answer-preserving minimization, success requires not only that the model produce the right answer, but also that the disclosed context itself support that answer.

Figure 1 illustrates this setting with a running example. The same private scientific context can require different evidence depending on the query. CEDAR searches over policy-valid disclosures and externalizes a task only when the released evidence is sufficient for a verified response.

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.

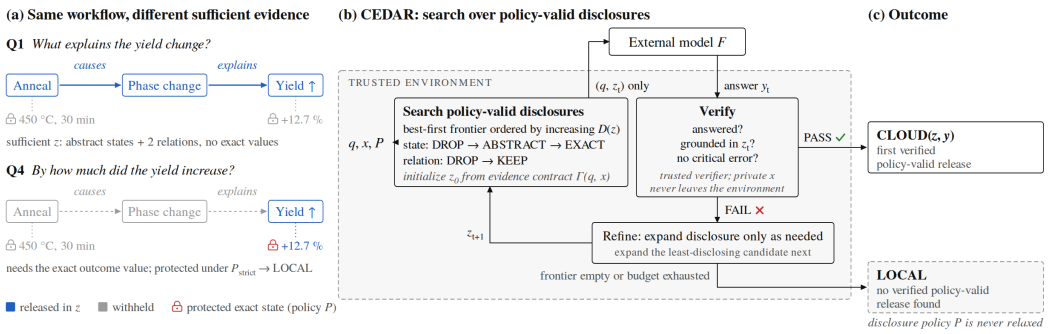


Figure 1: Overview of CEDAR. CEDAR searches over policy-valid disclosures to find sufficient evidence for each query and keeps the task local when no verified release is found.

We introduce CEDAR, a verifier-guided framework for approximating this objective without access to gold support annotations. CEDAR starts from a low-disclosure, policy-valid representation and uses trusted verification failures to guide selective disclosure expansion. If no sufficient policy-compliant external representation can be recovered, the unresolved computation remains in the trusted environment.

Across 1,600 policy–task cases, CEDAR achieves 0.578 end-to-end GTS, compared with 0.428 for task-conditioned One-Shot selection and 0.308 for answer-preserving data minimization. More importantly, this advantage is regime-dependent. Under the permissive policy, CEDAR and One-Shot are statistically indistinguishable (0.539 vs. 0.543 GTS). Under the strict policy, CEDAR reaches 0.618 GTS versus 0.314 for One-Shot while its final external specificity is only 0.043 of the raw scientific state and protected exact exposure remains zero. These results show that adaptive evidence recovery becomes most valuable when the information required by a query conflicts with what may be externalized.

Our study makes three contributions. First, we establish empirically that scientific evidence sufficiency is query-conditioned, sparse, and semantically selective. Second, we formulate Scientific Evidence Minimization as a grounded alternative to answer-preserving context minimization. Third, we introduce CEDAR, which uses verification-guided recovery and trusted-local fallback to adapt disclosure without access to gold evidence annotations.

2 Related Work

Evidence sufficiency and grounded scientific reasoning. Prior work distinguishes correct predictions from sufficient supporting evidence. ERASER evaluates whether selected rationales support model decisions (DeYoung et al., 2020), while SciFact and QASPER associate scientific claims or answers with supporting evidence (Wadden et al., 2020; Dasigi et al., 2021). ALCE evaluates whether generated claims are supported by cited evidence (Gao et al., 2023), and Joren et al. (2025) distinguish insufficient retrieved context from failures to use otherwise sufficient context. We ask a complementary question: rather than assessing whether a given context is sufficient, we ask which parts of a private scientific context must be disclosed for a particular query. Accordingly, sufficiency constrains the released representation itself and depends jointly on scientific-state specificity and evidential relations.

Minimal and adaptive context selection. Rationale selection seeks compact subsets that preserve predictions (Vafa et al., 2021), while prompt-compression methods such as Selective Context and LLMLingua reduce redundant context for inference efficiency (Li et al., 2023; Jiang et al., 2023a). Adaptive retrieval methods such as FLARE and Adaptive-RAG vary retrieved context across queries or during generation (Jiang et al., 2023b; Jeong et al., 2024), and ECoRAG compresses retrieved context according to evidentiality and retrieves additional evidence when the compressed context is insufficient (Jeong et al., 2025). These methods operate over retrieved or otherwise available context to improve retrieval, compression, or generation quality. CEDAR instead treats disclosure itself as

the optimization variable, searching over policy-valid representations of private scientific state while requiring the released representation itself to support the answer.

Data minimization and privacy-conscious delegation. Our closest optimization formulation is data minimization for LLM prompting: Zhou et al. (2026) search a privacy-ordered transformation space for a minimally revealing input that preserves response utility, while RUPTA iteratively revises text using privacy and utility feedback (Yang et al., 2025). Prompt-sanitization methods such as Rescriber and Casper reduce sensitive information before cloud execution (Zhou et al., 2025; Chong et al., 2024). Privacy-conscious delegation methods such as PAPILLON use local models to construct privacy-preserving requests for remote models (Siyan et al., 2025), while Privacy-R1 learns adaptive local–remote routing under privacy–utility objectives (Hui et al., 2026). CEDAR instead treats scientific evidence as the disclosure object: it searches over policy-valid scientific state and relations, requires the released representation to provide sufficient evidence for the answer, and retains execution locally when no verified policy-valid external release is found.

3 Query-Conditioned Minimum Sufficient Disclosure

3.1 Problem Formulation

Let $x = (S, R)$ denote a scientific context, where S contains scientific state—such as identities, numerical values, formulations, experimental conditions, and protocols—and R contains evidence relations among observations, controls, interventions, and scientific conclusions. Given a query q, the goal is not to preserve the entire context, but to identify the information required to answer that particular query.

We represent a released scientific context by z and let $D(z)$ denote its semantic disclosure cost, where smaller values correspond to less disclosure under the lexicographic ordering defined below. A disclosure policy P defines the set $\mathcal{Z}(x, P)$ of representations that may be released externally. The policy may designate selected state items as protected and restrict the levels of specificity at which they can be externally disclosed. For a target response model F, when a sufficient policy-compliant external representation exists, we define the ideal query-conditioned minimum sufficient disclosure as

$$z^*(q, x, P; F) = \arg \min_{z \in \mathcal{Z}(x,P)} D(z) \quad \text{s.t.} \quad \mathrm{GTS}(q, z; F) = 1, \tag{1}$$

where GTS denotes Grounded Task Success. The objective differs from ordinary context compression: it minimizes disclosed scientific information rather than token length, while requiring the resulting answer to remain grounded in the evidence that is actually released. The optimum is inherently query-conditioned. For the same experiment, a mechanistic question may require relations connecting perturbations, controls, and outcomes, whereas a quantitative question may require an exact numerical state.

3.2 Scientific Disclosure and Grounded Sufficiency

We instantiate z over two kinds of scientific information. Each state item $s_i \in S$ can be DROPped, ABSTRACTed, or retained EXACT, ordered as DROP $\prec$ ABSTRACT $\prec$ EXACT, while each evidence relation $r_j \in R$ is either DROP or KEEP. Rather than measuring disclosure by prompt length alone, we order candidate states lexicographically by $D(z) = (n_{\text{protected-exact}}, m_{\text{specificity}}, n_{\text{edges}}, n_{\text{chars}})$, where $m_{\text{specificity}}$ aggregates the disclosure levels of the released state items. We first minimize exact disclosure of protected state, then overall state specificity, retained relations, and finally serialized length. This defines an ordering over semantic disclosure states; two representations of similar length can therefore have substantially different disclosure costs.

Given an answer $y = F(q, z)$, we define Grounded Task Success as

$$\mathrm{GTS}(q, z; F) = \mathbb{1}\left[A(q, y) = 1 \wedge C(q, z, y) = 1 \wedge E(q, y) = 0\right], \tag{2}$$

where A denotes correctness of the central answer, C indicates that the answer is supported by the disclosed context, and E indicates a critical scientific error. This distinction matters because a model may produce the correct answer from parametric knowledge even when the released evidence is insufficient to justify it. The disclosure policy P determines which representations are externally feasible. If no $z \in \mathcal{Z}(x, P)$ satisfies the grounded-sufficiency constraint for the target model F, no

policy-compliant external solution to Eq. 1 exists. In this case, the appropriate outcome is trusted-local execution rather than relaxation of the disclosure policy. Finding a sufficient low-disclosure representation when one exists, without access to gold answers or gold support annotations, is the algorithmic problem addressed next.

4 CEDAR

CEDAR approximates the objective in Eq. 1 through bounded verifier-guided refinement. Starting from a low-disclosure policy-valid representation, it reveals additional scientific information only when the current release fails a grounded-sufficiency check. Unlike one-shot disclosure selection, CEDAR treats observed verification failures as online feedback for deciding whether and how to expand the released evidence, while never relaxing the active disclosure policy.

4.1 VERIFIER-GUIDED DISCLOSURE SEARCH

Before search, the trusted environment constructs a query-specific *evidence contract* $\Gamma(q, x)$ from the question and the full private scientific graph. The contract summarizes the kinds of evidence that may be needed for the query and serves only as a soft search prior. It is constructed without gold answers or gold support annotations, and evidence outside the contract may still be introduced during refinement.

CEDAR initializes a low-disclosure state z_0 under policy P. When the contract can be realized under the policy, the initializer constructs a contract-feasible state; otherwise, the requested evidence is projected to a policy-valid representation. If no valid seed can be constructed, the task is routed directly to trusted-local execution. Candidate states are explored in increasing order of the disclosure cost $D(z)$ defined in Section 3.2.

For each candidate z_t, the external model receives only (q, z_t) and produces $y_t = F(q, z_t)$. A trusted verifier with access to the private context x assesses whether the response addresses the task, whether its claims are supported by the released evidence, and whether it contains a critical scientific error. The verifier receives neither gold answers nor gold support annotations, and a candidate passes only when all grounded-sufficiency conditions are satisfied.

When verification fails, a trusted diagnoser proposes a small set of targeted repairs using the private context, current response, verifier feedback, and evidence contract. A repair may introduce a missing node or relation or increase the specificity of an already visible state item. Repairs are monotone and restricted to information present in the private scientific graph.

Each repair is projected through the active disclosure policy and relation-endpoint constraints before entering the search frontier. Search terminates when a candidate passes verification, no valid refinement remains, or the search budget is exhausted. CEDAR returns the first verified state reached by this bounded search and does not claim to recover the global optimum of Eq. 1. If no verified policy-compliant representation is found, the unresolved task is retained within the trusted environment. Algorithm 1 summarizes the procedure; implementation details, search limits, and runtime configuration are provided in Appendix C.

4.2 TRUST BOUNDARY

The private scientific graph, disclosure policy, evidence contract, verifier, diagnoser, and search state remain inside the trusted environment. The external response model receives only the query and the currently released representation. Gold answers and evaluation annotations are excluded from CEDAR search and are used only for post-hoc evaluation. LOCAL denotes completing the unresolved task entirely within the trusted environment using the private scientific state; LOCAL execution itself does not release additional private state to the external model. The search-time verifier is part of CEDAR's inference procedure and is distinct from the frozen post-hoc evaluator used to report the main results; the frozen evaluation protocol is detailed in Appendix D.

Algorithm 1 CEDAR: Verifier-Guided Disclosure Search

Require: Question q, private graph x, policy P, budget B
Ensure: $\textsc{Cloud}(z, y)$ or $\textsc{Local}$

$\Gamma \leftarrow \textsc{BuildContract}(q, x)$
$z_0 \leftarrow \textsc{InitializeSeed}(q, x, P, \Gamma)$
if $z_0 = \perp$ **then**
    return $\textsc{Local}$
$\mathcal{F} \leftarrow \{z_0\}$
while $\mathcal{F} \neq \emptyset$ and budget remains **do**
    $z \leftarrow \textsc{PopMin}(\mathcal{F}, D)$
    $y \leftarrow F(q, z)$
    $v \leftarrow \textsc{Verify}(q, z, y, x)$
    if $\textsc{Pass}(v)$ **then**
        return $\textsc{Cloud}(z, y)$
    $\Delta \leftarrow \textsc{Diagnose}(q, x, z, y, v, \Gamma)$
    for all $\delta \in \Delta$ **do**
        $z' \leftarrow \textsc{ProjectAndClose}(z, \delta, P)$
        if z' is valid and unseen **then**
            $\textsc{Push}(\mathcal{F}, z', D(z'))$
return $\textsc{Local}$

5 Experimental Setup

5.1 Benchmark and Evaluation Protocol

Our main benchmark contains 160 source-grounded scientific workflows spanning chemistry, materials science, physics, and biology, with 40 workflows per domain. Each workflow represents a coherent experiment, simulation, or mechanistic analysis from a single scientific source. We construct five questions per workflow: mechanism interpretation (Q1), control or alternative reasoning (Q2), source-grounded next experiment (Q3), quantitative comparison (Q4), and exact-state retrieval (Q5), yielding 800 tasks.

