Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evidence-aware retrieval and bounded orchestration for heterogeneous materials R&D

Seifallah Elfetni

seifallah.elfetni@ctc-germany.org

Center for the Transformation of Chemistry https://orcid.org/0000-0003-3615-627X

Theodore Tyrikos-Ergas

theodore.tyrikos-ergas@ctc-germany.org

Center for the Transformation of Chemistry

Research Article

Keywords: materials informatics, retrieval-augmented generation, evidence sufficiency, additive manufacturing, revision authority, numerical provenance

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11746659/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Evidence-aware retrieval and bounded orchestration for heterogeneous materials R&D

Seifallah Elfetni[1,*], Theodore Tyrikos-Ergas[1,*]

[1]Center for the Transformation of Chemistry, Puschstr. 6b, 04103 Leipzig, Germany

[*]Corresponding authors:
seifallah.elfetni@ctc-germany.org
theodore.tyrikos-ergas@ctc-germany.org

Abstract

Materials research and development (R&D) questions often require evidence distributed across measurements, process histories, and revision-controlled investigations. Relevance ranking alone cannot determine whether that evidence is sufficient, current, or numerically interpretable. We present an evidence-aware retrieval methodology with three elements: provenance-preserving representations, a question-specific evidence ledger for bounded orchestration, and layered evaluation of retrieval, claim admission, and response behavior. An IN718 additive-manufacturing testbed combines public sources from the National Aeronautics and Space Administration (NASA) and the National Institute of Standards and Technology (NIST) with an authored benchmark of 60 questions in 30 paired intents. Hybrid retrieval increased the top-five hit rate from 0.7885 to 0.9231, whereas reranking the Hybrid candidates reduced it to 0.8654: component effects are non-monotonic. Evidence-conditioned control exposed unresolved requirements but did not guarantee closure: every integrated case with multiple retrievals retained a residual. The fixed integrated system, executed once on 24 evaluation questions with prior component-level exposure, answered one abstention-labeled question and withheld three answerable ones, while all emitted citation identifiers were valid and all rule-rejected claims were excluded. Revision-controlled investigation examples and numerical probes show how historical authority, recorded event status, and calculation provenance constrain what an answer can establish. The methodology is inspectable: strong retrieval scores do not conceal response or semantic failures. The labels are authored, not expert-adjudicated; independent adjudication and an untouched evaluation cohort are the next validation axes.

Keywords: materials informatics; retrieval-augmented generation; evidence sufficiency; additive manufacturing; revision authority; numerical provenance

1 Introduction

A fatigue result is rarely sufficient to explain a materials-processing problem. An engineer may also need the specimen condition, the loading definition, the processing history, fracture observations, and the current interpretation of an investigation. These records answer different questions. A test report establishes an observation; a processing record establishes a recorded condition; a root-cause analysis interprets a collection of observations. Even when the records concern the same specimen or build, their information cannot be substituted indiscriminately. A preliminary hypothesis can remain relevant to the investigation after a later revision has superseded it, while a precisely recorded runout count still does not identify a fatigue-failure time. The information problem is therefore not simply to find related text, but to determine what the available records can support.

Materials informatics has developed several complementary routes for making scientific information reusable. Structured repositories such as the Materials Project, NOMAD, and the Materials Data Facility organize computational and experimental materials data, while the FAIR principles articulate requirements for finding, accessing, integrating, and reusing research objects [1–5]. The Materials Genome Initiative and materials data science have framed this reuse problem as a shared infrastructure challenge [6, 7], and additive-manufacturing benchmark initiatives similarly emphasize connecting measurements with the conditions under which they were obtained [8]. The emphasis is warranted for the alloy studied here: the mechanical and fatigue behavior of additively manufactured IN718 depends on defect population, surface condition, and post-processing, so a reported endpoint is interpretable only with its processing and loading context [9, 10]. The information required during a specific research and development (R&D) investigation, however, may also include local reports, test logs, revisions, and records that were never written as database entries. A question-answering interface must connect those sources without discarding the distinctions that make them scientifically interpretable.

Automated extraction addresses the conversion of published information into structured records. Chemistry-aware text mining, extraction of synthesis recipes, and embeddings learned from materials literature have demonstrated different ways to recover information that would otherwise remain difficult to search [11–13], and large language models now extract structured entity, property, and condition records directly from scientific text [14, 15]. Recent materials studies make the importance of representation particularly clear. Circi et al. evaluated polymer-

composite table extraction across image and text representations and distinguished recognition of property names from recovery of their values, units, and conditions [16]. Liu et al. used expert-grounded prompt optimization for high-entropy-alloy lattice-constant extraction and identified errors involving entity attribution, scientific terminology, and unit conversion [17]. Immanuel and Mahata connected alloy-property extraction, structured tables, natural-language querying, and downstream prediction within a reusable workflow [18]. Together, these studies show why a materials record must carry more than a number or a passage: its subject, conditions, and provenance are part of the information being extracted.

A further distinction separates stored observations from analytical results. Integrated computational materials engineering (ICME) links models, data, and experiments across disciplines and places explicit weight on the verification and validation of its tools [19, 20]. The compound knowledge graph of Aggour et al. connects factual, analytical, and expert knowledge and can assemble calculations when an answer is not directly stored [21]. In ICME decision support, Digonta et al. link predictive models and uncertainty-aware design formulations across interacting disciplines [22]. For retrieval-based assistance, the corresponding requirement is to preserve the difference between quoting a source-reported value and calculating a new quantity from selected records. Both can support an answer, but their provenance and admissibility conditions differ.

Retrieval-augmented generation (RAG) provides a general mechanism for conditioning language-model outputs on external information [23]. Its retrieval components act on different aspects of relevance. Lexical ranking retains exact terms and identifiers; dense retrieval uses learned representations; reciprocal-rank fusion combines ranked lists without requiring a common score scale; and cross-encoders rescore query–passage pairs [24–27]. Sparse and dense signals capture complementary evidence, and dense retrievers often require interpolation with a lexical score to reach their best passage-level effectiveness [28, 29]. Query rewriting and decomposition additionally change the requests presented to a retriever [30, 31]. Relevance ranking alone, however, does not determine whether a scientific question has been answered completely. A highly ranked result may supply the cycle count but omit the stress definition. A query transformation may discover a needed passage that is subsequently displaced from the short evidence window used for answering. Evaluating the assembled pipeline therefore requires both discovery and evidence-completeness measurements.

Conditional retrieval moves part of that design problem into the execution of a question. Adaptive-RAG uses question complexity to select a retrieval strategy, whereas corrective and self-reflective approaches incorporate assessments of retrieved information or generated content [32–34]. Tool-using language models extend conditional behavior to sequences of actions [35]. Scientific systems such as Coscientist, ChemCrow, and LLaMP connect such actions to domain resources, with LLaMP organizing materials requests through a supervisor and specialized retrieval or simulation capabilities [36–38]; retrieval-augmented and agent-based assistants have also been applied to materials literature and materials design [39, 40]. The issue considered here is narrower: how should subsequent retrieval depend on the evidence already available for a materials-R&D question, and how should unresolved requirements remain visible when further search does not close them?

Reliability also depends on the level at which it is measured. Citation-oriented evaluation, attribution measurement, atomic factual-support assessment, and automated RAG evaluation address related but distinct properties of an answer [41–44], and RAG benchmarks treat the rejection of unsupported questions as a capability separate from answer accuracy [45]. In a graphene-synthesis study, Cho et al. found that automated evaluators differed in their alignment with expert judgments and in their sensitivity to retrieval augmentation [46]. A system can consequently retrieve relevant information, cite valid identifiers, and still choose an inappropriate response mode or omit a supported premise correction. Conversely, withholding an unsupported number can be appropriate even when retrieval has located other useful context.

We investigate this problem through an evidence-aware methodology for heterogeneous IN718 additive-manufacturing R&D records. Its contribution has three parts. First, a provenance-preserving representation separates narrative evidence, typed numerical records, revision information, and derived calculations. Second, a question-specific evidence ledger makes support requirements explicit and supplies the state used by a bounded retrieval controller. Third, a layered evaluation separates ranking, assessed sufficiency, rule-based claim admission, response behavior, and numerical semantics. The experimental record combines a comparison of six retrieval configurations, question-conditioned and evidence-conditioned control policies, a fixed integrated-system evaluation, and separate numerical probes. Public sources from the National Aeronautics and Space Administration (NASA) and the National Institute of Standards and Technology (NIST) anchor the data environment, while authored internal-style records provide controlled revision, missing-evidence, and misleading-evidence cases. The study tests an inspectable form of materials information assistance; its central object is the relationship between evidence and response, not autonomous engineering authority.

2 Methods

2.1 Materials evidence and canonical representation

The experimental corpus contains 277 retrieval units drawn from NASA publications, NIST IN718 data, and a fictional internal R&D investigation. The public sources

contribute 217 units and the authored internal records contribute 60. Across representations, there are 146 narrative chunks and 131 structured units: 70 experiment summaries, one process record, four analysis records, and 56 test-log records. Documents are counted through their chunks, not as additional entries. Table 1 summarizes the evidence and evaluation populations.

The internal corpus comprises 12 linked records centered on builds B017 and B021. It includes material and process specifications, build and heat-treatment records, fatigue reports, written fracture observations, a nonconformance report, and two root-cause-analysis (RCA) revisions. All B017/B021 measurements and conclusions are authored fiction. Their purpose is to define controlled information relationships, including a preliminary hypothesis superseded by later evidence, without presenting synthetic values as new materials experiments. The public-source records and the internal investigation retain separate provenance.

Each retrieval unit has a stable identifier, a retrieval-facing text representation, a source reference, and type-specific metadata. We write this as $u_i = (\mathrm{id}_i, x_i, p_i, m_i)$, where x_i is the text used for discovery, p_i identifies its source, and m_i contains the available record and revision attributes. Numerical records remain accessible in typed form. Their retrieval text and embeddings help locate the record but do not replace its numerical fields. This distinction is consequential for linked data: the cycle count of a test-log object can describe its number of signal rows, whereas the fatigue endpoint resides in the associated experiment summary.