Each workflow is represented as $x = (S, R)$, separating scientific state from evidence relations. Gold evidence annotations are used only for post-hoc evaluation and controlled analysis and are never available to any method during transformation or search. CEDAR, One-Shot, and DM additionally receive no gold answers. Benchmark construction and task selection were completed independently of method outputs.

We evaluate every task under two pre-specified disclosure policies, P_{open} and P_{strict}, yielding 1,600 policy–task cases. The policies differ in the permitted specificity of designated protected state while sharing the disclosure space defined in Section 3.2. Because all source material is public, these designations simulate controlled scientific disclosure constraints rather than genuine confidentiality.

For the controlled study in Section 6.1, we use a separate set of 24 workflows from the same four domains. We manipulate state specificity and relational evidence and, on the 72 Q1–Q3 tasks, compare full relational context with a support-complete oracle, a relation-budget-matched random subset, and matched critical versus non-critical evidence removals. Benchmark construction and policy definitions are detailed in Appendix A, and the controlled intervention design is provided in Appendix B.

5.2 Models and Evaluation

CEDAR uses `openai/gpt-oss-120b` for trusted local computation, including evidence-contract generation, sufficiency verification, failure diagnosis, and LOCAL execution, running on three NVIDIA A100 40GB GPUs. External CLOUD execution uses `gpt-5.4-mini-2026-03-17`, which receives only the question and currently released representation. Structured local outputs are validated before use; retry and implementation details are given in Appendix C.

Condition	Rel.	Ans.	GTS
Full	100%	–	.764
Support	21.6%	.889	.750
Random	21.6%	.875	.333

Table 1: **Controlled evidence sufficiency on Q1–Q3.** Rel. denotes the retained relation fraction.

Sparse but selective. Support-complete retains only 21.6% of available relations while reaching 0.750 GTS, close to the 0.764 obtained with the full relational context. A random subset with the same relation budget reaches only 0.333.

Answer preservation is insufficient. Support-complete and random-matched contexts achieve similar answer accuracy (0.889 vs. 0.875), yet sharply different grounded success (0.750 vs. 0.333).

All methods use the same frozen post-hoc evaluation protocol. Q1–Q3 are evaluated with `gpt-5.6-terra` for answer correctness, released-context support, and critical scientific errors, while Q4–Q5 use deterministic task-specific scorers. GTS follows Eq. 2. For CEDAR, the same post-hoc criteria are applied to both CLOUD and LOCAL outputs. LOCAL execution may use the full private scientific state but does not itself constitute external disclosure. Full scoring rules and human validation of the frozen evaluator are reported in Appendix D.

We measure external disclosure using retained node, relation, and specificity fractions together with protected-exact exposure. For CEDAR, we additionally audit the union of scientific state exposed across the adaptive search trajectory. Confidence intervals use 20,000 paired workflow-cluster bootstrap replicates.

5.3 Baselines

We compare CEDAR with four fixed-release and two query-aware baselines. Policy-Max releases the maximally informative representation allowed by the active policy; Q-Only releases no scientific context; Direct Masking removes protected exact values while preserving the remaining content; and Uniform Abstraction applies policy-valid abstraction independently of the query.

One-Shot uses the same trusted local model, private structured context, question, and disclosure policy as CEDAR, but produces a single policy-valid disclosure assignment without candidate execution, sufficiency verification, diagnosis, refinement, or fallback search. It tests whether query-conditioned evidence selection is sufficient in one pass.

Data Minimization (DM) adapts the search framework of Zhou et al. (2026) to the same structured disclosure space. DM searches for a low-disclosure representation subject to preservation of the response produced from the raw full-exact scientific context, without evaluating whether the released context itself provides sufficient evidence for that response. The same response-preservation criterion is used across Q1–Q5.

Neither One-Shot nor DM receives CEDAR's evidence contract, verifier feedback, failure diagnosis, or gold evidence annotations. All externally released contexts are executed by the same response model under the same disclosure policies. Full baseline implementations and failure handling are provided in Appendix E.

6 Results

6.1 Evidence Sufficiency Is Query-Conditioned

Query-conditioned requirements. The controlled study reveals distinct evidence requirements across scientific queries. For mechanism interpretation, control reasoning, and next-experiment tasks (Q1–Q3), retaining correctly bound relational evidence while abstracting exact state outperforms retaining exact state without relations by $+0.625$ GTS; Table 1 further shows that only a sparse subset of these relations is needed. For quantitative comparison and exact-state retrieval (Q4–Q5), the direction reverses (-0.375 for structure minus state). Thus, no single representation is uniformly sufficient: reasoning-oriented queries depend more strongly on evidence structure, whereas quantitative and retrieval queries require greater state specificity.

These results establish the empirical premise behind CEDAR: scientific evidence minimization must be both query-conditioned and grounded in the evidence that remains disclosed. Neither disclosure

	Overall	Policy		Question Type				
Method	**GTS**	P_{open}	P_{strict}	Q1	Q2	Q3	Q4	Q5
Policy-Max	.648	.854	.443	.741	.919	.628	.481	.472
Q-Only	.004	.004	.004	.000	.006	.013	.000	.000
Direct Masking	.405	.405	.405	.656	.762	.544	.044	.019
Uniform Abstraction	.449	.449	.449	.731	.887	.625	.000	.000
DM	.308	.419	.196	.356	.403	.256	.225	.297
One-Shot	.428	.543	.314	.569	.672	.363	.209	.328
CEDAR	.578	.539	.618	.425	.656	.259	.741	.809

Table 2: **End-to-end GTS** over all 1,600 policy–task cases. CEDAR includes trusted-local execution when a sufficient policy-compliant external representation is not recovered.

amount nor preservation of the model's answer is sufficient to determine whether a representation contains the evidence required by the query. CEDAR therefore treats grounded sufficiency as the stopping criterion and expands a release only when the current representation fails verification. Additional controlled interventions and matched-deletion analyses are reported in Appendix B.

6.2 Policy Constraints Induce Distinct Operating Regimes

We next ask whether query-conditioned scientific evidence can be minimized without access to the gold support annotations used in the controlled study. Table 2 compares CEDAR with fixed-release baselines, DM, and One-Shot. One-Shot uses the same trusted local model, private context, and policy-constrained disclosure space as CEDAR, but produces a single disclosure proposal without verification, diagnosis, iterative refinement, or fallback.

CEDAR improves end-to-end grounded success. Across all 1,600 cases, CEDAR reaches 0.578 GTS, compared with 0.428 for One-Shot and 0.308 for DM. DM nevertheless preserves answer correctness well (0.876), reinforcing the distinction between preserving an output and preserving sufficient evidence for that output. Relative to One-Shot, CEDAR improves overall GTS by $+0.150$ (95% CI $[0.119, 0.181]$). Full pairwise comparisons and additional breakdowns are reported in Appendix F.

Disclosure policy determines the value of adaptation. Under P_{open}, CEDAR and One-Shot are statistically indistinguishable ($\Delta = -0.004$, 95% CI $[-0.044, 0.036]$), indicating that a single query-conditioned release is often sufficient when evidence can be externalized permissively. Under P_{strict}, the operating regime changes: CEDAR reaches 0.618 end-to-end GTS versus 0.314 for One-Shot ($\Delta = +0.304$, 95% CI $[0.271, 0.336]$), while the external-only Policy-Max baseline reaches 0.443.

This policy-dependent transition is central to CEDAR. When the evidence needed for a task can be represented within the disclosure boundary, reasoning can remain external over a compact release. When the required scientific state cannot be externalized at an admissible specificity, CEDAR preserves the policy boundary by retaining the unresolved computation in the trusted environment rather than releasing more sensitive state. Under P_{strict}, CEDAR operates at final external specificity 0.043, compared with 0.123 for One-Shot, while protected-exact exposure remains zero. These quantities measure information transmitted outside the trust boundary; scientific state retained for LOCAL execution remains inside the trusted environment.

6.3 Task Requirements Determine Adaptation and Execution

The policy effect becomes clearer when results are separated by scientific question type. Figure 3 reveals two distinct operating regimes associated with the evidence requirements of the query.

Reasoning-oriented queries benefit from relational evidence. For Q1–Q3, methods that preserve relational structure remain strong: Uniform Abstraction achieves high GTS while retaining abstract state and broad relational context, and One-Shot is competitive with CEDAR on these tasks. To-

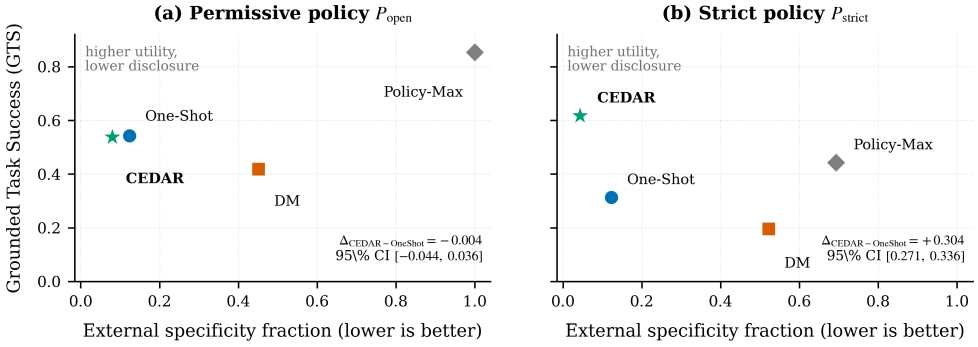


Figure 2: **Utility–disclosure regimes.** End-to-end GTS versus final external specificity under P_{open} and P_{strict}. CEDAR GTS includes trusted-local execution.

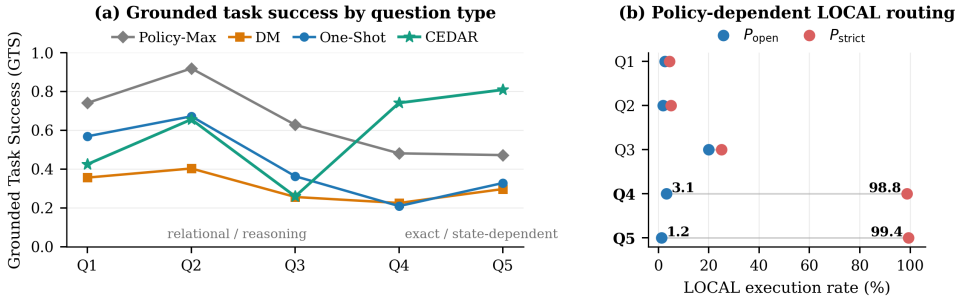


Figure 3: **Task-dependent adaptation and execution.** (a) End-to-end GTS by question type for the principal adaptive and high-disclosure reference methods; (b) CEDAR LOCAL execution rate under the permissive and strict disclosure policies.

gether with the controlled study, this indicates that mechanism, control, and next-experiment questions can often be resolved without exact scientific state when the relevant relational evidence remains available. The complete paired comparisons and verifier audit are reported in Appendices F and G.1.

State-dependent queries induce execution localization. For Q4 and Q5, where successful resolution requires quantitative or exact scientific state, CEDAR improves over One-Shot by $+0.531$ and $+0.481$ GTS, respectively. Under P_{strict}, LOCAL routing rises to 98.8% for Q4 and 99.4% for Q5, while remaining at or below 25% for Q1–Q3. Thus, the same disclosure policy induces qualitatively different operating regimes depending on the evidence required by the query: relational reasoning is usually externalized from abstract evidence, whereas exact state-dependent reasoning is predominantly retained within the trusted environment. This transition shows that evidence minimization concerns both what to release and where to compute. We next test whether refinement expands policy-compliant external execution.

6.4 Verification and Refinement Govern Execution Placement

We next separate the roles of verifier-gated recovery and iterative refinement on a balanced 160-case diagnostic subset. The Gate/LOCAL control applies CEDAR's frozen verifier once to the One-Shot release and routes rejected cases directly to trusted-local execution, without diagnosis or refinement. We additionally vary CEDAR's verifier-call budget $B \in \{1, 2, 4, 8\}$ under the same frozen configuration.

The two controls reveal complementary roles: verification governs stopping and execution placement, whereas refinement expands the set of policy-valid states that satisfy this criterion. Full mechanism diagnostics are reported in Appendix G.2.