Fig. 1 shows these representations alongside the executed information pathways. Retrieval produces source evidence for a question-specific ledger. Registered numerical operations produce derived receipts in a separate capability evaluation. The final response pathway uses model-guided retrieval and assessment followed by deterministic claim construction, admission, and assembly.

2.2 Questions, experimental populations, and exposure

The controlled benchmark contains 60 questions arranged as 30 paired intents. The two formulations of an intent remain in the same split: 36 questions in 18 intents constitute the development set, and 24 questions in 12 intents constitute the integrated-system evaluation set (labeled "challenge" in the figures). Annotations specify relevant evidence identifiers, reference requirements, answerability, and 26 hard-negative pairs. The labels were authored for this corpus and have not been independently adjudicated by materials experts.

Three response categories distinguish direct answers, premise corrections, and abstentions. A premise-correction question is answerable because a useful response can establish that a proposed interpretation is unsupported or needs qualification. An abstention question requests information absent from the authored record; treating abstention as a scored response outcome follows selective question answering, in which declining to answer is a measured decision rather than a failure to respond [47]. Development contains 28 direct-answer and eight premise-correction questions; the integrated evaluation contains four direct-answer, 12 premise-correction, and eight abstention questions. Ranking metrics exclude the eight abstention questions, leaving 52 eligible questions in 26 intents overall and 16 questions in eight intents for the integrated evaluation. The different compositions preclude treating development-to-evaluation score differences as improvement on a common population.

The integrated configuration was fixed after development and executed once on the 24 evaluation questions. These question texts had already appeared in earlier retrieval, query-transformation, routing, and corrective experiments. The evaluation therefore tests the fixed integrated system on its designated evaluation partition, not on a globally unseen question set. At execution, model roles received only their permitted question and evidence inputs; relevance labels, reference answers, hard-negative annotations, split labels, and offline metrics were excluded. Runtime outputs were fixed before their offline scoring.

Eight public-source sanity questions and eight numerical development probes provide additional, separate evaluations. The former address NASA/NIST evidence; the latter test predefined typed numerical requests. Neither contributes observations to the 60-question benchmark.

2.3 Retrieval configurations

Dense retrieval uses the recorded BAAI/bge-base-en-v1.5 model, 768-dimensional embeddings, and cosine ranking [48]. Hybrid retrieval combines dense and BM25 rankings. BM25 uses $k_1 = 1.5$ and $b = 0.75$, and each branch supplies up to 50 candidates [24]. Reciprocal-rank fusion (RRF) assigns

$$s(u_i \mid q) = \sum_{\ell:\, u_i \in \ell} \frac{1}{60 + \mathrm{rank}_\ell(u_i)},$$

where ℓ indexes the dense and lexical lists, ranks begin at one, and an absent unit contributes zero [26].

The cross-encoder (CE) configuration uses BAAI/bge-reranker-base to rescore query–evidence pairs. Dense + CE reranks 20 dense candidates, whereas Hybrid + CE reranks 50 fused candidates. These experiments compare the recorded pipelines: candidate depth and candidate composition differ between the two reranker applications and are not isolated.

Multi-query retrieval uses the original question and exactly three question-only reformulations, applies Hybrid retrieval to each, and combines their lists through a second RRF stage. Decomposition identifies whether a question has separable evidence needs. Atomic questions retain the original Hybrid query; decomposable questions use the original plus generated subqueries before fusion. The

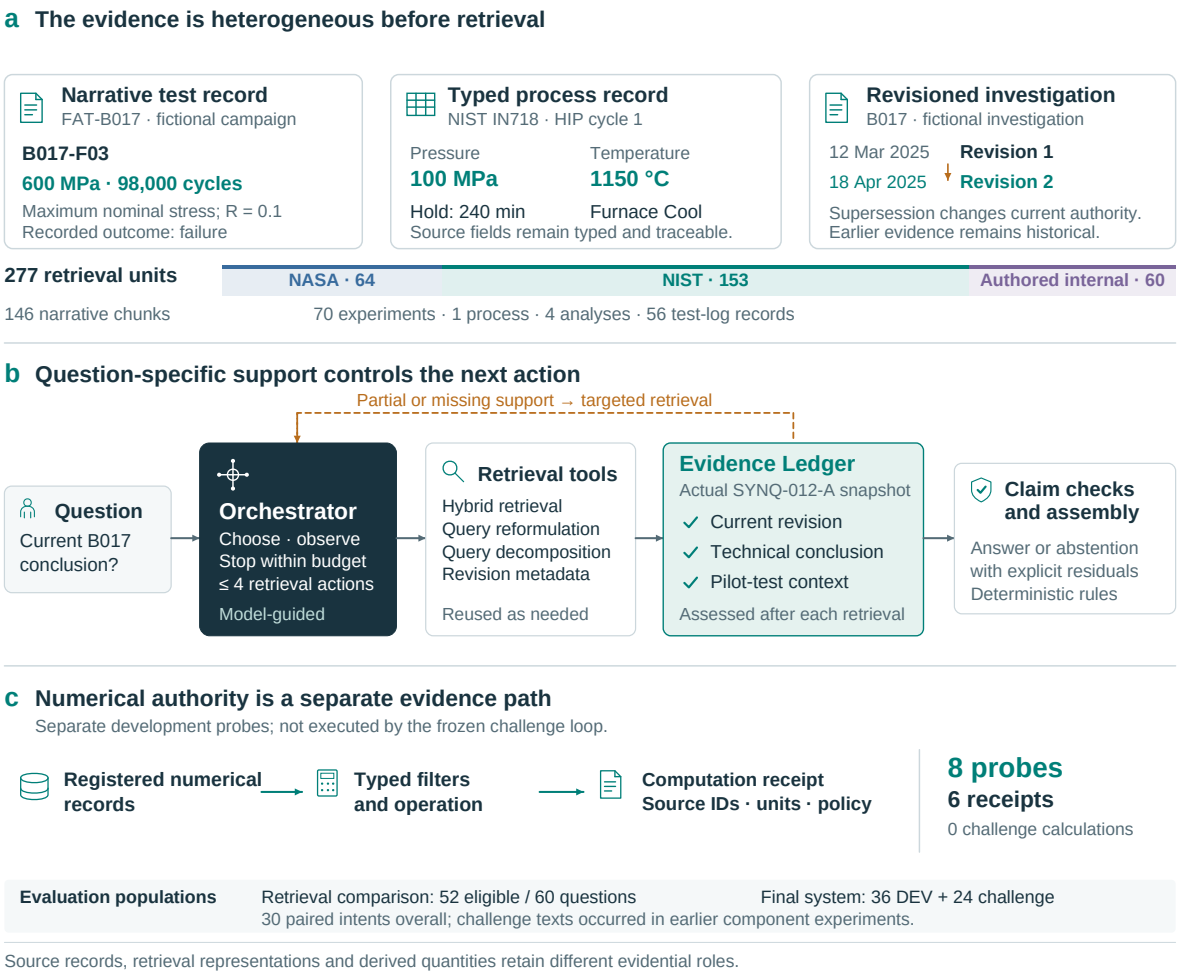


Fig. 1: Evidence-aware materials information architecture. (a) Narrative, typed numerical, and revision-controlled evidence examples from the corpus, with unit counts by representation (146 narrative chunks, 131 structured units) and by partition (NASA 64, NIST 153, authored internal 60). (b) The bounded model-guided controller, the question-specific evidence ledger (the SYNQ-012-A snapshot after Hybrid retrieval), and the deterministic claim checks and assembly; the feedback branch is conditional. (c) The separate numerical evidence path from registered records through typed operations to computation receipts (eight development probes, six receipts). DEV, development set.

Table 1: Evidence environment and evaluation design. B017/B021 records are fictional; labels are not expert-adjudicated. Corpus rows describe representations, not mutually exclusive document counts.

Element	Scale	Composition or interpretation
Evidence corpus	277 retrieval units	217 public-source; 60 authored internal units
Narrative representation	146 chunks	86 public-source; 60 authored internal chunks
Structured representation	131 units	70 experiment; 1 process; 4 analysis; 56 test-log records
Controlled questions	60 questions; 30 paired intents	32 direct-answer; 20 premise-correction; 8 abstention labels
Development partition	36 questions; 18 intents	28 direct-answer; 8 premise-correction
Integrated evaluation	24 questions; 12 intents	4 direct-answer; 12 premise-correction; 8 abstention
Ranking-eligible questions	52 overall; 16 integrated	26 overall and 8 integrated intent clusters
Misleading-evidence stressor	26 hard-negative pairs	Authored misleading evidence linked to benchmark questions
Separate evaluations	8 public questions; 8 numerical probes	Excluded from the controlled benchmark denominators

fixed-method comparisons reuse the recorded transformations. In the integrated controller, selected transformation actions instead generate their queries during execution. These are local implementations of query transformation, not reproductions of the cited systems [30, 31].

2.4 Evidence requirements and bounded orchestration

An assessment describes the support needed for a question as requirements. At retrieval step t, requirement $r_{j,t} = (d_{j,t}, s_{j,t}, I_{j,t})$ contains a description, an assessed status, and supporting identifiers. The ledger L_t combines these requirements with overall sufficiency, ordered selected

identifiers, and a residual summary. Supported, partial, and missing states refer to the supplied evidence pool. Requirements can be reformulated on reassessment; a reused requirement number does not establish semantic identity across steps.