Condition	GTS	CLOUD
One-Shot	.438	–
+ Gate/LOCAL	.706	.556
CEDAR $B = 1$	.706	.000
CEDAR $B = 2$	.681	.613
CEDAR $B = 4$	.694	.787
CEDAR $B = 8$	.694	.775

Table 3: **Execution-control diagnostic.** CLOUD denotes termination with a verifier-accepted external representation.

Verification enables reliable fallback. Adding only CEDAR's verifier gate and LOCAL recovery raises One-Shot GTS from 0.438 to 0.706, while 55.6% of cases still complete through CLOUD.

Refinement expands external execution. Increasing the search budget raises verified CLOUD completion to 78.7% at $B = 4$ while maintaining 0.694 GTS. Refinement therefore primarily increases the set of tasks that can be resolved from a policy-compliant external representation before LOCAL execution is required.

6.5 Low External Disclosure Persists Across Adaptive Search

We finally examine whether CEDAR's low-disclosure operating point persists when intermediate search candidates are included in the accounting. Under P_{strict}, CEDAR's final external release retains only 0.076 of nodes, 0.046 of relations, and 0.043 of raw specificity mass, compared with 0.245, 0.149, and 0.123 for One-Shot, respectively, while protected-exact exposure remains zero. Accounting for the full adaptive trajectory increases these quantities only moderately: across the 795 strict-policy cases with complete traces, the union of all externally transmitted candidates contains 0.128 of nodes, 0.072 of relations, and 0.071 of specificity mass. This search-wide exposure remains below even the final-release exposure of One-Shot, DM, and Policy-Max.

These quantities measure information crossing the trust boundary rather than private computation performed inside it. LOCAL execution may use the full private scientific state, but does not itself add external disclosure; candidate evidence transmitted before fallback is already included in the search-wide audit. Full policy-level exposure and external-call accounting are reported in Appendix G.4.

7 Conclusion and Discussion

We introduced Scientific Evidence Minimization, which seeks to minimize scientific information disclosed to an external model while preserving sufficient evidence for grounded reasoning. Our controlled study shows that evidence sufficiency is query-conditioned and selective: reasoning-oriented questions depend more strongly on relational structure, whereas quantitative and exact-state questions require greater specificity, and answer correctness alone does not guarantee adequate support. We operationalized this formulation with CEDAR, combining query-conditioned selection, verification-guided recovery, and trusted-local execution. Its benefit is regime-dependent: CEDAR and One-Shot perform similarly under permissive disclosure, while CEDAR achieves substantially higher grounded success under strict disclosure.

Evidence minimization therefore concerns both what information is released and where computation occurs. When sufficient evidence can be represented within the disclosure policy, CEDAR externalizes a compact representation; when required scientific state cannot be released at an admissible specificity, reasoning remains inside the trusted environment. Our mechanism diagnostics further show that verification primarily provides reliable execution placement, while refinement expands the set of tasks that can be completed from policy-compliant external evidence. Because adaptive search may transmit multiple candidates before this decision, disclosure must be measured over the interaction trajectory rather than only the final release.

Limitations and future directions. CEDAR assumes a trusted environment, a structured evidence space, and simulated disclosure policies over public sources, and currently evaluates one external/local model configuration with an imperfect search-time verifier. Future work should examine richer scientific workflows, stronger verification, cross-model generalization, and learned or human-specified policies. Overall, our results suggest that scientific reasoning under controlled disclosure requires jointly deciding what evidence to release and where reasoning should occur.

AI USE STATEMENT

We used generative AI tools to assist with research ideation and experimental design, literature discovery and retrieval, construction and formatting of source-grounded benchmark artifacts, implementation and debugging of experimental code, and drafting and editing portions of the manuscript. All AI-assisted benchmark content was checked against the underlying scientific sources, experimental outputs were validated through the frozen evaluation and quality-control procedures described in the paper, and AI-assisted manuscript content and citations were reviewed by the authors. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies controlled disclosure of scientific information. The benchmark is constructed from public primary-research articles; protected-state annotations simulate confidentiality constraints and do not imply that the underlying publications are private. The experiments do not use personal or sensitive human data. Human validation is limited to evaluating scientific model outputs and their evidential support. Our results should therefore be interpreted as evidence about controlled scientific disclosure under the stated threat model, rather than as a guarantee of privacy for real confidential scientific deployments.

REPRODUCIBILITY STATEMENT

We provide the details needed to reproduce the experiments throughout the main paper and appendix. Appendix A specifies benchmark construction, annotations, policies, and quality control; Appendix B describes the controlled interventions and statistical protocol; Appendix C gives CEDAR's implementation, search limits, and model configuration; Appendix D specifies the frozen evaluation and human validation procedures; and Appendix E documents all baseline implementations. Additional statistical results and system audits are reported in Appendices F and G.

REFERENCES

Chun Jie Chong, Chenxi Hou, Zhihao Yao, and Seyed Mohammadjavad Seyed Talebi. Casper: Prompt sanitization for protecting user privacy in web-based large language models. *arXiv preprint arXiv:2408.07004*, 2024. doi: 10.48550/arXiv.2408.07004. URL https://arxiv.org/abs/2408.07004.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365/.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

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, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Zheng Hui, Yijiang River Dong, Sanhanat Sivapiromrat, Ehsan Shareghi, and Nigel Collier. Privacy-R1: Privacy-aware multi-LLM agent collaboration via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pp. 45924–45937, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.2130. URL https://aclanthology.org/2026.acl-long.2130/.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

Yeonseok Jeong, Jinsu Kim, Dohyeon Lee, and Seung-won Hwang. ECoRAG: Evidentiality-guided compression for long context RAG. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26607–26628, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.1365. URL https://aclanthology.org/2025.findings-acl.1365/.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13358–13376. Association for Computational Linguistics, 2023a. doi: 10.18653/v1/2023.emnlp-main.825. URL https://aclanthology.org/2023.emnlp-main.825/.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Jjr2Odj8DJ.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6342–6353. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.391. URL https://aclanthology.org/2023.emnlp-main.391/.

Li Siyan, Vethavikashini Chithrra Raghuram, Omar Khattab, Julia Hirschberg, and Zhou Yu. PAPILLON: Privacy preservation from internet-based and local language model ensembles. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3371–3390, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.173. URL https://aclanthology.org/2025.naacl-long.173/.

Keyon Vafa, Yuntian Deng, David Blei, and Alexander Rush. Rationales for sequential predictions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10314–10332. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.807. URL https://aclanthology.org/2021.emnlp-main.807/.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. Robust utility-preserving text anonymization based on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28922–28941. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1404. URL https://aclanthology.org/2025.acl-long.1404/.

Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. Rescriber: Smaller-LLM-powered user-led data minimization for LLM-based chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–28. ACM, 2025. doi: 10.1145/3706598.3713701. URL https://doi.org/10.1145/3706598.3713701.

Jijie Zhou, Niloofar Mireshghallah, and Tianshi Li. Operationalizing data minimization for privacy-preserving LLM prompting. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=rpcnvW33EG.

A Benchmark Construction, Auditability, and Example

A.1 Source and Workflow Construction

The main benchmark contains 160 independent scientific workflows spanning chemistry, materials science, biology, and physics, with 40 workflows per domain. Each workflow is grounded in a single primary-research paper and represents one coherent experiment, simulation, observational analysis, or mechanistic study. We exclude review and survey articles and do not combine scientific evidence from multiple papers within a single workflow. Each workflow is paired with five question types, yielding 800 scientific tasks and 1,600 policy–task cases after evaluation under the two disclosure policies.

Workflow selection was completed before the main CEDAR and baseline experiments. CEDAR outputs, baseline outputs, and downstream model performance were not used to select or prune the final workflows. The retained 160 workflows all match their registered source identifiers and DOIs, with no duplicate DOI or paper title in the final benchmark.

To avoid concentrating the benchmark on a single experimental style, each domain covers three source families:

Domain	**Experimental families**
Materials	Structure–property (13), processing–treatment (13), degradation–performance (14)
Chemistry	Reaction–catalysis (14), spectroscopy–mechanism (13), synthesis–condition control (13)
Biology	Perturbation–phenotype (13), pathway–mechanism (13), treatment/condition–response (14)
Physics	Measurement–phenomenon (13), device–field response (13), model–parameter comparison (14)

Table 4: **Domain composition of the main benchmark.** Each domain contains 40 source-grounded workflows.

Each selected workflow must contain sufficient source-grounded information to support multiple scientific reasoning tasks over the same underlying study. The benchmark therefore preserves experimental state, measurements and observations, comparisons or interventions, candidate explanations, and the relations connecting these elements, rather than reducing each source to a single question–answer pair.

A.2 Scientific State and Evidence-Relation Annotation

Each workflow is represented as a structured scientific context $x = (S, R)$. The state set S contains scientific entities and attributes whose presence and specificity can be controlled during disclosure. These include experimental conditions, numerical measurements, treatments, outcomes, scientific states, candidate explanations, locations, and measurement procedures. The relation set R records how visible scientific state items are connected.

Each state node contains a node identifier, scientific type, exact representation, abstract representation, source-evidence identifiers, and, when applicable, one or more registered protected fields. A protected field records the exact value, its field type, and an abstract scientific role. Each relation

Object	Core annotation	Purpose
State node	ID, type, exact text, abstract text, protected fields, evidence IDs	Represents scientific information whose visibility and specificity can be controlled independently.
Relation edge	ID, source, relation, target, evidence IDs	Represents the source-grounded evidential connection between visible state items.

Table 5: **Core benchmark representation.** Scientific state and evidence relations are represented separately so that specificity and relational structure can be manipulated independently.

contains a relation identifier, source node, relation type, target node, associated source evidence, and an annotation rationale.

This separation is important because preserving the relevant scientific entities does not necessarily preserve the relations required to justify a conclusion. Conversely, relational evidence may remain informative even when exact scientific values are abstracted. Appendix B uses this representation to independently manipulate state specificity and evidence structure.

Protected-field annotations define the controlled disclosure experiment. Because the underlying scientific papers are public, these annotations simulate information that would be treated as confidential in a private scientific workflow rather than asserting that the publication itself is secret.

A.3 Question and Support Construction

Each workflow contains five questions designed to require qualitatively different uses of the same scientific record:

- **Q1: Mechanism interpretation** asks which mechanism or scientific explanation is best supported by the observations.
- **Q2: Control or alternative reasoning** asks how a comparison, control, or contrasting observation bears on an alternative explanation.
- **Q3: Source-grounded next experiment** asks which follow-up experiment would most directly resolve an uncertainty using procedures already used or explicitly motivated by the source.
- **Q4: Quantitative comparison** requires an explicit numerical comparison or calculation from source-reported values.
- **Q5: Exact-state retrieval** requires recovery of a registered exact scientific value together with the correct scientific binding.

For every question, the benchmark stores a source-grounded critical answer, acceptable alternatives, required nodes, required relations, required scientific content, known critical-error conditions, and non-critical evidence. Q4 additionally stores its numerical inputs, operation, and expected result. Q5 registers the target node, protected field type, and exact expected value.

These annotations are evaluation metadata, not CEDAR inputs. Gold answers and gold support annotations are excluded from CEDAR's evidence-contract construction, verification, diagnosis, and disclosure search. One-Shot is constructed under the same gold-free condition. The baseline-specific utility signal used by adapted Data Minimization is described separately in Appendix E.4.

Because Q1–Q5 share the same workflow, the benchmark directly tests whether evidence sufficiency changes with the query. A workflow is retained only when the five questions do not all reduce to the same annotated support set.

A.4 Disclosure Policies

Each of the 800 tasks is instantiated under two pre-specified disclosure policies, P_{open} and P_{strict}.

Item	$\mathbf{P}_{\text{open}}$	$\mathbf{P}_{\text{strict}}$
Protected state	All available levels	Exact prohibited; abstract allowed when available
Unprotected state	All available levels	All available levels
Evidence relation	Drop / Keep, subject to visible endpoints	Drop / Keep, subject to visible endpoints

Table 6: **Frozen disclosure-policy semantics.** The strict policy constrains exact disclosure only for registered protected state.

A node can be represented at DROP, ABSTRACT, or EXACT when a distinct abstract representation exists. If the exact and abstract strings are identical, ABSTRACT is not treated as a distinct disclosure action. Relations use the binary levels DROP/KEEP.

Under P_{open}, all available node specificity levels are permitted. Under P_{strict}, EXACT is prohibited only for nodes containing registered protected fields; unprotected nodes retain their ordinary available levels. If a protected node has no distinct abstract representation, the strict policy permits only DROP. Relations may be retained under either policy, subject to endpoint consistency: an edge can be visible only when both of its endpoint nodes are visible.

The same policy validation and endpoint-closure rules are applied across all structured methods. CEDAR does not relax the active disclosure policy during search. When a sufficient external representation cannot be constructed within the policy, unresolved computation can instead remain within the trusted environment.

A.5 DETAILED WORKFLOW EXAMPLE

We illustrate the benchmark using workflow `MAIN_MAT_SRC001`, derived from the 2024 materials-science paper *Operando real-space imaging of a structural phase transformation in the high-voltage electrode* $Li_xNi_{0.5}Mn_{1.5}O_4$ (DOI: 10.1038/s41467-024-55010-6).

The selected experiment studies how the internal phase boundary of a single $\mathrm{Li}_x\mathrm{Ni}_{0.5}\mathrm{Mn}_{1.5}\mathrm{O}_4$ nanoparticle evolves during Li-intercalation-driven discharge. The workflow combines the operando imaging sequence with coherent- and semi-coherent-interface microelasticity analysis, displacement-field defect analysis, a Burgers-vector consistency check, and the source's resolution-limited discussion of potentially unresolved interphases.

The observed Li-rich phase nucleates and grows within the initially Li-poor particle while the phase boundary changes from curved to planar. The later planar state contains an array of misfit dislocations, supporting a coherent-to-semi-coherent interpretation rather than a boundary that remains fully coherent throughout.

The complete workflow contains 20 evidence atoms, 32 state nodes, 26 relation edges, and 17 distinct relation types. Table 7 shows the subset most directly involved in Q1–Q5.

The relation graph binds these state items into evidence structures. For Q1, the observed curved interface precedes the planar interface (R4: N3 $\rightarrow$ N4). The coherent-interface candidate predicts an overall spherical geometry (R7), which fits the observed curved interface (R8), whereas the semi-coherent candidate predicts a planar preferred geometry (R11), which fits the later planar interface (R12). The early continuous displacement field supports the coherent candidate (R15), while the later misfit-dislocation array supports the semi-coherent candidate (R16) and is linked to the transition to the planar state (R21).

A different branch of the same graph supports Q3. The reported absence of interphases is explicitly bound to the current spatial resolution (R22). Increased diffraction signal increases spatial resolution (R23), and higher spatial resolution enables detection of the potentially present interphase (R24). Operando BCDI is linked to the structural transformation being measured (R25). Q5 uses yet another binding: R17 connects the [110] Burgers-vector state N18 to the 5.1-radian phase-discontinuity measurement N19.

Table 8 shows the resulting query-specific support annotations.

ID	Type	Exact representation	Abstract representation
N3	State	curved interface	a curved phase boundary
N4	State	planar interface	a planar phase boundary
N5	Candidate	coherent interface	a fully lattice-continuous interface candidate
N6	Candidate	semi-coherent interface	an interface candidate with localized coherency loss
N9	Measurement	8% difference between the maximum and minimum $B(n)$ directions	the coherent-model directional spread in $B(n)$
N12	Measurement	197% difference between the lowest and highest $B(n)$ directions	the semi-coherent-model directional spread in $B(n)$
N15	Measurement	continuous displacement field without singularities	a continuous defect-free displacement-field map
N17	State	array of misfit dislocations	an interfacial dislocation array
N18	State	Burgers vector along [110]	the Burgers-vector direction used in the consistency check
N19	Measurement	phase discontinuity of 5.1 radians	the phase-discontinuity magnitude from the Burgers-vector consistency check
N25	Condition	current spatial resolution	the spatial resolution available in the reported experiment
N26	Measurement	no interphases observed	absence of an observed interphase signal
N27	Candidate	potentially present interphases	a possible interphase candidate
N28	Condition	increased diffraction signal	a stronger diffraction signal
N29	Measurement	higher spatial resolution	improved spatial resolution
N31	State	discontinuous structural phase transformation in $Li_xNi_{0.5}Mn_{1.5}O_4$	the discontinuous structural transformation under study
N32	Measurement	operando Bragg Coherent Diffractive Imaging (BCDI)	the operando coherent X-ray imaging measurement used in the study

Table 7: **Selected state nodes from `MAIN_MAT_SRC001`.** The exact and abstract representations are taken from the frozen benchmark annotation.

This example makes the benchmark's central distinction concrete. Q1 and Q2 depend on multiple relational bindings connecting observations, models, and candidate explanations. Q3 requires a different evidence branch describing measurement resolution and a source-motivated follow-up experiment. Q4 requires two exact quantitative inputs and their model bindings. Q5 requires only two state nodes and one relation, but the requested value itself is a registered protected field.

For this workflow, N18 registers the exact state label `[110]` and N19 registers the exact numeric value `5.1 radians` as protected fields. Therefore, P_{open} permits the exact Q5 evidence to be externalized, whereas P_{strict} prohibits those exact protected representations. This illustrates why evidence sufficiency cannot be separated from the active disclosure policy: a compact sufficient representation may exist under one policy but not under another.

A.6 Quality Control and Benchmark Auditability

Benchmark selection and quality control were completed before observing main method performance. The frozen source-identity audit confirms that all 160 retained workflows match their registered source identifiers and DOIs, with zero source-identity mismatch, duplicate DOI, or duplicate paper title.

The final benchmark additionally satisfies the following structural requirements:

- 160 workflows, with 40 workflows from each of four scientific domains;
- exactly five question types per workflow, yielding 160 instances of each of Q1–Q5;
- an explicit quantitative comparison for every Q4 task;
- a registered exact-state target corresponding to a real protected field for every Q5 task;
- non-empty critical answers and source-grounded evaluation annotations;

Q	Task	Question	Required evidence
Q1	Mechanism	What interface-evolution mechanism is best supported by the operando imaging and microelasticity analysis for the observed change from a curved to a planar phase boundary?	Nodes: N3, N4, N5, N6, N10, N13, N15, N17. Edges: R4, R7, R8, R11, R12, R15, R16, R21.
Q2	Control / alternative	What does the early-versus-late displacement-field comparison show about the alternative that the interface remains coherent throughout the transformation?	Nodes: N4, N5, N6, N14, N15, N16, N17. Edges: R13, R14, R15, R16, R21.
Q3	Next experiment	Which next experiment would most directly resolve whether a transient interphase exists despite not being observed in the reported data?	Nodes: N25, N26, N27, N28, N29, N31, N32. Edges: R22, R23, R24, R25.
Q4	Quantitative	By how many percentage points does the directional difference between the lowest and highest $B(n)$ values increase from the coherent-interface calculation to the semi-coherent calculation?	Nodes: N5, N6, N9, N12. Edges: R6, R10. Inputs: 8%, 197%; target: 189 percentage points.
Q5	Exact state	What exact phase-discontinuity value is reported as being consistent with a Burgers vector along [110]?	Nodes: N18, N19. Edge: R17. Exact target: 5.1 radians.

Table 8: **Query-conditioned support within one scientific workflow.** The same experiment yields five questions with different required state, relations, and specificity.

- at least one non-critical evidence item per workflow; and
- non-identical support requirements across the five questions of each workflow.

For the detailed example above, the frozen workflow-level QC records 20 evidence atoms, 32 state nodes, 26 relations, and 17 distinct relation types, and verifies the presence of a control or comparison, competing explanations, explicit Q4 source values, a protected Q5 target, non-critical evidence, and question-dependent support sets.

The source-identity audit establishes bibliographic correspondence between each retained workflow and its intended paper and DOI; it should not be read as a claim that every annotation in all 160 workflows was independently re-derived through a second exhaustive line-by-line review of every source paper. The structured benchmark instead preserves provenance fields that make individual evidence atoms, answer annotations, and support bindings auditable.

We use public primary-research sources precisely to make scientific provenance, reference answers, and evidence annotations inspectable and reproducible. The resulting disclosure policies simulate confidential scientific deployment rather than asserting that the source publications themselves are private. Potential parametric familiarity with a public source does not by itself satisfy the evaluation criterion: if a model produces a correct answer from prior knowledge but the released context does not support that answer, then $C(q, z, y) = 0$ and the instance receives $\mathrm{GTS} = 0$ under Eq. 2.

The benchmark should therefore be interpreted as a controlled testbed for a specific question: given the same source-grounded scientific record, which scientific state and evidence relations must be disclosed for a particular query, at what specificity, and under which disclosure policy?

B Controlled Evidence-Sufficiency Study

B.1 Study Design

The controlled study uses a separate set of 24 scientific workflows, with six workflows from each of chemistry, materials science, biology, and physics. Each workflow contains the same five question types as the main benchmark, yielding 120 tasks. These workflows are disjoint from the 160-

workflow benchmark used for the main CEDAR comparison; the controlled study is used only to characterize the evidence-sufficiency phenomenon.

We conduct two complementary interventions. First, we independently vary scientific-state specificity and relational structure across all five question types. Second, on the 72 Q1–Q3 reasoning tasks, we hold the state representation fixed and manipulate the identity and amount of retained relational evidence. These interventions are constructed from the frozen benchmark annotations and are not outputs of CEDAR.

B.2 State–Structure Interventions

To separate state specificity from relational evidence, we construct the following representations of each workflow while holding the underlying scientific source and question fixed:

- **Full**: exact scientific state with all correct L0–L2 relations;
- **Structure-Only**: abstract scientific state with all correct L0–L2 relations;
- **State-Only**: exact scientific state with relations removed; and
- **Neither**: abstract scientific state with relations removed.

The key comparison is therefore not between a long and a short context, but between representations that preserve different kinds of scientific information. For Q1–Q3, Structure-Only exceeds State-Only by $+0.6250$ GTS (95% CI $[0.5278, 0.7222]$). For Q4–Q5, the same contrast reverses to -0.3750 (95% CI $[-0.5208, -0.2292]$). The resulting difference-in-differences is $+1.0000$ (95% CI $[0.8056, 1.1944]$).

This interaction is the basis for treating scientific evidence requirements as query-dependent: relational structure is especially important for the reasoning-oriented questions, whereas quantitative comparison and exact-state retrieval depend more strongly on state specificity.

B.3 Relation Identity and Selectivity

A possible alternative explanation for the Structure-Only result is that models benefit merely from relation labels or graph-like formatting, rather than from correctly bound scientific evidence. We test this with three additional Q1–Q3 conditions, all using abstract state:

- **Structure-Noninferential**: retains the correct direct and source-analytic relations (L0–L1) while excluding interpretive L2 relations;
- **Structure-Shuffled**: retains the relation types but corrupts their source–target bindings; and
- **Relation-Inventory**: exposes the available relation labels without their source–target bindings.

Structure-Noninferential exceeds Structure-Shuffled by $+0.4306$ GTS (95% CI $[0.2917, 0.5694]$) and Relation-Inventory by $+0.3611$ (95% CI $[0.2222, 0.5000]$). Thus, the observed benefit is not explained by the presence of relation vocabulary or structured formatting alone; the identity of the scientific endpoints and their correct bindings materially affect grounded success.

B.4 Sparse Evidence and Matched Deletions

We next ask whether reasoning-oriented questions require the full relational graph once the relevant form of evidence is known. This analysis uses Q1–Q3 only and fixes the scientific state at the abstract representation. Here, **Full** refers specifically to retaining all available L0–L1 relations under this fixed abstract-state condition.

We compare:

- **Support-Complete**: an annotation-derived subset that satisfies all frozen support requirements while minimizing the retained relations;

- **Random-Matched**: a random subset matched to Support-Complete in both total relation count and L0/L1 composition;
- **Critical-Group-Drop**: a minimal deletion that breaks one selected critical support group while preserving the remaining required groups; and
- **Matched-Noncritical-Drop**: a deletion matched in size and relation-level composition that preserves the required support groups.

Support-Complete is an oracle diagnostic used only in this controlled analysis; its annotated support sets are never supplied to CEDAR during inference. It retains 21.6% of available relations on average (median 19.3%).

The matched conditions are intended to separate evidence identity from evidence quantity. Support-Complete and Random-Matched therefore receive the same relation budget, while the critical and non-critical deletion conditions remove matched amounts of relational evidence. The latter paired analysis is defined for the 67 tasks for which both deletion conditions were instantiated.

Despite nearly identical answer accuracy under Support-Complete and Random-Matched (0.8889 versus 0.8750), their GTS values are 0.7500 and 0.3333, respectively. More broadly, across the state–structure study, 284 of 572 representations that fail the frozen representation-level support requirements nevertheless produce the correct central answer. These cases motivate evaluating evidential support separately from answer correctness.

B.5 Statistical Analysis

The workflow, rather than the individual question, is the statistical resampling unit. Confidence intervals use 20,000 workflow-cluster bootstrap replicates with seed 20260828, preserving all questions from the same workflow within each bootstrap sample.