The question-conditioned router chooses among Hybrid, Multi-query, and Decomposition using question text alone. Its recorded output supplies the integrated controller with an initial suggestion, not a compulsory first action. The fixed corrective policy assesses the initial evidence and permits one alternate strategy when the bundle is judged insufficient. Its correction uses preserved strategy outputs. The bounded controller instead receives the question q, ledger L_t, observable history H_t, and remaining budgets B_t, and proposes

$$a_t = \pi(q, L_t, H_t, B_t).$$

The policy π is a model-guided proposal under a fixed prompt and typed action schema. Permitted executable actions are Hybrid search, Multi-query search, Decomposition search, and revision-metadata lookup, together with stopping proposals. Search calls return up to ten contexts for assessment. A retrieval updates the ordered pool by stable identifier and obtains a new assessment:

$$E_{t+1} = \text{deduplicate}_{\text{id}}(E_t \parallel T(a_t)),$$
$$L_{t+1} = \text{assess}(q, E_{t+1}).$$

Here $\parallel$ denotes concatenation: first-seen evidence remains ordered, and new results are appended in tool-rank order. Revision lookup supplies an observation to the history; in the evaluated loop it does not add a retrieval unit or trigger reassessment on its own.

The controller is limited to six decision steps, four retrieval actions, and two query-transform calls. A transformation requested after its allowance is exhausted falls back to Hybrid with the same query. Exact normalized duplicate actions are rejected. Retrieval stops on a sufficient ledger, an explicit finish proposal, or exhausted decision or retrieval capacity. Final evidence selection retains at most five ledger-selected units. Model-guided action selection, assessment, and dynamic transformation used the recorded gpt-5.5 low-effort configuration; ranking mechanics, validation, pool merging, and budget enforcement were deterministic given those outputs. Algorithm 1 summarizes the executed control and response sequence.

2.5 Numerical operations and claim admission

The numerical interface accepts a registered dataset, validated filters, a permitted operation, numerical fields, and any grouping and unit or censoring information. Filtering, descriptive summaries, and explicit group comparisons are supported. A successful calculation returns a computation receipt containing the normalized request, source identifiers, sample counts, result, units, warnings, and implementation provenance. No arbitrary executable expression is supplied by a model. Unsupported operations, incompatible inputs, and zero-denominator comparisons are handled as structured errors.

The eight numerical probes use predefined requests, not calls to a Scientific Analyst model. Scientific analysis appeared in the broader action schema but was not executable in the generic loop evaluated here, and no calculation was invoked by the integrated-system evaluation. The numerical results therefore establish a separately tested interface, not integrated autonomous selection of a calculation.

After retrieval, deterministic functions construct at most eight candidate claims from ledger requirements and short selected-source excerpts. A keyword-based taxonomy distinguishes numerical, revision, limitation, and other claim categories. Admission requires supplied evidence identifiers for non-limitation claims, a computation receipt for numerical-classified claims, and appropriate currentness for a current or approved revision claim. Limitations are admitted from the residual ledger state. Other claims can pass the rule set through evidence membership without a separate entailment assessment. Accordingly, *supported* denotes an internal admission disposition, not an independent scientific judgment.

The assembler emits at most five admitted non-limitation claims with citations. It withholds an answer when no such claim remains, and otherwise includes available residual statements. This creates three distinct endpoints: assessed retrieval sufficiency, admission disposition, and response mode. Valid citations are checked for membership in the supplied and admitted evidence. The final checker is not a separate language-model critic, and the rule requiring a receipt does not distinguish every source-reported number from newly calculated arithmetic.

2.6 Evaluation and statistical analysis

For eligible question q, let G_q be its annotated relevant-evidence set and $R_q^{(k)}$ the first k ranked or selected units. Hit@k is $\mathbf{1}[G_q \cap R_q^{(k)} \neq \varnothing]$, while Recall@$k$ is $|G_q \cap R_q^{(k)}|/|G_q|$. Mean reciprocal rank (MRR) averages the reciprocal rank of the first relevant item. Binary normalized discounted cumulative gain at rank 10 (nDCG@10) weights a relevant item at rank r by $1/\log_2(r+1)$ and normalizes by the ideal ordering. All scores are macro-averaged over eligible questions. Missing items beyond the saved candidate depth are not assigned invented ranks. Assessor-selected evidence and unassessed retrieval rankings are reported separately because selection changes the information and objective of the ordering step.

Uncertainty intervals were computed post hoc from the preserved question-level results, using 10,000 intent-cluster bootstrap replicates and seed 20260919. Both formulations of a sampled intent remain together, and paired method differences use the same resampled clusters. The

Algorithm 1. Bounded retrieval and response construction.

Inputs: question, recorded route suggestion, registered retrieval/revision tools, and allowances of six controller decisions, four retrievals, and two transformations.

Outputs: selected evidence, assessed requirements, admitted claims, assembled response or abstention, and action history.

1. Initialize the evidence pool, ledger, action history, and normalized-signature set.
2. At each remaining decision step, stop if the ledger is sufficient; otherwise request and validate a controller action.
3. Honor a finish proposal. If a requested retrieval would exceed its allowance, stop acquisition with a residual.
4. Replace a transformation beyond its allowance with Hybrid retrieval on the same query; reject an exact normalized duplicate action.
5. Execute the permitted action and append its observation to the history.
6. For retrieval, merge returned units by stable identifier and reassess the pool. For revision lookup, retain the observation without a new assessment.
7. On termination or decision-budget exhaustion, retain at most five ledger-selected evidence identifiers.
8. Construct at most eight candidate claims and apply the membership, numerical-receipt, and revision-admission rules.
9. Assemble at most five admitted non-limitation claims and available residual text; withhold the answer when none remains. Validate emitted identifier membership.

The controller, assessor, and selected query transformations are model-guided; execution checks, merging, revision lookup, claim construction and admission, and assembly are deterministic.

analysis covers five contrast families: Dense + CE minus Dense, Hybrid minus Dense, Hybrid + CE minus Hybrid, Multi-query minus Hybrid, and Decomposition minus Hybrid. Reported 95% percentile intervals are pointwise, not simultaneous. They describe variation across the sampled intents conditional on the recorded outputs and annotations; repeated model-generation variability is not estimated.

Integrated-system response behavior is compared with the authored answerability labels, counting premise corrections as answerable. We report the raw answer/withhold confusion matrix, claim dispositions, emitted citation membership, hard-negative selection, and residual states. These response-mode labels do not grade the completeness or truth of each assembled answer. The additional lexical reference-token evaluator is retained as an implementation diagnostic rather than used as a scientific-correctness endpoint. Post hoc trace analysis distinguishes newly retrieved identifiers that support final ledger requirements from identifiers actually emitted as new answer citations. Failure localization uses observable records from all four response-mode disagreements and all nine multi-retrieval cases.

3 Results

3.1 Retrieval composition has non-monotonic effects on evidence coverage

Hybrid retrieval increased top-five evidence coverage relative to Dense retrieval on the 52 eligible questions (Fig. 2; Table 2). Hit@5 rose from 0.7885 to 0.9231, corresponding to a relevant item appearing in the first five positions for 48 rather than 41 questions. Recall@5 increased from 0.5625 to 0.6635. The paired differences were 0.1346, with a 95% intent-cluster interval of [0.0385, 0.2500], for Hit@5 and 0.1010 [0.0369, 0.1683] for Recall@5.

The ranking metrics show why those two improvements should not be reduced to a single score. Hybrid increased nDCG@10 from 0.5665 to 0.6543, with paired difference 0.0878 [0.0244, 0.1534]. MRR increased from 0.6402 to 0.6894, but its paired interval included zero. Cross-encoder reranking produced a different pattern. Applied to Dense candidates, it raised the four reported point estimates, including Hit@5 to 0.8846. Applied after Hybrid fusion, the same reranker reduced Hit@5 to 0.8654, Recall@5 to 0.5962, and nDCG@10 to 0.5789. The Hybrid + CE minus Hybrid nDCG contrast was −0.0754 [−0.1417, −0.0011]. This is a pipeline-dependent result under the differing candidate depths specified in Methods, not evidence for an intrinsic ordering of reranking methods.

Query transformation also changed the balance between early coverage and rank quality. Multi-query achieved the highest MRR and nDCG@10 point estimates among the six configurations, 0.7121 and 0.6711, while its Hit@5 and Recall@5 remained below Hybrid. Its paired differences from Hybrid included zero for all four displayed metrics. Decomposition similarly combined an MRR of 0.7008 with lower top-five coverage than Hybrid. The Recall@k curves in Fig. 2b show the broader discovery context for these outcomes: finding evidence at a later position and retaining it inside the answering window are different achievements.

The specimen question in Fig. 2d makes that distinction concrete. A complete response about B017-F03 needs both the recorded fatigue result and the section defining the reported stress. Dense ranks the result third and the conditions twelfth. Dense + CE places them second and fourth, bringing both into the top five. Hybrid ranks the result first but the conditions sixth; Hybrid + CE moves them to third and tenth. This example was selected

Table 2: Retrieval quality of the six fixed configurations, macro-averaged over 52 eligible questions in 26 paired intents (abstention-labeled questions excluded). CE denotes cross-encoder reranking.

Configuration	Hit@5	Recall@5	MRR	nDCG@10
Dense	.7885	.5625	.6402	.5665
Dense + CE	.8846	.6154	.6837	.6030
Hybrid	.9231	.6635	.6894	.6543
Hybrid + CE	.8654	.5962	.6539	.5789
Multi-query	.8654	.6442	.7121	.6711
Decomposition	.8846	.6234	.7008	.6447

descriptively, but its two evidence requirements are fixed by the question. Across aggregate metrics and this case, the experimental lesson is non-monotonic composition: each added stage changes what evidence is prioritized, and the relevant objective is coverage of the question's needs rather than the number of components used.

3.2 Evidence-conditioned control exposes gaps without guaranteeing closure

Question-only routing favored Decomposition for 42 of the 60 benchmark questions, assigning 16 to Hybrid and two to Multi-query (Fig. 3a). Its selected strategy agreed with the offline evaluation-only oracle for 22 questions; the oracle is a hindsight selector, not a runtime target. A question-conditioned policy thus concentrated its decisions on one strategy despite the mixed effects of Decomposition in the fixed-method comparison.