For the sparse-evidence analysis, we predefine three paired contrasts. First, Full versus Support-Complete tests whether the sparse oracle preserves GTS within a non-inferiority margin of 0.10. The observed difference,

$$\mathrm{GTS}_{\mathrm{Full}} - \mathrm{GTS}_{\mathrm{Support}} = 0.0139,$$

has a 95% CI of $[-0.0833, 0.0972]$; the upper confidence bound remains below the predefined margin.

Second, Support-Complete exceeds Random-Matched by $+0.4167$ GTS (95% CI $[0.3056, 0.5278]$), testing evidence identity at a matched relation budget. Third, Matched-Noncritical-Drop exceeds Critical-Group-Drop by $+0.1389$ GTS (95% CI $[0.0278, 0.2500]$), testing whether removal of annotated critical evidence is more consequential than removal of a matched amount of non-critical evidence.

These controlled comparisons characterize the evidence-sufficiency phenomenon independently of CEDAR; the method comparison in the main experiments uses the separate 160-workflow benchmark.

C CEDAR Implementation Details

C.1 Threat Model and Trust Boundary

We assume an honest-but-curious external model provider. The provider follows the inference protocol but may inspect every representation transmitted during CEDAR's adaptive search and may use the disclosed information together with public scientific knowledge to infer protected scientific state.

The trusted environment contains the full scientific graph $x = (S, R)$, the disclosure policy P, the evidence contract, the search frontier, the search-time verifier, the failure diagnoser, and all unreleased scientific state. The external response model receives only the scientific question q and the currently released representation z_t. Gold answers, annotated support sets, reference answers, and post-hoc evaluation judgments are never available to the external model or to the CEDAR search procedure.

Because CEDAR may evaluate multiple external candidates, the adversary is assumed to observe the entire sequence of externally transmitted representations rather than only the terminal release. We therefore separately report search-wide disclosure over the union of all transmitted scientific state in Appendix G.4.

Our evaluation quantifies protocol-level scientific disclosure across this trust boundary, including which scientific state, relations, specificity, and serialized content are transmitted to the external provider. It does not provide a guarantee against arbitrary downstream inference from the released evidence. Cross-query composition, joint inference over released representations and generated responses, and reconstruction attacks using additional external databases or auxiliary knowledge are outside the present scope.

LOCAL execution remains entirely inside the trusted environment and may access the full private scientific state. It does not itself release additional private state to the external model. Compromise of the trusted environment and network side channels are outside the scope of this work.

C.2 EVIDENCE CONTRACT AND INITIALIZATION

CEDAR first constructs a query-specific evidence contract $\Gamma(q, x)$ from the scientific question and the full private graph. The contract describes scientific obligations that may be needed for a grounded answer and may include alternative realizations when multiple evidence paths could satisfy the same obligation.

The contract generator does not receive the gold answer, annotated required nodes or relations, acceptable support alternatives, teacher responses, or the disclosure policy. This separation is deliberate: $\Gamma(q, x)$ is intended to describe the scientific evidence requirement of the query, while the policy determines which realizations of that requirement may cross the trust boundary. The contract is therefore a search prior rather than a sufficiency oracle.

The initializer combines $\Gamma(q, x)$ with the frozen disclosure lattice and the active policy P. When the contract can be directly realized under the policy, it constructs a low-disclosure contract-feasible seed. Otherwise, the requested evidence is projected to a policy-valid representation using the same policy-projection and endpoint-closure rules used during later refinement. If no valid external seed can be constructed, CEDAR proceeds to LOCAL execution without sending a candidate representation to the external model.

C.3 VERIFICATION AND FAILURE DIAGNOSIS

For a candidate representation z_t, the external response model produces

$$y_t = F(q, z_t).$$

A trusted search-time verifier then evaluates (q, z_t, y_t, x). The verifier may inspect the full private scientific graph but does not receive the gold answer or the annotated gold support set.

Its structured decision contains three substantive checks—whether the task is answered, whether the answer is supported by the released context, and whether a critical scientific error is present—together with a failure type. A candidate is accepted only when

$$\text{TASKANSWERED} \wedge \text{RELEASEDCONTEXTSUPPORT} \wedge \neg\text{CRITICALERROR}.$$

Failure is categorized as MISSING-EVIDENCE, INSUFFICIENT-SPECIFICITY, UNSUPPORTED-CONCLUSION, or CONTRADICTORY-EVIDENCE; NONE is reserved for passing candidates.

When verification fails, a trusted diagnoser receives the private graph, current representation, external response, verifier feedback, and evidence contract. It proposes targeted disclosure repairs by identifying graph items that could address the observed failure. A repair may reveal a previously withheld node or relation, or increase the specificity of an already visible node. The diagnoser is not restricted to obligations explicitly anticipated by the initial contract: evidence outside $\Gamma(q, x)$ may be proposed when the observed failure indicates that it is necessary. However, every proposed item must refer to an existing node or relation in the frozen private graph.

The search-time verifier is an operational inference component rather than the evaluator used to report final GTS. All reported results use the independent post-hoc protocol in Appendix D; agreement between the two is audited in Appendix G.1.

C.4 POLICY-CONSTRAINED MONOTONE REFINEMENT

CEDAR permits only monotone increases in external disclosure. Scientific state follows

$$\textsc{Drop} \rightarrow \textsc{Abstract} \rightarrow \textsc{Exact},$$

subject to the levels permitted by P, while evidence relations follow

$$\textsc{Drop} \rightarrow \textsc{Keep}.$$

Previously released evidence is not retracted along a refinement path.

Every proposed repair δ is deterministically converted into a valid candidate through

$$z' = \mathrm{ProjectAndClose}(z, \delta, P).$$

Policy projection removes or limits disclosure operations that violate the active policy. Endpoint closure enforces graph consistency: a retained relation can be released only when both of its endpoint nodes are visible at an admissible specificity. Invalid states, duplicate states, and previously visited states are discarded before further evaluation.

The policy is never relaxed in response to a verification failure. In particular, if an exact protected value is prohibited under P_{strict}, repeated failure cannot promote that value to EXACT. When the information required for an externally grounded answer cannot be represented within the active policy, CEDAR changes the execution location rather than weakening the disclosure constraint.

C.5 BOUNDED SEARCH AND EXECUTION

Candidate representations are explored with bounded best-first search using the lexicographic disclosure cost $D(z)$ defined in Section 3. The frontier prioritizes the lowest-disclosure unexplored representation currently available. After a failed candidate is diagnosed, its policy-valid monotone refinements are inserted into the frontier and ordered by the same disclosure criterion.

The frozen main configuration permits at most 40 search-time verifier calls per case and a maximum refinement depth of four. Search terminates when the verifier accepts a candidate, when no valid refinement remains, or when either search limit is reached. A verified candidate is returned through CLOUD execution; otherwise, the unresolved task is retained for LOCAL execution. Exhausting the search budget never triggers relaxation of the active disclosure policy.

CEDAR therefore does not claim to solve the combinatorial disclosure problem to global optimality. The diagnoser proposes only a bounded subset of possible repairs, and the method returns the first verified state encountered by the bounded disclosure-ordered search. The resulting representation should therefore be interpreted as an operational approximation to the ideal minimum-disclosure objective rather than a certificate of global minimality.

All contract, verifier, and diagnoser outputs use structured schemas over the frozen graph-identifier inventory. Outputs are checked for schema validity, field types, and valid node and relation identifiers. When a structured output is malformed, CEDAR permits at most two format-only retries after the initial generation, for at most three total attempts. A retry may restate the frozen schema constraints and valid identifier inventory, but it cannot introduce scientific feedback, gold answers, annotated support sets, or evidence-selection hints. Outputs that remain invalid after the retry budget is exhausted are treated as unsuccessful rather than being manually corrected.

C.6 MODEL AND RUNTIME CONFIGURATION

All trusted language-model components of CEDAR—evidence-contract generation, search-time verification, failure diagnosis, and LOCAL execution—use `openai/gpt-oss-120b`. The model is executed inside the trusted environment on three NVIDIA A100 40 GB GPUs. Private scientific state used by these components therefore remains inside the trusted environment.

Trusted structured generation uses low reasoning effort. The local backend enforces role-specific minimum output budgets of 4,000 tokens for evidence-contract generation, 2,500 tokens for search-time verification, and 4,000 tokens for failure diagnosis. The backend uses a default random seed of 0; other sampling parameters are left at their backend defaults.

External candidate execution uses `gpt-5.4-mini-2026-03-17`. Each external call receives only the scientific question and the serialized candidate representation selected by the current search state, with a maximum output budget of 800 tokens. The external model does not receive the private graph, disclosure policy, evidence contract, verifier output, diagnoser output, search frontier, or gold evaluation metadata.

Model roles are held fixed across the main evaluation. The trusted local model constructs evidence requirements, verifies candidate sufficiency, diagnoses failures, and handles local execution, whereas the external model produces scientific responses from released representations. The independent post-hoc evaluator is not part of this inference loop and is described separately in Appendix D.

D Post-Hoc Evaluation

All reported main results are computed with a frozen post-hoc evaluation pipeline that is separate from CEDAR's search-time verifier. The evaluation depends on question type: Q1–Q3 use an independent LLM judge because the scientific answers are open-ended, whereas Q4–Q5 use deterministic, annotation-bound scorers. The same post-hoc criteria are applied to all methods after generation and are not used to define CEDAR's reported performance during search.

D.1 Q1–Q3 LLM Evaluation

Q1–Q3 require open-ended mechanism interpretation, control reasoning, and source-grounded experimental proposals. We therefore evaluate these questions with `gpt-5.6-terra` using a frozen structured judge with `reasoning_effort=low` and a maximum output budget of 2,500 tokens. For each evaluated response, the judge receives four inputs:

1. the scientific question;
2. the scientific context available for evaluation;
3. the generated candidate answer; and
4. the trusted source-grounded reference answer.

The evaluator produces three binary judgments:

- **Answer correctness** (A): whether the response gives the correct substantive scientific answer relative to the trusted reference;
- **Released-context support** (C): whether the scientific claims made in the response are supported by the context available to the response, rather than only by information present in the trusted reference; and
- **Critical error** (E): whether the response contains a material scientific error that invalidates or substantially alters the intended conclusion.

The frozen evaluator explicitly separates correctness from evidence support. The trusted reference is used to assess the substantive correctness of the answer, but information appearing only in the trusted reference cannot be counted as evidence supporting the response. In particular, the evaluator is instructed to judge support only from the scientific context provided with the candidate response.

The structured output schema contains `answer_preserved`, `context_support`, `critical_error`, `failure_type`, and `failure_description`. We map `answer_preserved` to A and retain the other two binary judgments directly. For Q1–Q3,

$$\mathrm{GTS} = A \wedge C \wedge \neg E. \tag{3}$$

The post-hoc evaluator is independent of the search-time verifier described in Section 4. Search-time verifier decisions, diagnoses, and acceptance labels are not supplied to the final evaluator. Their agreement is analyzed separately in Appendix G.1.

D.2 Q4–Q5 Deterministic Evaluation

Q4 and Q5 admit deterministic evaluation because the benchmark registers explicit quantitative inputs or an exact scientific target together with the evidence required to support that target. We therefore separate *answer correctness* from *evidence availability* without using an LLM judge.

Q4: quantitative comparison. Each Q4 instance specifies the source-grounded calculation inputs, the operation to be performed, the expected result, acceptable answer forms, the required state nodes, and the required evidence relations. The deterministic answer scorer extracts quantitative values from the generated response and compares them against the registered result and its acceptable alternatives using the frozen numerical normalization and tolerance rules. Unit representations are canonicalized before matching so that equivalent formatting does not by itself cause a failure.

Evidence support is evaluated independently from the answer text. A Q4 response is considered supported only when:

1. every annotated required node is visible in the evaluated context;
2. every registered calculation-input node is present at exact specificity with its required source value; and
3. every annotated required relation is retained.

Thus, a model that happens to produce the correct numerical answer without the required numerical evidence does not receive grounded success.

Q5: exact-state retrieval. Each Q5 instance specifies an exact-state target, its field type and expected value, required nodes, required relations, and acceptable answer variants. The deterministic answer scorer checks whether the response contains a normalized complete match to the registered exact state or an accepted equivalent form.

Evidence support requires that:

1. every required node is visible;
2. the registered target node is present at exact specificity;
3. the target value is recoverable from that exact node; and
4. every required relation is retained.

For both Q4 and Q5, the deterministic evaluator therefore separates obtaining the correct output from having the scientific evidence necessary to derive or retrieve that output. Let A_{det} denote deterministic answer correctness and C_{det} denote annotation-bound evidence support. We compute

$$\mathrm{GTS}_{Q4,Q5} = A_{\mathrm{det}} \wedge C_{\mathrm{det}}. \tag{4}$$

No separate LLM-based critical-error judgment is required for these two question types; the final evaluation records $E = 0$ and uses the deterministic answer and support checks directly.

D.3 Grounded Task Success and Failure Accounting

The final evaluation combines the two evaluation families into the common Grounded Task Success criterion introduced in Eq. 2. Table 9 summarizes the scoring procedure.

For CEDAR, the same grounded-success criterion is applied regardless of execution route. For a CLOUD case, evidence support is judged against the scientific representation released to the external model. For a LOCAL case, the unresolved task is executed inside the trusted environment using the private scientific context, and support is evaluated against the scientific context available to

Task	Answer evaluation	Evidence-support evaluation	Grounded success
Q1–Q3	Independent `gpt-5.6-terra` judgment against the trusted reference	Independent judgment using only the scientific context supplied for evaluation	$A \wedge C \wedge \neg E$
Q4	Deterministic match to the registered quantitative result	All required nodes visible, calculation inputs exact, and required relations retained	$A_{\text{det}} \wedge C_{\text{det}}$
Q5	Deterministic normalized match to the registered exact-state target	Required nodes visible, exact target recoverable, and required relations retained	$A_{\text{det}} \wedge C_{\text{det}}$

Table 9: **Frozen post-hoc evaluation protocol.** Q1–Q3 use an independent structured LLM evaluator, whereas Q4–Q5 use annotation-bound deterministic scoring.

that trusted execution. The private context used by LOCAL execution is not counted as external disclosure.

This distinction separates two quantities that are intentionally treated differently in our evaluation. GTS measures whether the final computation is scientifically correct and grounded in the evidence available at its execution location. Disclosure metrics instead measure only scientific information that crosses the trust boundary. Accordingly, a LOCAL result can receive GTS $= 1$ while having no terminal external release. Any candidate representations transmitted before the LOCAL routing decision are still counted in the search-wide disclosure audit in Appendix G.4.

All aggregate GTS results retain the full pre-specified evaluation population. A case receives GTS $= 1$ only when all success conditions for its question family are satisfied. Terminal execution or method failures are not removed from the denominator and receive GTS $= 0$. We do not impute successful answers for failed executions. Method-specific failure counts and conditional robustness analyses are reported in Appendix F.

D.4 HUMAN VALIDATION PROTOCOL

We additionally validate the frozen Q1–Q3 post-hoc evaluator against independent human judgments. This analysis is intended to assess the reliability of the automated A, C, and E judgments and is not used to modify evaluator prompts, benchmark annotations, or the reported main results. Q4–Q5 are excluded because their final scores are computed deterministically.

We construct a balanced validation set of 120 Q1–Q3 cases, with one case sampled from each combination of five evaluated methods (POLICY-MAX, UNIFORM ABSTRACTION, DM, ONE-SHOT, and CEDAR), three question types, two disclosure policies, and four scientific domains, yielding $5 \times 3 \times 2 \times 4 = 120$ cases. Stratified sampling uses the fixed seed 20260925. Cases without an annotatable candidate response, including terminal method failures that produced no answer, are excluded from the eligible pool and replaced within the same method–question–policy–domain stratum. The resulting sample contains 24 cases per method, 40 per question type, 60 per policy, and 30 per domain, with all 120 cases drawn from distinct workflows. For CEDAR and DM, we additionally verify that the context and answer shown to human annotators exactly match the corresponding inputs used by the canonical final post-hoc evaluator.

Two annotators independently evaluate every sampled case. To prevent the trusted reference from influencing evidence-support judgments, annotation is split into two stages.

Stage 1: released-context support. Annotators receive only the scientific question, the evaluation context, and the candidate answer. They assign `context_support` (YES/NO) according to whether the evaluation context itself contains sufficient scientific evidence to justify the substantive claims made in the answer. Annotators are explicitly instructed not to use outside knowledge to fill missing evidence. Thus, an answer may be scientifically plausible or correct while still receiving $C = 0$ if the supplied context does not support it.

	Human–Human		Evaluator–Human		
Judgment	**Agree.**	**κ**	**Agree.**	**Prec.**	**Rec. / F1**
Answer A	94.2	.846	86.7	.878	.956 / .915
Support C	86.7	.636	73.3	.985	.677 / .802
Critical error E	98.3	.825	80.8	.185	.833 / .303
GTS	91.7	.819	78.3	.952	.720 / .819

Table 10: **Human validation of the frozen Q1–Q3 post-hoc evaluator.** Human–Human columns report exact pre-adjudication agreement (%) and Cohen's κ. Evaluator–Human columns use the adjudicated human labels as reference and report exact agreement (%), positive-class precision, recall, and F1. The positive class is YES for A, C, and E, and successful grounded task completion for GTS.

Stage 2: answer correctness and critical error. Only after all Stage-1 judgments are completed and frozen do annotators receive the scientific question, candidate answer, and trusted source-grounded reference. They independently assign `answer_preserved` and `critical_error` using the same definitions as the frozen post-hoc evaluator. Paraphrases and shorter semantically equivalent answers are accepted for A, whereas $E = 1$ is reserved for scientifically material errors that change, invalidate, or contradict the intended conclusion.

The annotation files do not expose the evaluated method, disclosure policy, CEDAR execution route, or automated evaluator labels. Inter-annotator agreement is computed before adjudication. Cases on which the two annotators disagree are subsequently resolved by a third adjudicator using the same predefined stage-specific rubric. The third adjudicator is not shown the original annotators' labels or the automated evaluator decision. This requires adjudication of 16 support judgments, seven answer-correctness judgments, and two critical-error judgments, corresponding to 21 unique cases. Final human GTS is then derived mechanically as

$$\mathrm{GTS}_{\mathrm{human}} = A_{\mathrm{human}} \wedge C_{\mathrm{human}} \wedge \neg E_{\mathrm{human}}. \tag{5}$$

D.5 HUMAN VALIDATION RESULTS

Table 10 reports both pre-adjudication inter-annotator agreement and agreement between the frozen automated evaluator and the final adjudicated human labels.

After adjudication, humans assign $A = 1$ to 90 of 120 cases, $C = 1$ to 96 cases, and $E = 1$ to six cases, yielding 82 human-grounded successes (68.3%). Before adjudication, the two annotators already agree on 94.2% of answer-correctness judgments, 86.7% of support judgments, and 98.3% of critical-error judgments. Agreement on the derived GTS label is 91.7% with Cohen's $\kappa = .819$.

The frozen evaluator shows a systematic asymmetry relative to the adjudicated human labels. For released-context support, 65 of its 66 positive judgments are also judged supported by humans, corresponding to 98.5% precision. However, it rejects 31 additional cases that humans consider supported, resulting in lower support recall (67.7%). A similar pattern appears at the combined GTS level: 59 of 62 evaluator-positive cases agree with human GTS, giving 95.2% precision, while 23 additional human-success cases are rejected by the evaluator. Thus, disagreement on grounded success is predominantly in the conservative direction rather than from accepting responses that humans judge unsupported.

Critical-error judgments show the same conservative tendency more strongly. Humans identify six critical-error cases after adjudication, whereas the automated evaluator flags 27. It captures five of the six human-identified errors but additionally marks 22 cases as critical errors, explaining the high recall but low positive-class precision for E. Taken together, the human study indicates that the frozen evaluator does not obtain its reported GTS results by permissively accepting unsupported responses; its principal disagreement with humans is instead stricter rejection of otherwise acceptable cases.

E BASELINE IMPLEMENTATION DETAILS

This section specifies implementation choices needed to distinguish the baselines from CEDAR and to clarify which information is available to each method during release construction. All methods are evaluated on the same frozen workflow instances and disclosure policies. Unless stated otherwise, baseline construction does not use the gold evidence annotations employed by the post-hoc evaluator.

E.1 SHARED DISCLOSURE SPACE

The structured baselines operate over the same scientific workflow representation and policy specification as CEDAR. Scientific state uses the ordered levels

$$\textsc{Drop} < \textsc{Abstract} < \textsc{Exact},$$

subject to the levels permitted for each node by the active policy, and evidence relations use

$$\textsc{Drop} < \textsc{Keep}.$$

Policy validation, rendering, disclosure-cost computation, and endpoint closure are shared across methods. In particular, a retained relation is valid only when both endpoint nodes are visible, and no structured baseline may promote a node to a specificity prohibited by the active policy. Consequently, differences among CEDAR, One-Shot, and the structured search baseline arise from how a release is selected rather than from different representation languages or policy semantics.

Direct Masking is the only exception to the node-level abstraction mechanism: it is defined as a serialization-level masking baseline over the full scientific graph. It retains the surrounding scientific content and relation serialization while replacing registered protected exact values with a mask token. This baseline is included to distinguish direct-value suppression from query-conditioned evidence selection.

E.2 FIXED-RELEASE BASELINES

The four fixed-release baselines are deterministic and require no model-based release selection.

Policy-Max. Policy-Max constructs the maximally informative representation permitted by the active disclosure policy. Each node is assigned its highest admissible specificity and each relation is retained whenever permitted by endpoint closure. Under P_{open}, this approaches the full exact scientific graph; under P_{strict}, protected exact state remains prohibited.

Q-Only. Q-Only releases no private scientific context. The external response model receives only the scientific question. It therefore measures how much of the task can be answered without disclosure of the workflow-specific evidence.

Direct Masking. Direct Masking preserves the full serialized scientific context and all evidence relations but replaces registered protected exact values with `<MASKED>`. It performs no query-dependent selection and introduces no semantic abstraction of the remaining scientific evidence.

Uniform Abstraction. Uniform Abstraction maps scientific state to the policy-permitted abstract representation wherever such a representation is available and retains all relations permitted by endpoint closure. The same abstraction rule is applied independently of the question. It therefore tests whether uniform semantic generalization is sufficient without identifying which evidence or specificity is required by a particular task.

E.3 ONE-SHOT STRUCTURED DISCLOSURE

One-Shot tests whether query-conditioned disclosure can be selected directly in a single semantic decision, without CEDAR's verification-and-repair loop. It uses the same trusted local model as CEDAR, `openai/gpt-oss-120b`, and receives the scientific question, full private graph, active disclosure policy, and policy-valid disclosure levels.

Algorithm 2 One-Shot Structured Disclosure

Require: Question q, private graph $\mathcal{G}$, disclosure policy P
Ensure: Released representation z or method failure
$\mathcal{G}_P \leftarrow \text{ATTACHALLOWEDLEVELS}(\mathcal{G}, P)$
$\hat{s} \leftarrow \text{LLM}_{\text{local}}(q, \mathcal{G}_P)$
if $\hat{s}$ violates the required output schema **then**
    $\hat{s} \leftarrow \text{FORMATREPAIR}(q, \mathcal{G}_P, \hat{s})$
if $\hat{s}$ remains structurally invalid **then**
    return FAILURE
$\hat{s} \leftarrow \text{PROJECTANDCLOSE}(\hat{s}, P)$
$z \leftarrow \text{RENDER}(\hat{s})$
return z

The planner produces exactly one complete disclosure assignment. It does not receive CEDAR's evidence contract, an external candidate response, a sufficiency-verifier decision, failure diagnosis, downstream utility feedback, or post-hoc evaluation labels. The proposed assignment is deterministically projected onto the active policy and endpoint-closure constraints before rendering.

Schema or identifier errors may trigger format-only repair. Such repair provides only the structural violation needed to obtain a valid serialized assignment and does not provide information about answer quality, scientific evidence sufficiency, privacy performance, or which evidence should be added. If the proposal remains structurally invalid after the allowed repair procedure, the case is recorded as a method failure rather than being semantically regenerated using downstream feedback.

Only after the single disclosure assignment has been finalized is the released representation executed by the external response model. Thus, One-Shot isolates one-pass query-aware evidence selection from CEDAR's adaptive use of observed verification failures.

Algorithm 2 contains exactly one semantic disclosure decision. The optional repair corrects serialization or identifier errors only; it does not provide answer-quality, privacy, utility, or evidence-sufficiency feedback. The resulting representation is executed once by the external response model, with no semantic refinement or trusted-local fallback.

E.4 Adapted Data Minimization

We adapt response-preserving data minimization to the same structured scientific disclosure space used by CEDAR. The purpose of this baseline is to distinguish minimization subject to *output preservation* from CEDAR's requirement that the released representation itself contain sufficient scientific evidence for the answer.

DM does not use the trusted GPT-OSS-120B model, CEDAR's evidence contract, search-time verifier, failure diagnosis, released-context support criterion, or trusted-local fallback. Instead, it searches directly over policy-valid scientific-state representations using downstream answer preservation as its utility signal.