Adding evidence assessment exposed a different control variable: whether the current bundle was sufficient. The fixed corrective policy judged 57 of 68 cases sufficient initially and flagged 11, including one public-source question. Seven corrections used Hybrid and four used Multi-query; none produced a sufficient final ledger (Fig. 3b). Seven of the 11 flagged cases were authored abstention cases, so this is not an 11-case failure to recover facts known to exist.

The integrated controller could formulate later queries around missing reports, test modalities, or evidential scope. Of its 24 evaluation cases, 15 stopped after one retrieval and nine used multiple retrievals. All nine retained an insufficient final retrieval ledger. Fig. 3c displays the assessment states after each retrieval rather than only the terminal count. New supporting identifiers entered the final ledger in two cases, but new final-answer citations appeared in only one. In the B021-fractography case, the added context concerned the limits of available fracture evidence. In the creep case, it concerned the scope of heat-treatment and specification records. Neither supplied the missing measurement itself.

These traces distinguish an executed targeted action from successful evidence closure. They also distinguish what an assessor recognizes from what reaches the answer: the added ledger support in one case was not among the final emitted citations. The observed benefit is therefore an inspectable sequence of requests and residuals, not demonstrated superiority over a matched fixed-policy counterfactual. Explicit requirements enable targeted search, but they also make it possible to show when the search remains unsuccessful.

3.3 Integrated reliability separates valid evidence from response-mode agreement

The integrated system answered 13 of the 16 questions labeled answerable or premise-correctable and withheld responses to seven of the eight abstention-labeled questions. The remaining outcomes were three false abstentions and one answer on an abstention-labeled case. Fig. 4 aligns these response modes with each question's retrieved evidence, candidate claims, and final admission state. The resulting answer precision is 13/14, answer recall 13/16, and abstention precision 7/10. These are response-mode outcomes under the authored labels; they do not certify the scientific completeness of 13 answers.

Evidence coverage alone would give a more favorable but incomplete account. Every one of the 16 ranking-eligible questions had at least one annotated relevant item among its five selected units, giving Hit@5 = 1.0000. Recall@5 was 0.7917, MRR 0.9375, and nDCG@10 0.8016. No annotated hard negative appeared in the final selected evidence. A relevant selection nevertheless coexisted with withheld responses, so retrieval sufficiency and answer assembly require separate inspection even when aggregate ranking measures are high.

The 84 candidate claims provide another view of the same system. Rule-based checking assigned supported status to 51 non-limitation claims and 17 limitation statements. Fifteen numerical-classified claims were left unverified because no computation receipt was supplied, and one revision claim was marked superseded. None of these 16 rejected claims was admitted as supported factual content. Of the 51 supported non-limitation claims, 49 were assembled; two exceeded the five-claim output limit. Three supported limitation statements appeared as residual text.

All 39 emitted citation-identifier occurrences passed the membership checks. Final admission states were ten without a recorded residual, 11 with residuals, and three with no admissible answer. These states do not coincide with the 14 answered and ten withheld response modes. A supported limitation can contribute to a residual state without permitting a non-limitation answer, and an assembled scope statement can be labeled an answer even when it declines to provide the requested quantity.

Table 4 localizes the four disagreements. Two false abstentions arose when broad numerical typing required a receipt for a source-supported statement: a causal-sufficiency limitation in SYNQ-014-B and a runout interpretation in SYNQ-025-A. A third, SYNQ-020-B, represented the missing fracture report as an unresolved requirement and produced a limitation-only bundle, preventing a useful premise correction. The false-answer

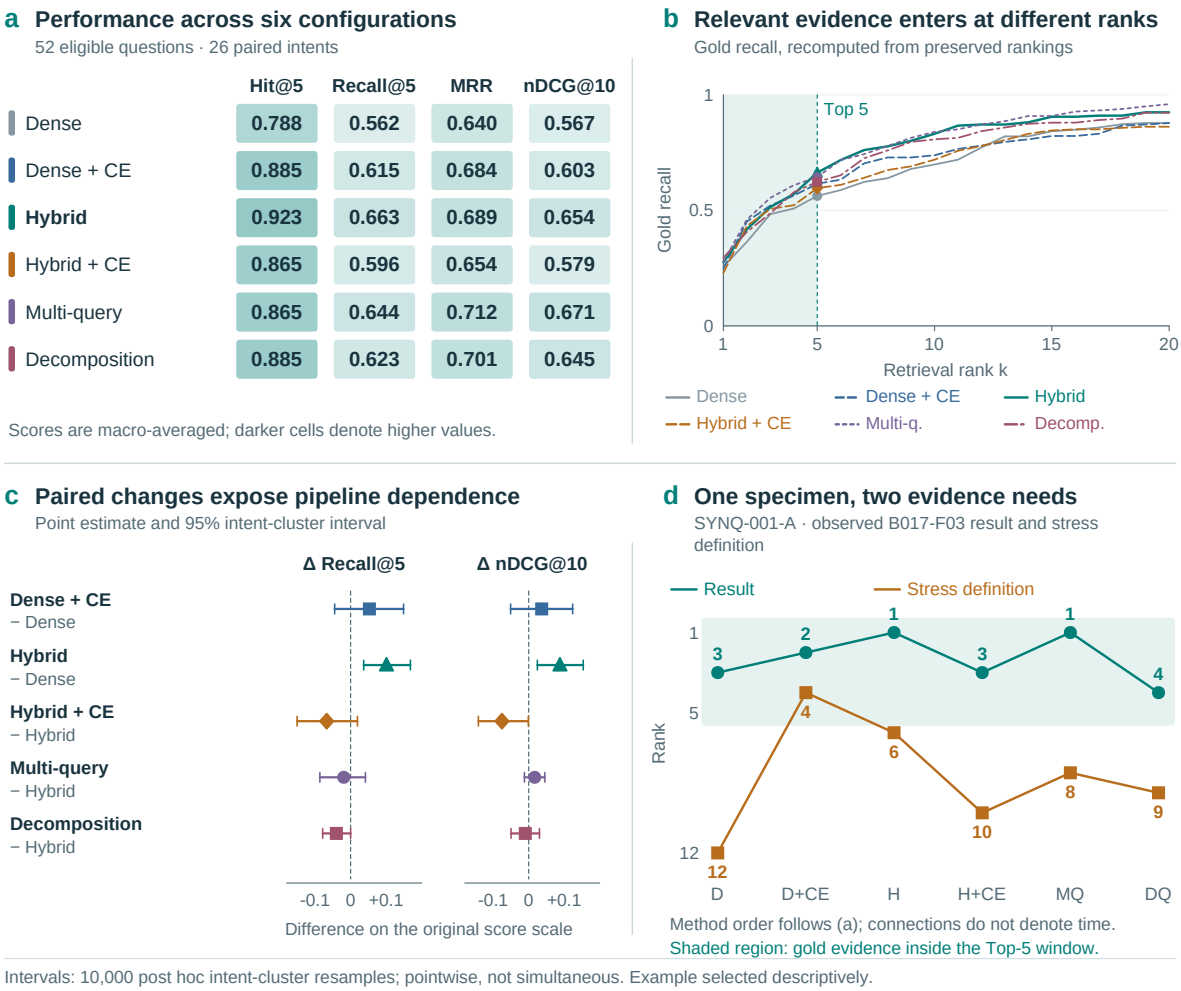


Fig. 2: Retrieval composition changes discovery, ranking, and evidence completeness differently. (a) Macro-averaged scores of the six configurations on the 52 ranking-eligible questions; darker cells are higher scores. (b) Gold-evidence Recall@k for k = 1–20 from the preserved rankings; the shaded prefix is the top five. (c) Paired changes in Recall@5 and nDCG@10 for the five contrast families; whiskers are the 95% intent-cluster bootstrap percentile intervals. (d) Ranks of the B017-F03 result and of its stress-definition section in SYNQ-001-A under each configuration; the shaded band is the top five. D, H, MQ, DQ: Dense, Hybrid, Multi-query, Decomposition; CE: cross-encoder reranking.

classification in SYNQ-022-B was different: the emitted response explained that the available records did not establish the requested creep result, but the answer-mode flag disagreed with the abstention label. It was not an invented creep-rupture measurement. These cases expose different failure mechanisms that a single correctness score would obscure.

Execution used 48 retrieval actions and 132 model calls across the 24 questions, with averages of 2.00 and 5.50, respectively (Table 3). The model calls comprise 64 controller decisions, 48 assessments, ten rewrites, and ten decompositions; revision lookups and final assembly are separate deterministic operations. The system selected 67 evidence items in total.

3.4 Revision authority and numerical semantics determine permissible interpretation

The fictional B017 investigation shows how a relevant record can lose authority for one claim while retaining value for another (Fig. 5). At 600 MPa maximum nominal stress and $R = 0.1$, the four authored B017 observations fall below the 200,000-cycle pilot gate; their descriptive median is 123,000 cycles, compared with 487,500 cycles for the four nominally matched B021 observations. Written fracture observations for three B017 specimens describe surface-connected origins with two-dimensional projected extents. These records establish the designed evidence environment, not a replicated experiment isolating a build-process effect.

The investigation initially proposed powder contamination. Subsequent source-dated chemistry and fracture records precede revision 2, which supersedes the earlier

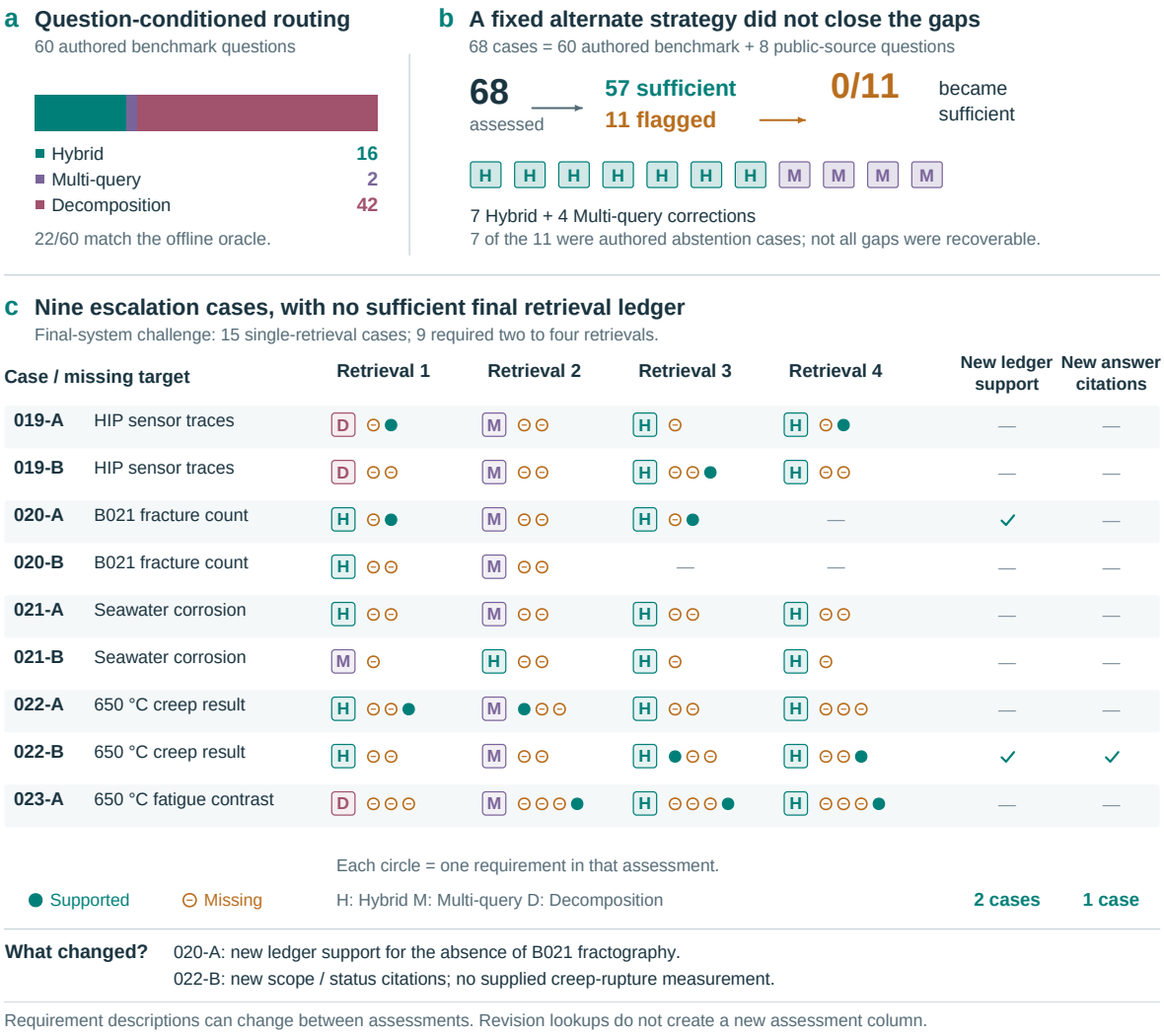


Fig. 3: Observed evidence states show the limits of additional retrieval. (a) Question-only routes for the 60 benchmark questions and agreement with the evaluation-only oracle. (b) Outcomes of the fixed corrective policy on its 68 assessed cases. (c) The nine multi-retrieval challenge cases: each column is one retrieval followed by an assessment, H/M/D boxes name the action (Hybrid, Multi-query, Decomposition), and each circle is one requirement status. Reassessment can reformulate requirements, so circle position carries no meaning across columns. Right-hand marks: new final-ledger support (two cases) and newly emitted final-answer citations (one case).

Table 3: Integrated-system response, evidence, and execution outcomes. Response-mode agreement is distinct from graded answer correctness; premise corrections count as answerable.

Level	Endpoint	Result	Population or meaning
Response behavior	Answer on an answerable question	13/16	Authored answerable/premise-correction questions
Response behavior	Withheld on an abstention question	7/8	Authored abstention questions
Response behavior	False abstention / false answer	3 / 1	24 response-mode comparisons
Selected evidence	Hit@5 / Recall@5	1.0000 / 0.7917	16 ranking-eligible questions
Selected evidence	MRR / nDCG@10	0.9375 / 0.8016	16 ranking-eligible questions
Selected evidence	Annotated hard negative selected	0/24	Questions with any selected hard negative
Evidence state	Retrieval residual	9/24	Final retrieval ledger
Attribution	Emitted identifier membership	39/39	Identifier occurrences, not entailment judgments
Execution	Retrieval actions	48; 2.00/question	24 questions
Execution	Model calls	132; 5.50/question	Controller, assessor, rewrite, and decomposition calls
Execution	Selected evidence items	67; 2.79/question	Total selected items across questions

analysis and identifies near-surface, surface-connected lack-of-fusion features as the primary supported contributor within the pilot investigation. Asked for the current conclusion at the 1 May 2025 snapshot, the system used one Hybrid retrieval. Its first five returned items included both RCA revisions; final selection retained revised status, nonconformance scope, and revised conclusion at returned ranks one, three, and five. The historical revi-

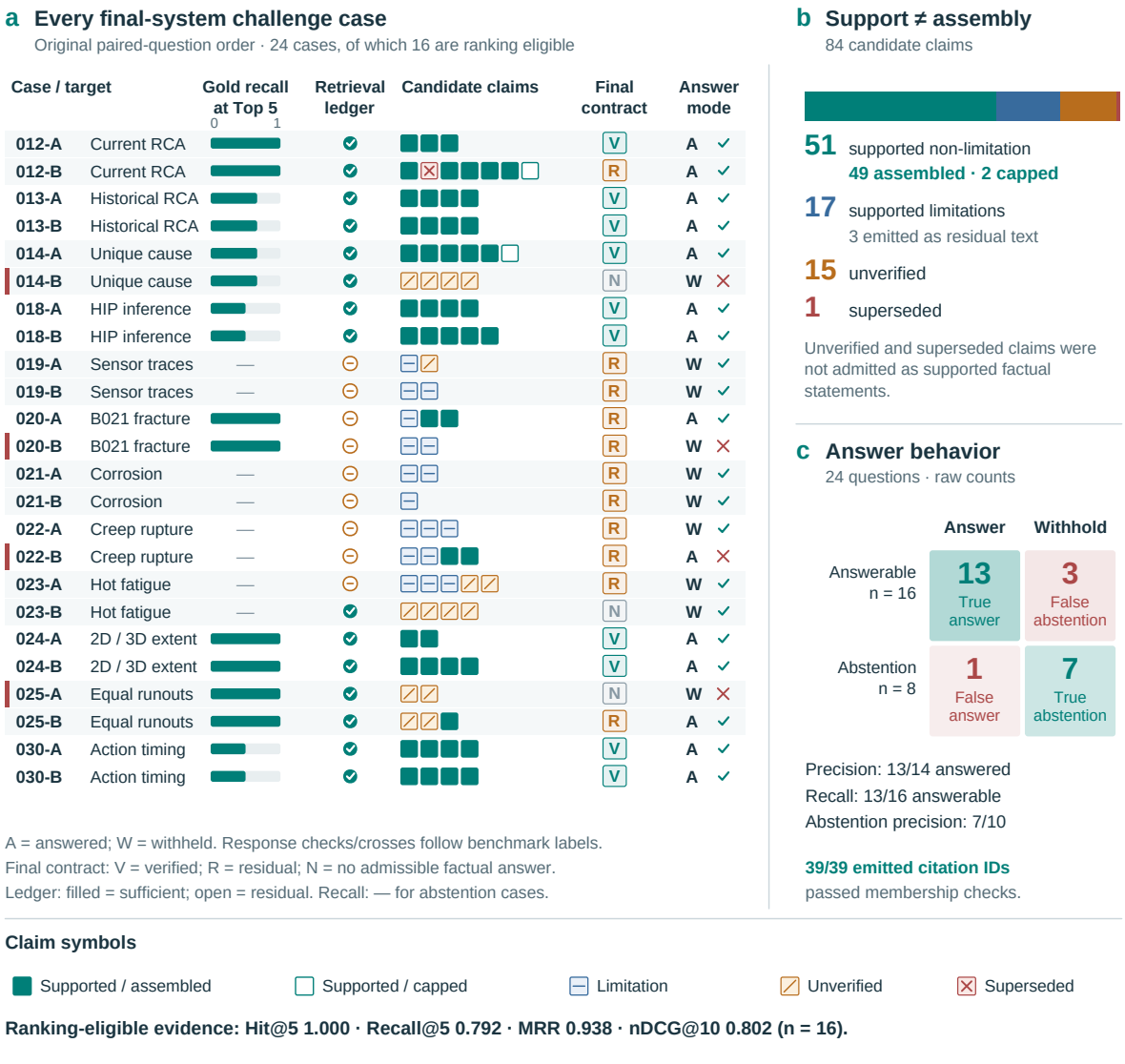


Fig. 4: Evidence selection, claim disposition, and answer behavior for every challenge case. (a) The 24 cases in preserved paired-question order; short IDs keep the SYNQ number and A/B suffix. Top-5 gold recall bars: the 16 ranking-eligible cases only. Ledger symbol: filled, sufficient; open, residual. Claim glyphs: one per candidate claim (84), in original order, keyed in the legend. Contract: V, VERIFIED; R, VERIFIED_WITH_RESIDUAL; N, NOT_ANSWERABLE_FROM_EVIDENCE. Answer mode: A, answered; W, withheld; checks and crosses mark agreement with the authored label, and the four disagreements are marked at the left edge. (b) Dispositions of the 84 candidate claims. (c) Answer/withhold confusion counts against the authored labels (n = 24).