For each task, DM first generates a frozen reference response from the raw full-exact private scientific graph, independently of the active disclosure policy. Candidate representations are executed with the same external response model, and a structured utility judge determines whether the candidate response preserves the substantive answer of this reference response without a material contradiction or omission. The same response-preservation criterion is used for Q1–Q5. Gold answers, gold evidence annotations, released-context support, and the final GTS evaluator are not used during DM search. The raw full-exact reference call is used only to instantiate the response-preservation baseline objective; accordingly, the disclosure statistics reported for DM characterize its returned policy-valid release rather than end-to-end exposure incurred by constructing this reference.

The search coordinates are policy-protected scientific-state nodes whose active policy permits more than one disclosure level. Unprotected nodes remain at their policy-maximal representation. Relations are not independent search coordinates. Instead, they remain at their policy-maximal level whenever both endpoints are visible and are deterministically dropped when either endpoint is removed.

Algorithm 3 Adapted Data-Minimization Search

Require: Question q, private graph $\mathcal{G}$, policy P, response model F
Ensure: Policy-valid released representation z

$z_{\text{raw}} \leftarrow \textsc{RenderFullExact}(\mathcal{G})$
$y_{\text{ref}} \leftarrow F(q, z_{\text{raw}})$
$s_{\max} \leftarrow \textsc{PolicyMax}(\mathcal{G}, P)$
$\mathcal{V} \leftarrow$ searchable protected nodes
$\mathcal{F} \leftarrow \emptyset$
for $v \in \mathcal{V}$ **do**
    $s' \leftarrow \textsc{SetLevel}(s_{\max}, v, \textsc{Drop})$
    $s' \leftarrow \textsc{ProjectEdges}(s', P)$
    if $\textsc{Preserve}(F(q, \textsc{Render}(s')), y_{\text{ref}})$ **then**
        $\ell_v \leftarrow \textsc{Drop}$
    else
        $s' \leftarrow \textsc{SetLevel}(s_{\max}, v, \textsc{Abstract})$
        $s' \leftarrow \textsc{ProjectEdges}(s', P)$
        if $\textsc{Preserve}(F(q, \textsc{Render}(s')), y_{\text{ref}})$ **then**
            $\ell_v \leftarrow \textsc{Abstract}$
        else
            $\mathcal{F} \leftarrow \mathcal{F} \cup \{v\}$
$s_0 \leftarrow \textsc{ApplySelectedLevels}(s_{\max}, \{\ell_v\})$
$Q \leftarrow$ priority queue containing s_0, ordered by $D(s)$
$\mathcal{S} \leftarrow \emptyset$
while $Q \neq \emptyset$ **and** $|\mathcal{S}| < 512$ **do**
    $s \leftarrow \textsc{PopMin}(Q)$
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{s\}$
    if $\textsc{Preserve}(F(q, \textsc{Render}(s)), y_{\text{ref}})$ **then**
        return $\textsc{Render}(s)$
    for $v \in \mathcal{V} \setminus \mathcal{F}$ **do**
        $s' \leftarrow \textsc{IncreaseOneLevel}(s, v, P)$
        $s' \leftarrow \textsc{ProjectEdges}(s', P)$
        if s' is valid and unvisited **then**
            $\textsc{Push}(Q, s', D(s'))$
return $\textsc{Render}(s_{\max})$

The search proceeds in two stages. In Stage 1, each searchable protected node is independently probed from the policy-maximal state. DM first tests DROP; if answer preservation fails, it tests ABSTRACT. Nodes for which neither reduction preserves the reference response remain frozen at their policy-maximal level. The successful reductions are then combined to form the initial joint state for Stage 2.

Stage 2 performs a privacy-prioritized best-first search over this joint disclosure state. Candidate representations are ordered by the same lexicographic disclosure cost $D(z)$ used throughout the paper. At each expansion, one non-frozen protected node is increased to the next policy-permitted disclosure level, from DROP to ABSTRACT or from ABSTRACT to EXACT, and relation visibility is recomputed by deterministic endpoint projection. The first candidate whose external response preserves the frozen reference response is returned.

To bound computation on the larger joint disclosure lattice, Stage 2 expands at most 512 states per instance. If this budget is exhausted before an answer-preserving state is found, DM falls back to the policy-maximal policy-valid representation. The same fallback is used defensively if the search frontier is exhausted. DM never relaxes the active disclosure policy and never invokes trusted-local execution.

Algorithm 3 preserves the defining distinction between DM and CEDAR. DM may adaptively evaluate multiple candidate releases, but its search criterion asks only whether the model's substantive answer is preserved from the raw full-exact reference response. It does not test whether the released representation itself contains the scientific evidence needed to justify that answer. Final answer cor-

Method	Scope	Answer	Support	CritErr	GTS
DM	Overall	.876	.338	.173	.308
	P_{open}	.874	.453	.179	.419
	P_{strict}	.879	.224	.168	.196
One-Shot	Overall	.652	.449	.115	.428
	P_{open}	.831	.564	.111	.543
	P_{strict}	.473	.334	.119	.314

Table 11: **Component breakdown for the adaptive baselines.** Answer denotes final answer correctness, Support denotes support from the evaluation context, and CritErr is the critical-error rate. GTS requires a correct and supported answer without a critical error.

rectness, released-context support, critical errors, and GTS are therefore assessed only afterward by the common frozen post-hoc evaluation pipeline.

E.5 Common Execution and Failure Handling

All externally executed methods use the same response model, `gpt-5.4-mini-2026-03-17`. Fixed-release baselines execute their deterministically constructed context once. One-Shot executes its single projected release once. DM may execute multiple candidate releases while evaluating its search utility, but it does not receive CEDAR's grounded sufficiency signal.

No baseline is given CEDAR's trusted-local fallback as a recovery mechanism. A terminal One-Shot generation or validation failure is retained as a method failure. If DM exhausts its bounded Stage 2 search or search frontier without finding an answer-preserving representation, it returns the policy-maximal policy-valid representation rather than invoking trusted-local execution.

All final method outputs are subsequently evaluated with the same frozen post-hoc protocol described in Appendix D. Search-time utility decisions made by One-Shot or DM are not reused as final GTS labels. Cases with terminal execution or method failure remain in the original evaluation denominator and receive GTS $= 0$; no failed case is removed or replaced after evaluation.

F Additional Results and Statistical Analysis

F.1 Full Evaluation-Metric Breakdown

Table 11 decomposes the final GTS of the two model-based release baselines whose construction criteria do not directly optimize the frozen grounded-success metric. All quantities are computed by the same frozen post-hoc evaluators used for the main results. For Q4–Q5, the critical-error component is zero by construction under the deterministic evaluation protocol.

The decomposition clarifies why utility-preserving release construction without an explicit released-context support criterion is insufficient for scientific evidence minimization. DM retains high answer correctness under both policies, but context support drops from 0.453 under P_{open} to 0.224 under P_{strict}, with the corresponding GTS decreasing from 0.419 to 0.196. The same separation is visible, though less extreme, for One-Shot. Thus, a low-disclosure representation may still induce the correct substantive answer while failing to contain the scientific evidence needed to justify that answer.

F.2 Pairwise Statistical Comparisons

All confidence intervals in this section use 20,000 paired workflow-cluster bootstrap replicates with seed 20260923. Workflows, rather than individual policy–question cases, are resampled so that all evaluations derived from the same scientific workflow remain in the same bootstrap cluster. Differences are reported as the first method minus the second.

The complete comparison reinforces the regime-dependent interpretation in the main text. CEDAR and One-Shot are statistically indistinguishable in aggregate under P_{open}, whereas their difference becomes large under P_{strict}. The question-level contrasts further show that this advantage is con-

Comparison	Scope	Δ GTS	95% CI
CEDAR − One-Shot	Overall	$+.150$	$[+.119, +.181]$
	P_{open}	$-.004$	$[-.044, +.036]$
	P_{strict}	$+.304$	$[+.271, +.336]$
	Q1	$-.144$	$[-.234, -.056]$
	Q2	$-.016$	$[-.094, +.063]$
	Q3	$-.103$	$[-.166, -.044]$
	Q4	$+.531$	$[+.484, +.578]$
	Q5	$+.481$	$[+.431, +.531]$
One-Shot − DM	Overall	$+.121$	$[+.086, +.155]$
	P_{open}	$+.124$	$[+.078, +.170]$
	P_{strict}	$+.118$	$[+.084, +.150]$
CEDAR − DM	Overall	$+.271$	$[+.234, +.307]$
CEDAR − Policy-Max	Overall	$-.070$	$[-.098, -.042]$
	P_{open}	$-.315$	$[-.350, -.279]$
	P_{strict}	$+.175$	$[+.144, +.208]$

Table 12: **Paired GTS comparisons on the 160-workflow benchmark.** Intervals are obtained by paired workflow-cluster bootstrap.

	Q1	Q2	Q3	Q4	Q5
P_{open}	.412	.650	.256	.525	.850
P_{strict}	.438	.662	.263	.956	.769

Table 13: **CEDAR GTS by disclosure policy and question type.** Each cell contains 160 cases.

centrated on Q4–Q5 rather than being uniform across task types. In contrast, One-Shot exceeds DM under both policies, indicating that direct semantic release selection from the private scientific graph can outperform response-preserving minimization even without iterative verification.

F.3 Policy–Question and Domain Breakdown

The main table reports policy and question marginals separately. Their interaction is shown in Table 13. Each cell contains 160 policy–question cases.

For Q1–Q3, CEDAR's GTS changes only modestly between the two disclosure policies. The operating regime changes much more strongly for exact evidence-dependent tasks. In particular, Q4 reaches 0.956 GTS under P_{strict}, where the corresponding task is predominantly retained for trusted-local execution rather than forcing policy-prohibited state into the external release. The associated routing counts are reported separately in Appendix G.

Table 14 reports results by scientific domain. Each domain contains 40 workflows, corresponding to 400 policy–task cases. We include Policy-Max as the high-disclosure reference and DM as the answer-preserving adaptive baseline.

The CEDAR–DM difference remains positive in all four domains, ranging from $+0.255$ to $+0.290$ GTS. The gap between CEDAR and the policy-maximal reference also varies only moderately across domains. The main qualitative pattern is therefore not attributable to a single scientific discipline; the substantially larger variation occurs across disclosure policies and question requirements.

F.4 Baseline Failure Accounting

One-Shot is the baseline for which structured release generation can terminate without producing a valid representation. Among the 1,600 evaluation cases, 1,519 produce a valid One-Shot release and 81 terminate as method-level generation failures after the permitted format-only repair. These failures are retained in the end-to-end denominator and receive GTS $= 0$.

Domain	Policy-Max	DM	CEDAR	CEDAR − DM
Biology	.670	.303	.593	+.290
Chemistry	.610	.273	.543	+.270
Materials	.658	.293	.560	+.268
Physics	.655	.363	.618	+.255

Table 14: **End-to-end GTS by scientific domain.** Each domain contributes 40 workflows and 400 policy–task cases.

One-Shot quantity	Value
Total evaluation cases	1,600
Successful releases	1,519
Method-level failures	81
Failure rate	5.1%
End-to-end GTS	.428
GTS conditional on successful release	.451

Table 15: **One-Shot failure accounting.** Conditional GTS is reported only as a robustness diagnostic; the primary end-to-end result retains all 1,600 cases.

Removing these failed generations raises One-Shot GTS from 0.428 to 0.451, but does not eliminate the difference from CEDAR. The aggregate CEDAR–One-Shot comparison therefore cannot be explained solely by malformed One-Shot outputs. The primary results nevertheless use the full pre-specified population rather than the conditional subset.

The deterministic fixed-release baselines do not require model-based release construction. DM likewise returns a policy-valid representation when its bounded search terminates, using the policy-maximal state as the fallback rather than introducing a trusted-local execution path. Thus, search termination and method-level generation failure are kept distinct from the scientific correctness and evidence-support failures measured by the common post-hoc evaluator.

G CEDAR System Audits

This section audits four aspects of CEDAR that are not fully visible from terminal end-to-end scores: the relationship between the search-time verifier and the independent evaluator, the respective roles of verifier-gated recovery and iterative refinement, the routing behavior induced by different queries and disclosure policies, and the cumulative external disclosure created by adaptive search.

G.1 Search-Time Verifier Audit

CEDAR uses a trusted search-time verifier to decide whether the current released representation is sufficiently informative or whether additional scientific evidence should be considered. This verifier is part of the inference procedure and is distinct from the frozen post-hoc evaluator used to report all main-paper GTS results. In particular, search-time verifier outputs are never reused as final evaluation labels.

To characterize how closely the operational search signal approximates the independent grounded-sufficiency criterion, we compare the terminal search-time verifier decision against the frozen Q1–Q3 post-hoc evaluation on CEDAR cases that terminate with an externally evaluated candidate. Table 16 reports disagreement by question type.