Table 4: Observable mechanisms in the four response-mode disagreements, localized post hoc from preserved traces, not from reruns of a repaired system.

Case	Expected response	Observed mode	Primary locus	Recorded mechanism
SYNQ-014-B	Premise correction: causal sufficiency	Withheld	Claim admission	Broad numerical typing required a receipt for a source-supported causal limitation.
SYNQ-020-B	Premise correction: missing B021 fracture report	Withheld	Requirement framing and assembly	A limitation-only bundle prevented the supported correction that the requested report was unavailable.
SYNQ-022-B	Abstention: creep-rupture result	Answered	Response-mode/ assembly ambiguity	The text limited the records' scope rather than supplying a creep measurement, but its answer flag disagreed with the abstention label.
SYNQ-025-A	Premise correction: runout interpretation	Withheld	Claim admission	A source-status lookup was classified as numerical and rejected without a computation receipt.

sion ranked second but was not selected to establish the current conclusion. No separate revision-lookup action occurred. Authority was carried by the retrieved record and its use in the ledger, rather than inferred from relevance rank alone.

The public NIST examples expose a different informa-

tion boundary (Fig. 6). The typed hot isostatic pressing (HIP) record contains a pressure of 100 MPa, a temperature of 1150 °C, a hold time of 240 min, argon-medium information, and furnace cooling. Yet the record's saved Hybrid, Multi-query, and Decomposition ranks are 57, 45, and 56. The final Corrective public-case bundle selected no evidence and retained six missing requirements. The typed cooling field was also absent from the retrieval-facing text. This is an observable representation omission alongside a retrieval failure, not an isolated causal explanation for that failure. The public case has no final Agentic execution record.

For the numerical probe, the registered 1-border group contains 20 NIST experiment summaries. Removing three explicit runouts leaves 17 records, of which four lack numerical cycle endpoints. The 13 retained values reproduce the receipt median of 971,292 cycles. Because an unrecorded runout flag does not establish a failure event, the reported quantity is a pooled descriptive median of retained endpoints, not a censored-population survival estimate of the kind that would require explicit event status for every record [49]. The canonical summaries do not specify the stress definition or stress ratio needed for a fixed-loading fatigue interpretation.

The broader predefined numerical set contains eight probes: six successful receipts, one unsupported-operation request, and one intentional rejection. The six completed calculations agreed with the source-based recomputation recorded for the probe evaluation. That agreement does not remove a unit-metadata defect in the group-ratio receipt: its output retains the input unit of cycles although a ratio of like-unit medians is dimensionless. This defect and the fatigue example separate three requirements that are easily conflated: reproducible arithmetic, correctly typed output, and a scientifically interpretable estimand.

4 Discussion

4.1 Matching the information objective to the component

The retrieval comparison identifies a practical distinction between a more elaborate pipeline and a better evidence set. Hybrid provided a strong reference for top-five coverage in this corpus, while reranking and query transformation changed different aspects of the ranking. The two-section fatigue question explains why the objective matters: promoting the specimen result does not necessarily promote the section that defines how its stress should be read. A short context must retain complementary information, not simply several highly similar descriptions of the same observation.

This observation is consistent with the different functions assigned to retrieval components in the literature. Dense retrieval, lexical relevance, fusion, and pairwise reranking each define an ordering mechanism; query rewriting changes the request presented to those mechanisms [24–27, 30]. The present comparison adds a controlled materials-evidence example in which their composition changes discovery and contextual completeness differently. A possible explanation for the negative Hybrid-reranking result is that rescoring alters the balance between complementary passages. The experiment does not isolate that explanation; it does show that a reranker benefit observed with one candidate pool did not transfer to the other.

The corresponding design implication is to specify the information objective before selecting components. First-hit retrieval may be appropriate for a direct lookup; a stress-definition question requires joint support; an investigation question requires both scientific content and current authority. Multi-level evaluation detects when optimizing one object, such as the first relevant rank, leaves another, such as the required evidence bundle, incomplete.

4.2 Evidence state changes the controller's task

Question-only routing chooses a strategy before seeing its result. An evidence-conditioned controller can instead act on a named gap: a missing report, a loading condition, a revision relationship, or a limit on what the current records establish. This is the methodological role of the ledger. It converts an unstructured request for "more context" into an observable information need, while the action budget defines when acquisition must stop.

Adaptive, corrective, and self-reflective retrieval systems provide distinct antecedents to such conditional control [32–34]. Our fixed correction and bounded controller instantiate local policies rather than reproducing those systems. Their value in this study is that the intermediate evidence states are available for inspection. The all-residual escalation result would be hidden by an account that reported only successful execution of the actions. Likewise, merging ledger gains with emitted citations would overstate what the additional retrieval delivered.

Residual information is useful only when its meaning is clear. A gap in the supplied bundle is not proof of absence from the entire corpus; absence from an authored corpus is not proof of absence from scientific knowledge. The public HIP case illustrates the former distinction, because the source record exists but does not reach the assessed selection. The designed missing-measurement cases illustrate the latter. An effective interface should identify the level at which evidence is missing and communicate any supported remainder of the answer. The false-abstention cases show that preserving a residual and communicating that remainder are separate capabilities.

4.3 Admission rules need semantic distinctions, not only stricter gates

The final rule set made claim removal reproducible: numerical-classified content without a receipt was with-

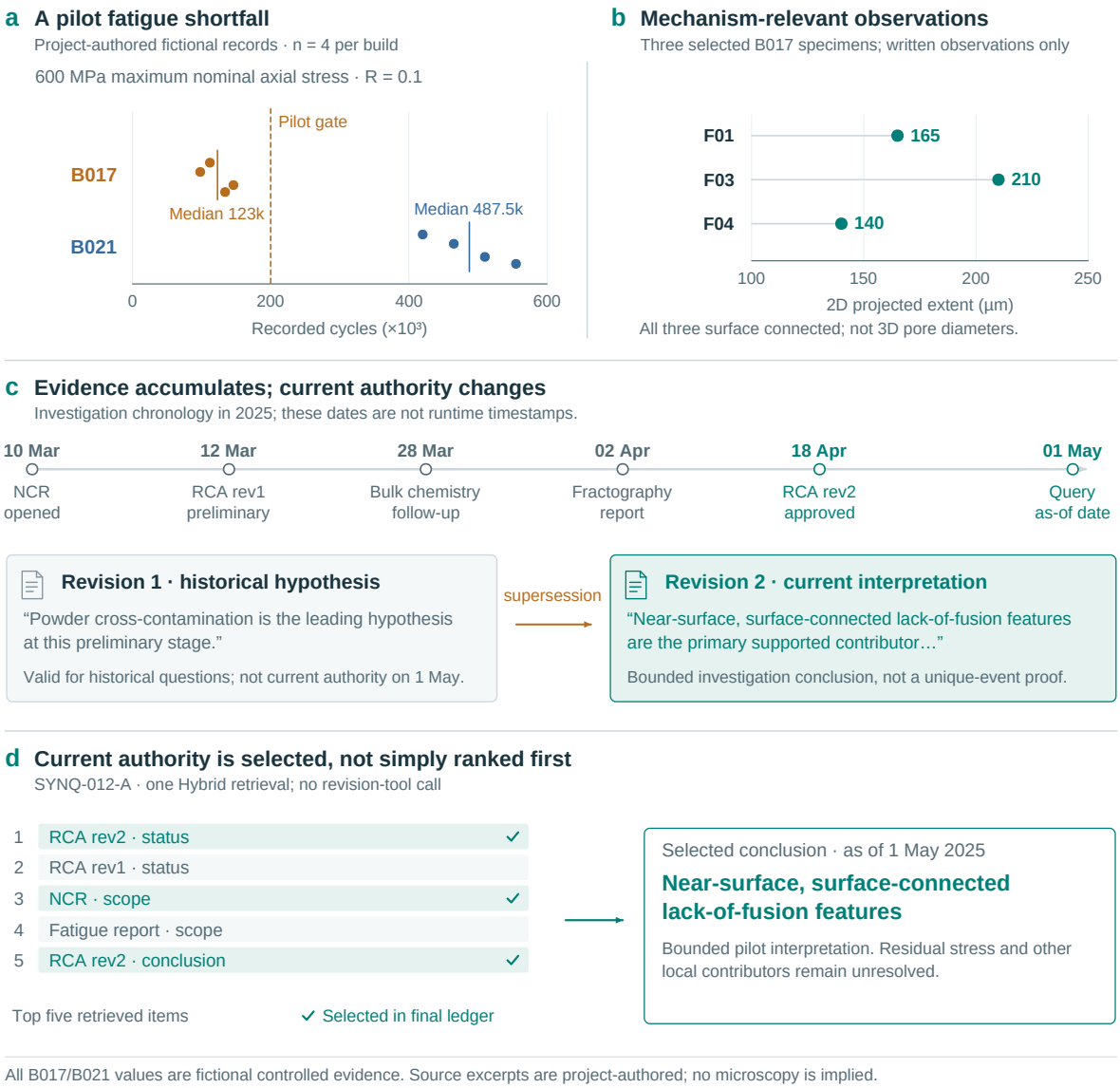


Fig. 5: Claim-dependent revision authority in the fictional B017 investigation. (a) Recorded cycles to failure of the four 600 MPa (R = 0.1) tests per build; points are individual failures (vertical offsets only separate observations), vertical marks are medians, and the vertical line is the pilot gate. (b) Two-dimensional projected extents of the three examined B017 fracture origins, all surface connected. (c) Source-dated chronology of the investigation with short excerpts from revisions 1 and 2 (the second truncated). NCR, nonconformance report. (d) The first five returned items for SYNQ-012-A; highlighted rows are the selected items (ranks 1, 3, and 5). All B017/B021 values are controlled fiction.

held, superseded support could not establish a current revision claim, and emitted citations belonged to the permitted evidence. Their failure modes nevertheless show that stricter admission is not automatically more useful assistance. A source-supported statement can be blocked when a coarse numerical category treats quoting a result or interpreting a runout as if it required a new calculation.

The needed distinction is between observed numerical evidence, derived quantities, and statements about the scope of a measurement. Each has a different admission condition. Quoting a cycle endpoint requires an attributable source and an appropriate test definition; computing a median additionally requires a defined population and exclusion policy; explaining that a runout is censored requires correct interpretation of the recorded status. A single receipt requirement cannot substitute for all three. The proposed distinction follows from the observed errors; it is a direction for a subsequent implementation, not a repair credited to the evaluated system.

Citation and factual-support research similarly distinguishes attribution from the adequacy of an assertion [41, 43]. Scientific evaluation adds a domain interpretation layer, as illustrated by the expert comparisons in Cho et al. and the entity and unit failures studied by Liu et al. [17, 46]. The present false-answer case also shows that response labels need semantic care: an answer-mode flag can disagree with an abstention label even when the text explicitly denies that the requested measurement is avail-

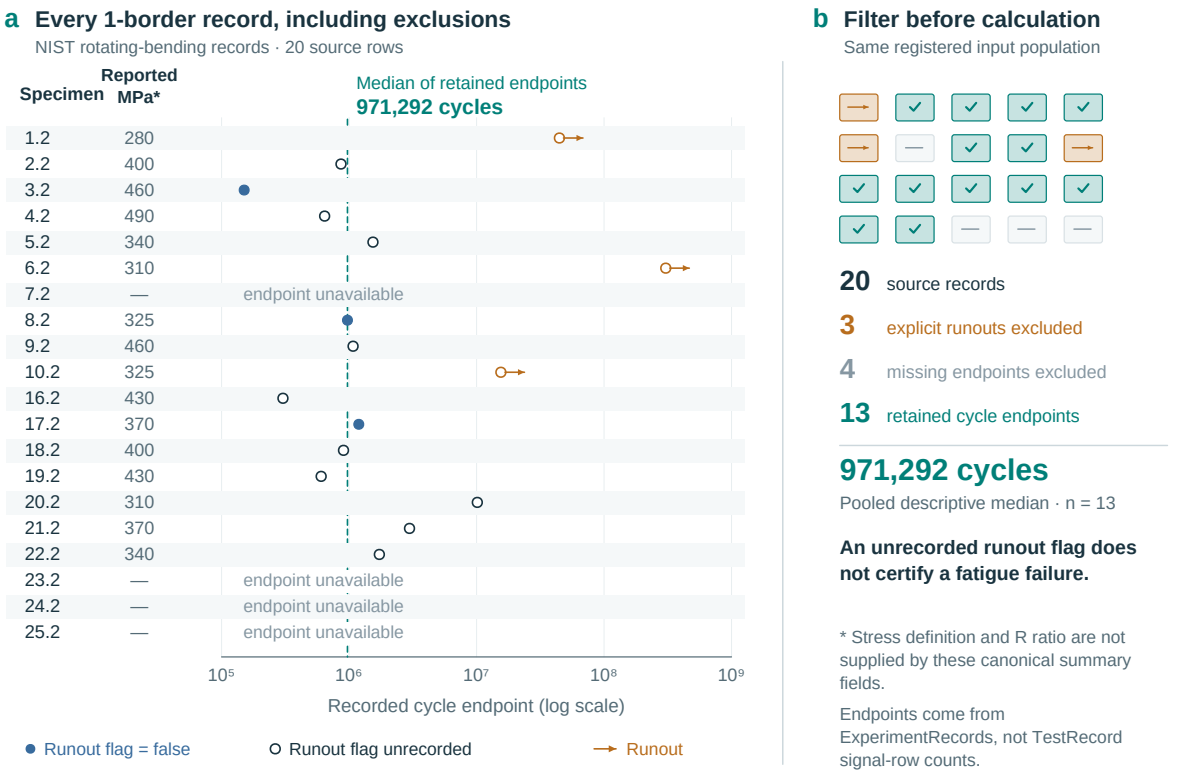


Fig. 6: Numerical provenance and unrecorded metadata constrain interpretation. (a) The 20 NIST 1-border experiment records of the registered computation input, in source row order, by recorded cycle endpoint (log scale) and reported stress. Filled blue markers: runout flag explicitly false; open dark markers: flag unrecorded; orange right-pointing symbols: the three explicit runouts (lower bounds); missing endpoints are shown as absent, not as zero. (b) The frozen exclusion rule, the 13 retained values, and their median. (c) The typed public HIP record (field values and medium string reproduced verbatim), its saved ranks under Hybrid, Multi-query, and Decomposition, and the final Corrective PUBSAN-005 selection. Sources: NIST DOI 10.18434/mds2-3044; project transformation DOI 10.5281/zenodo.22837229.

able. Such cases warrant expert assessment of usefulness and correctness, rather than assuming that a machine label or a valid citation settles both.

4.4 Materials authority belongs to the claim being made

The revision case separates the history of an investigation from the authority for its current conclusion. An early RCA can properly answer what investigators believed at issue, while the later RCA supports what was approved at a subsequent snapshot. Neither document alone turns an engineering interpretation into a replicated causal finding. The relevant relation is therefore between a claim, its time and scope, and the evidence supporting it. A global rule that simply discards every superseded document would lose valid historical information.

Numerical provenance has the same structure. An endpoint, its status, a filtered record set, and a statistic over that set are different evidence objects, and recording how each was derived from the others is the task that provenance models in e-science were designed for [50, 51]. The NIST calculation is reproducible, yet interpretation changes when unrecorded status is confused with confirmed failure or a pooled statistic is treated as a fixed-stress property. The ratio-unit defect shows that

correct arithmetic can also coexist with an incorrectly labeled output. These observations complement materials information-extraction studies in which the value, entity, unit, and measurement conditions have to be recovered together [16, 17].

Aggour et al. give factual and analytical knowledge separate representations within a compound graph [21]. The present method applies a related distinction at the boundary between retrieved source evidence and registered numerical receipts, and adds explicit sufficiency and response outcomes. Relative to LLaMP's database and simulation-oriented assistant architecture, the emphasis here is on mixed internal-style records, revision authority, and the limits of the evidence delivered to a question [38]. These are complementary scientific-assistance settings, not comparable leaderboard entries. The transferable contribution is an explicit accounting of what is known, what was computed, what remains unresolved, and what the system actually communicates.

5 Limitations

Within a controlled IN718 evidence environment with authored labels, the study establishes that retrieval composition, evidence-conditioned control, claim admission, and numerical provenance yield separable, inspectable outcomes. Four boundaries define that scope. The labels have no independent materials-expert adjudication. The 24 integrated-evaluation questions had prior component-level exposure, so the single fixed execution does not test unseen questions. No matched fixed-Hybrid end-to-end comparator exists, so orchestration is not credited with a causal end-to-end advantage. Reproduction uses the 252-unit redistributable projection of the 277-unit corpus; the 25 reference-only units require authorized access, and exact reruns depend on hosted-model availability and the unavailable historical embedding snapshot. The rule-based response stage, the predefined numerical probes, and the post hoc intervals carry the scope limits stated at their points of use. Expert adjudication, an untouched evaluation cohort, and a matched end-to-end comparison are the next axes; the separated endpoints and preserved traces are the instruments for them.

6 Conclusions

This study presents an evidence-aware approach to information assistance over heterogeneous materials-R&D records. The method preserves narrative and numerical source identity, represents question-specific support requirements, and uses those requirements to guide bounded retrieval. Its evaluation keeps evidence coverage, sufficiency, claim admission, response mode, and numerical interpretation distinct.

Three findings define the result. Retrieval components were not monotonically additive: Hybrid improved top-five coverage over Dense, while subsequent reranking and query transformation produced different trade-offs. Explicit evidence state made unresolved requirements observable, but neither the fixed corrective policy nor the multiple-retrieval integrated cases guaranteed closure. Finally, traceable admission and computation exposed limitations that favorable aggregate scores would conceal, including false abstention from broad claim typing and ambiguity in numerical status or output units.

The materials examples explain why these distinctions matter. Revision-controlled investigations require authority to be evaluated against the time and scope of a claim; numerical records require the reported quantity, event status, and derivation to remain identifiable. The resulting contribution is a reproducible method for examining the relation between materials evidence and system responses within the demonstrated testbed; its utility beyond that setting is the object of the next validation axes named above.

7 Data and code availability

The public project repository is available at `https://github.com/SFETNI/materials-rag-lab`. The archived project dataset and reproducibility materials are available through Zenodo, DOI 10.5281/zenodo.22837229 [52]. The reported experiments use the full 277-unit corpus. The redistributable projection contains 252 units; reconstruction of the remaining 25 reference-only units requires authorized access to the corresponding sources. The original NIST numerical source is identified by DOI 10.18434/mds2-3044. Source-specific rights remain applicable.

Acknowledgments

This work was supported by the Federal Ministry of Research, Technology and Space through the project "StStG CTC – Wissen schafft Perspektiven für die Region: Center for the Transformation of Chemistry gGmbH" (funding reference 03WSP1573).

Declarations

Conflict of interest. The authors declare that they have no conflicts of interest.

Author contributions. S.E. designed the study, developed the evidence-aware retrieval methodology and software, performed the experiments, analyzed the data and wrote the manuscript. T.T.-E. supervised the project and assisted with the review and quality check of the results and the manuscript. All authors discussed the results and approved the final manuscript.