The disagreement is strongly question-dependent. Q2 shows substantially closer correspondence between the operational verifier and the independent post-hoc criterion, whereas Q1 and especially Q3 are more difficult to certify during search. The support-specific disagreements follow the same pattern, indicating that much of the gap for Q1 and Q3 arises from whether the released representation contains sufficient evidence for the generated answer.

Question	Audited cases	GTS disagreement	Support disagreement
Q1	309	57.6%	50.5%
Q2	309	34.0%	33.0%
Q3	239	74.1%	71.1%

Table 16: **Search-time verifier audit on terminal externally accepted cases.** All audited cases were accepted by the search-time verifier. GTS disagreement reports the fraction for which the independent frozen post-hoc evaluator does not assign grounded success, while support disagreement isolates the released-context support component. These quantities characterize false acceptance among terminal externally accepted cases and do not measure false rejection over the full set of search candidates.

These results should not be interpreted as replacing the final evaluator with the search-time verifier. Rather, they quantify the gap between an online signal available without gold support annotations and the stricter evaluation used after generation. The main results therefore always use the independent post-hoc evaluator. Improving verifier calibration, particularly for open-ended relational and next-experiment reasoning, is a direct way to make the operational search more closely approximate the ideal GTS-constrained objective.

G.2 Execution-Control and Refinement Diagnostic

Section 6.4 separates two operational components of CEDAR: verifier-gated execution placement and iterative evidence refinement. We perform this diagnostic on a balanced 160-case subset containing one case per workflow, with 40 cases from each scientific domain, 32 from each question type, and 80 from each disclosure policy. The diagnostic subset is used only for mechanism analysis and does not replace the full 1,600-case main evaluation.

Gate/LOCAL control. To isolate the effect of execution gating from adaptive refinement, we augment the frozen One-Shot baseline with CEDAR's frozen search-time verifier. The verifier is applied exactly once to the One-Shot released representation and its corresponding external response. Cases accepted by the verifier retain the original One-Shot representation and response and terminate through CLOUD. Rejected cases are routed directly to trusted-local execution, with no diagnosis or semantic refinement. One-Shot proposals that fail to produce a valid external representation are likewise routed to LOCAL. Thus, this control adds CEDAR's verifier gate and local recovery while preserving One-Shot's single semantic disclosure decision.

The control routes 89 of 160 cases to CLOUD and 71 to LOCAL. Its end-to-end GTS is 0.706, compared with 0.438 for One-Shot on the same subset. Among the 89 verifier-accepted CLOUD cases, the independent frozen post-hoc evaluator assigns GTS 0.742. This comparison shows that verifier-gated execution placement accounts for a substantial part of the reliability gained over an ungated one-shot release.

Budgeted refinement. We next execute CEDAR independently with verifier-call budgets $B \in \{1, 2, 4, 8\}$. All model roles, prompts, disclosure policies, and post-hoc evaluation procedures are held fixed. These are independently executed budget-capped runs rather than nested prefixes of a single deterministic trajectory; cross-budget case transitions are therefore interpreted descriptively rather than as literal within-run recovery paths.

Table 17 reports the complete policy and question-type breakdown. Under the minimal $B = 1$ condition, all 160 cases terminate locally and GTS is 0.706. Increasing the budget raises verified CLOUD execution to 61.3% at $B = 2$, 78.7% at $B = 4$, and 77.5% at $B = 8$, while aggregate GTS remains between 0.681 and 0.694. Of the 47 cases unsuccessful under the $B = 1$ run, 15, 21, and 22 are successfully completed through CLOUD under the independently executed $B = 2$, $B = 4$, and $B = 8$ conditions, respectively.

The effect of refinement is therefore primarily visible in execution placement rather than in a monotonic increase in aggregate GTS. Relative to the Gate/LOCAL control, the $B = 4$ run increases verified external completion from 55.6% to 78.7%, while retaining comparable aggregate grounded

Condition	CLOUD	P_{open}	P_{strict}	Q1	Q2	Q3	Q4	Q5
Gate/LOCAL	.556	.738	.675	.719	.875	.375	.750	.813
CEDAR $B = 1$	.000	.738	.675	.594	.844	.313	.906	.875
CEDAR $B = 2$	.613	.688	.675	.625	.906	.375	.719	.781
CEDAR $B = 4$	.787	.725	.663	.781	.844	.375	.688	.781
CEDAR $B = 8$	.775	.675	.713	.813	.813	.375	.688	.781

Table 17: **Detailed execution-control diagnostic on the balanced 160-case subset.** CLOUD is the fraction of cases terminating with a verifier-accepted external representation. Remaining columns report end-to-end GTS by policy and question type. Budget conditions are independently executed runs.

Condition	CLOUD	Spec.	Chars
Gate/LOCAL	.556	7.854	716.9
CEDAR $B = 2$	.613	6.082	509.3
CEDAR $B = 4$	.787	7.246	599.5
CEDAR $B = 8$	.775	6.500	560.5

Table 18: **Release size conditional on verified CLOUD execution.** Spec. is mean raw specificity mass and Chars is mean serialized context length, computed only over cases terminating through CLOUD. These quantities characterize the mechanism diagnostic and are distinct from the normalized end-to-end and search-wide disclosure measures reported in Appendix G.4.

success. This distinction is consistent with CEDAR's design: the verifier determines whether the current released representation is sufficient, whereas refinement searches for additional policy-valid representations before the task is retained locally.

Release size of externally completed cases. The larger CLOUD fraction is not explained simply by larger released contexts. Table 18 reports raw specificity mass and serialized context length conditional on CLOUD termination. Compared with the Gate/LOCAL control, CEDAR at $B = 4$ externally completes substantially more cases while using a smaller average released representation per externally completed case.

Taken together, the diagnostic separates two complementary mechanisms. Verifier-gated routing prevents an insufficient external representation from being treated as the only available execution path, while iterative refinement increases the set of tasks for which a verifier-accepted, policy-compliant external representation can be found before trusted-local execution is required. The full-benchmark routing behavior is reported next, and Appendix G.4 separately audits whether the additional candidate evaluations increase cumulative external disclosure.

G.3 ROUTING AND EXECUTION BREAKDOWN

CEDAR can terminate in three ways. A CLOUD case terminates with a verified policy-compliant external representation and its corresponding response. A LOCAL case retains the unresolved computation inside the trusted environment when no sufficient external representation is recovered within the active policy and search budget. An execution failure is retained as an end-to-end failure and receives GTS 0.

Across the 1,600 main cases, CEDAR routes 1,173 cases to CLOUD, 418 to LOCAL, and records 9 execution failures. The policy split is shown in Table 19.

The aggregate shift under P_{strict} is concentrated in specific question types rather than being uniform across the benchmark. Table 20 gives the complete policy–question breakdown.

The contrast between Q1–Q3 and Q4–Q5 mirrors the evidence requirements identified in the controlled study. Under P_{strict}, only 2 of 160 Q4 cases and 1 of 160 Q5 cases remain CLOUD; the remaining 158 Q4 and 159 Q5 cases execute locally. Thus, when a question directly requires pro-

Policy	Cloud	Local	Failure
P_{open}	750	46	4
P_{strict}	423	372	5
Overall	1173	418	9

Table 19: **CEDAR routing by disclosure policy.** All execution failures remain failures in end-to-end evaluation.

	P_{open}			P_{strict}		
Question	**Cloud**	**Local**	**Fail**	**Cloud**	**Local**	**Fail**
Q1	156	4	0	153	7	0
Q2	157	3	0	152	8	0
Q3	124	32	4	115	40	5
Q4	155	5	0	2	158	0
Q5	158	2	0	1	159	0

Table 20: **CEDAR routing by policy and question type.** Reasoning-oriented Q1–Q3 remain predominantly external, whereas quantitative and exact-state tasks undergo an almost complete shift to trusted-local execution under the strict disclosure policy.

tected numerical or exact-state information, CEDAR preserves the external disclosure boundary by changing the execution location rather than relaxing the policy.

For completeness, the strict-policy final disclosure value reported in the main paper is an end-to-end external-disclosure quantity. Averaged over all 800 strict-policy cases, including LOCAL cases that terminate with no final external release, final specificity is 0.043. Conditional on the 423 cases that actually terminate CLOUD, final specificity is 0.081. The corresponding final node fractions are 0.076 end-to-end and 0.144 conditional on CLOUD. Protected-exact exposure is zero in both views.

These conditional values clarify the interpretation of the privacy result: the low end-to-end disclosure level reflects both selective external representations and policy-aware execution placement. It is not a measure of how much private state is internally available to trusted-local computation.

G.4 SEARCH-WIDE DISCLOSURE AND EXTERNAL-CALL ACCOUNTING

A terminal disclosure score alone can underestimate exposure for an adaptive method because CEDAR may transmit several candidate representations before returning a CLOUD answer or retaining the unresolved computation locally. We therefore reconstruct each available search trajectory and compute the union of all state and relation elements transmitted to the external response model.

Of the 1,600 main cases, complete exposure accounting is available for 1,591: 1,589 cases have reconstructed search traces and 2 are direct LOCAL cases with no external search. The remaining 9 execution failures lack a complete reconstructable trace and are excluded only from the search-wide exposure denominator; they remain failures in all end-to-end utility results. The policy-specific accounted denominators are 796 of 800 for P_{open} and 795 of 800 for P_{strict}.

Table 21 contrasts terminal release exposure with the union over all externally transmitted search candidates.

Under P_{strict}, accounting for every transmitted candidate increases the observed node fraction from 0.076 to 0.128, relation fraction from 0.046 to 0.072, and specificity from 0.043 to 0.071. Crucially, protected-exact exposure remains zero throughout the complete strict-policy search trajectory. Figure 4 visualizes the corresponding final-versus-search-wide operating point.

Iterative search also creates repeated transmission cost even when successive candidates overlap. Table 22 therefore reports the mean number of externally evaluated candidates and cumulative character transmission in addition to the union exposure.

Policy	Scope	Nodes	Exact	Relations	Specificity	Protected Exact	Chars
P_{open}	Final release	.116	.044	.065	.080	.058	.139
P_{open}	Search union	.127	.050	.071	.089	.063	.152
P_{strict}	Final release	.076	.009	.046	.043	.000	.094
P_{strict}	Search union	.128	.013	.072	.071	.000	.161

Table 21: **Final versus search-wide external disclosure.** Final-release values use all 800 cases per policy. Search-union values use the 796 open-policy and 795 strict-policy cases with complete exposure accounting. All quantities are normalized to the corresponding full scientific graph or full-context serialization.

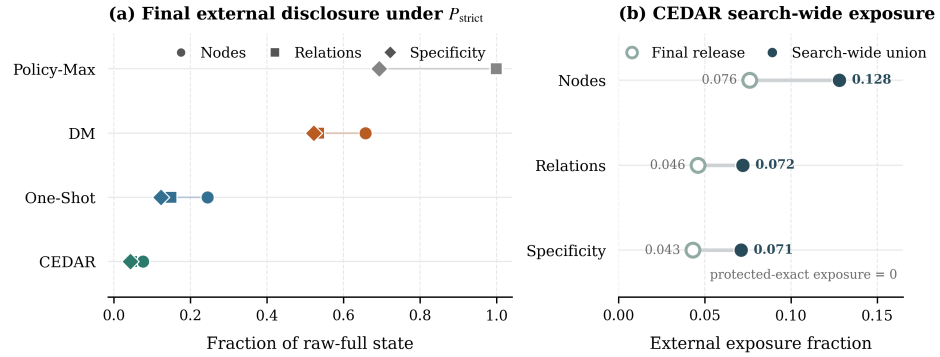


Figure 4: **External disclosure under P_{strict}.** (a) Final release; (b) union of CEDAR candidates transmitted during search for cases with complete traces.

Policy	Cases	Mean calls	Cumulative chars
P_{open}	796	2.62	.284
P_{strict}	795	2.67	.301
Overall	1591	2.64	.293

Table 22: **External-call accounting across CEDAR search.** Cumulative character exposure sums the normalized character mass transmitted across successive external candidate evaluations and therefore counts repeated transmission, unlike the search-union measure.

The distinction between terminal, union, and cumulative exposure is important. Final release measures what remains externally visible at termination; search union measures which scientific information was exposed at least once during the adaptive trajectory; cumulative transmission additionally captures repeated external communication. Reporting all three prevents the privacy benefit of adaptive search from being attributed solely to a small terminal representation.

These audits reveal two distinct operating regimes. Reasoning-oriented tasks are usually resolved externally from selective evidence, whereas strict quantitative and exact-state tasks are predominantly retained within the trusted environment. Importantly, the resulting reduction in external disclosure persists even when intermediate candidates and repeated external calls are included in the accounting.