References

[1] Jain A, Ong SP, Hautier G, et al (2013) Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. APL Mater 1:011002. https://doi.org/10.1063/1.4812323

[2] Draxl C, Scheffler M (2018) NOMAD: The FAIR concept for big data-driven materials science. MRS Bull 43:676–682. https://doi.org/10.1557/mrs.2018.208

[3] Blaiszik B, Chard K, Pruyne J, et al (2016) The Materials Data Facility: Data services to advance materials science research. JOM 68:2045–2052. https://doi.org/10.1007/s11837-016-2001-3

[4] Wilkinson MD, Dumontier M, Aalbersberg IJ, et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18

[5] Jacobsen A, de Miranda Azevedo R, Juty N, et al (2020) FAIR principles: Interpretations and implementation considerations. Data Intell 2:10–29. https://doi.org/10.1162/dint_r_00024

[6] Warren JA (2018) The Materials Genome Initiative and artificial intelligence. MRS Bull 43:452–457. https://doi.org/10.1557/mrs.2018.122

[7] Kalidindi SR, De Graef M (2015) Materials data science: Current status and future outlook. Annu Rev Mater Res 45:171–193. https://doi.org/10.1146/annurev-matsci-070214-020844

[8] Levine LE, Lane BM, Heigel JC, et al (2020) Outcomes and conclusions from the 2018 AM-Bench measurements, challenge problems, modeling submissions, and conference. Integr Mater Manuf Innov 9:1–15. https://doi.org/10.1007/s40192-019-00164-1

[9] Hosseini E, Popovich VA (2019) A review of mechanical properties of additively manufactured Inconel 718. Addit Manuf 30:100877. https://doi.org/10.1016/j.addma.2019.100877

[10] Sanaei N, Fatemi A (2021) Defects in additive manufactured metals and their effect on fatigue performance: A state-of-the-art review. Prog Mater Sci 117:100724. https://doi.org/10.1016/j.pmatsci.2020.100724

[11] Swain MC, Cole JM (2016) ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. J Chem Inf Model 56:1894–1904. https://doi.org/10.1021/acs.jcim.6b00207

[12] Kononova O, Huo H, He T, et al (2019) Text-mined dataset of inorganic materials synthesis recipes. Sci Data 6:203. https://doi.org/10.1038/s41597-019-0224-1

[13] Tshitoyan V, Dagdelen J, Weston L, et al (2019) Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571:95–98. https://doi.org/10.1038/s41586-019-1335-8

[14] Dagdelen J, Dunn A, Lee S, et al (2024) Structured information extraction from scientific text with large language models. Nat Commun 15:1418. https://doi.org/10.1038/s41467-024-45563-x

[15] Polak MP, Morgan D (2024) Extracting accurate materials data from research papers with conversational language models and prompt engineering. Nat Commun 15:1569. https://doi.org/10.1038/s41467-024-45914-8

[16] Circi D, Khalighinejad G, Chen A, et al (2024) How well do large language models understand tables in materials science? Integr Mater Manuf Innov 13:669–687. https://doi.org/10.1007/s40192-024-00362-6

[17] Liu S, Booth TR, Ji Y, et al (2026) Expert-grounded automatic prompt engineering for extracting lattice constants of high-entropy alloys from scientific publications using large language models. Integr Mater Manuf Innov 15:290–306. https://doi.org/10.1007/s40192-026-00460-7

[18] Immanuel J, Mahata A (2025) Enhancing materials data workflows through object-oriented design and large language models. Integr Mater Manuf Innov 14:561–572. https://doi.org/10.1007/s40192-025-00431-4

[19] Allison J, Backman D, Christodoulou L (2006) Integrated computational materials engineering: A new paradigm for the global materials profession. JOM 58(11):25–27. https://doi.org/10.1007/s11837-006-0223-5

[20] Cowles B, Backman D, Dutton R (2012) Verification and validation of ICME methods and models for aerospace applications. Integr Mater Manuf Innov 1:3–18. https://doi.org/10.1186/2193-9772-1-2

[21] Aggour KS, Detor A, Gabaldon A, et al (2022) Compound knowledge graph-enabled AI assistant for accelerated materials discovery. Integr Mater Manuf Innov 11:467–478. https://doi.org/10.1007/s40192-022-00286-z

[22] Digonta HMDA, Ghasemzadeh M, van Beek A, et al (2026) Design for ICME: A data-driven decision support framework for quantifying and managing uncertainty. Integr Mater Manuf Innov 15:528–552. https://doi.org/10.1007/s40192-026-00468-z

[23] Lewis P, Perez E, Piktus A, et al (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020), pp 9459–9474, URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[24] Robertson S, Zaragoza H (2009) The probabilistic relevance framework: BM25 and beyond. Found Trends Inf Retr 3:333–389. https://doi.org/10.1561/1500000019

[25] Karpukhin V, Oguz B, Min S, et al (2020) Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 6769–6781, https://doi.org/10.18653/v1/2020.emnlp-main.550

[26] Cormack GV, Clarke CLA, Buettcher S (2009) Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 758–759, https://doi.org/10.1145/1571941.1572114

[27] Nogueira R, Cho K (2019) Passage re-ranking with BERT. arXiv:1901.04085

[28] Luan Y, Eisenstein J, Toutanova K, et al (2021) Sparse, dense, and attentional representations for text retrieval. Trans Assoc Comput Linguist 9:329–345. https://doi.org/10.1162/tacl_a_00369

[29] Wang S, Zhuang S, Zuccon G (2021) BERT-based dense retrievers require interpolation with BM25 for effective passage retrieval. In: Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR), pp 317–324, https://doi.org/10.1145/3471158.3472233

[30] Ma X, Gong Y, He P, et al (2023) Query rewriting in retrieval-augmented large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 5303–5315, https://doi.org/10.18653/v1/2023.emnlp-main.322

[31] Khot T, Trivedi H, Finlayson M, et al (2023) Decomposed prompting: A modular approach for solving complex tasks. In: International Conference on Learning Representations (ICLR 2023), URL https://openreview.net/forum?id=_nGgzQjzaRy

[32] Jeong S, Baek J, Cho S, et al (2024) 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, https://doi.org/10.18653/v1/2024.naacl-long.389

[33] Yan SQ, Gu JC, Zhu Y, et al (2024) Corrective retrieval augmented generation. arXiv:2401.15884

[34] Asai A, Wu Z, Wang Y, et al (2024) Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In: International Conference on Learning Representations (ICLR 2024), URL https://openreview.net/forum?id=hSyW5go0v8

[35] Yao S, Zhao J, Yu D, et al (2023) ReAct: Synergizing reasoning and acting in language models. In: International Conference on Learning Representations (ICLR 2023), URL https://openreview.net/forum?id=WE_vluYUL-X

[36] Boiko DA, MacKnight R, Kline B, et al (2023) Autonomous chemical research with large language models. Nature 624:570–578. https://doi.org/10.1038/s41586-023-06792-0

[37] Bran AM, Cox S, Schilter O, et al (2024) Augmenting large language models with chemistry tools. Nat Mach Intell 6:525–535. https://doi.org/10.1038/s42256-024-00832-8

[38] Chiang Y, Hsieh E, Chou CH, et al (2025) LLaMP: Large language model made powerful for high-fidelity materials knowledge retrieval. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 25189–25221, https://doi.org/10.18653/v1/2025.emnlp-main.1280

[39] Ansari M, Moosavi SM (2024) Agent-based learning of materials datasets from the scientific literature. Digit Discov 3:2607–2617. https://doi.org/10.1039/D4DD00252K

[40] Buehler MJ (2024) Generative retrieval-augmented ontologic graph and multiagent strategies for interpretive large language model-based materials design. ACS Eng Au 4:241–277. https://doi.org/10.1021/acsengineeringau.3c00058

[41] Gao T, Yen H, Yu J, et al (2023) Enabling large language models to generate text with citations. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 6465–6488, https://doi.org/10.18653/v1/2023.emnlp-main.398

[42] Rashkin H, Nikolaev V, Lamm M, et al (2023) Measuring attribution in natural language generation models. Comput Linguist 49:777–840. https://doi.org/10.1162/coli_a_00486

[43] Min S, Krishna K, Lyu X, et al (2023) FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 12076–12100, https://doi.org/10.18653/v1/2023.emnlp-main.741

[44] Es S, James J, Espinosa Anke L, et al (2024) RAGAs: Automated evaluation of retrieval augmented generation. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pp 150–158, https://doi.org/10.18653/v1/2024.eacl-demo.16

[45] Chen J, Lin H, Han X, et al (2024) Benchmarking large language models in retrieval-augmented generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 17754–17762, https://doi.org/10.1609/aaai.v38i16.29728

[46] Cho ZH, Osvaldo M, Doloi S, et al (2026) Large language models in materials science: Assessing RAG evaluation frameworks through graphene synthesis. RSC Adv 16:11306–11313. https://doi.org/10.1039/D5RA09726F

[47] Kamath A, Jia R, Liang P (2020) Selective question answering under domain shift. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp 5684–5696, https://doi.org/10.18653/v1/2020.acl-main.503

[48] Beijing Academy of Artificial Intelligence (2023) bge-base-en-v1.5: Model card. Hugging Face model card, URL https://huggingface.co/BAAI/bge-base-en-v1.5, accessed 12 October 2026

[49] Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481. https://doi.org/10.1080/01621459.1958.10501452

[50] Simmhan YL, Plale B, Gannon D (2005) A survey of data provenance in e-science. ACM SIGMOD Rec 34(3):31–36. https://doi.org/10.1145/1084805.1084812

[51] Moreau L, Clifford B, Freire J, et al (2011) The Open Provenance Model core specification (v1.1). Future Gener Comput Syst 27:743–756. https://doi.org/10.1016/j.future.2010.07.005

[52] Elfetni S (2026) Materials RAG Lab dataset: IN718 additive-manufacturing evidence and benchmark for progressive RAG (version 0.1.0). Zenodo, https://doi.org/10.5281/zenodo.22837229, dataset. Source-specific rights remain applicable