SAME PAPER, DIFFERENT DECISION: A CONTROLLED DIAGNOSTIC OF TASK-CONDITIONED SCIENTIFIC EVIDENCE READINESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific papers are not intrinsically ready or unready for information extraction: their usability depends on the evidence required by the downstream task. We formulate task-conditioned scientific evidence readiness, in which a system maps a paper–task pair to one of three operational actions: READY, REVIEW_REQUIRED, or ABSTAIN. To isolate task conditioning from document variation, we construct a paired diagnostic set from eight Ti–6Al–4V laser powder bed fusion papers and three extraction tasks that differ only in their material-subtype requirement, yielding 24 correlated evaluations. Holding the paper fixed, the frozen AI-assisted author-reference action changes across tasks for six of eight papers. We compare four frozen prompting and control configurations. Direct prompting matches the reference action on 16/24 pairs, while the full structured configuration matches 14/24, indicating that additional structure does not consistently improve decisions in this testbed. An evidence-grounded configuration makes no false READY decisions among its 18 valid outputs but has execution failures on 6/24 pairs, exposing a safety–coverage tradeoff. Error analysis further shows that an output can record a task-relevant evidence conflict yet select an inconsistent operational action. These errors motivate a stage-wise diagnostic perspective separating evidence-state recognition from task-dependent action mapping; the stages are not independently evaluated here. Our results motivate pair-conditioned evaluation for scientific extraction systems whose outputs trigger downstream use or human review.

1 INTRODUCTION

Scientific papers are not intrinsically ready or unready for extraction. Their usability depends jointly on document evidence and the downstream extraction specification. A paper may name a material without establishing a required grade, report heat treatment without a linked cooling condition, or give a tensile value for a specimen group whose processing history is unclear. A fluent structured record can conceal these discontinuities. Materials-extraction studies document the practical difficulty of obtaining accurate, supported records from scientific text (Dagdelen et al., 2024; Ghosh et al., 2024; Polak & Morgan, 2024).

Consider an unchanged Grade 23, extra-low-interstitial (ELI) Ti–6Al–4V paper. It can be compatible with an ELI task and hard-incompatible with a non-ELI task. Removing the subtype constraint removes that conflict, while a separate missing test condition may still require review. The document stays fixed, but the task changes and the appropriate operational action can change with it. A paper-only suitability label cannot represent this relation.

Prior work supplies essential components: scientific information extraction evaluates records and relations; grounded QA conditions answers on a paper and information need; evidence sufficiency studies support conditions; and abstention and selective prediction consider declining or deferring decisions. We do not claim to introduce those concepts or the use of three operational actions. The relevant question concerns how a specified scientific source supports a downstream extraction requirement.

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.

Our narrower focus is a fixed-document, controlled downstream-requirement intervention combined with within-document readiness transition evaluation. The prior work discussed below gives less attention to this paired design for scientific structured extraction. Holding the document fixed makes it possible to inspect whether changing a requirement changes the readiness action at the intended boundary. We define readiness over a paper–task pair, with `READY`, `REVIEW_REQUIRED`, and `ABSTAIN` describing operational use under a specified task policy.

The current study pairs eight independent Ti–6Al–4V papers with three controlled subtype variants, yielding 24 correlated evaluations. The frozen AI-assisted author-reference action changes for 6/8 papers. Four frozen prompting and control configurations obtain Direct 16/24, Schema-only 13/24, Evidence-grounded 13/24, and Full 14/24 end-to-end exact decisions. Additional structure does not consistently improve decisions in this set. Error analysis suggests a further diagnostic distinction between evidence-state recognition and action mapping: in R09, Full records a hard subtype conflict yet chooses review rather than the reference action of abstention.

Our contributions are a task-conditioned formulation with explicit operational semantics, a small paired diagnostic set exposing within-paper transitions, and a descriptive comparison with qualitative error analysis. The evidence-to-action interpretation is motivated by recorded outputs; it is not a separately validated stage-wise method or a claim about the dominant source of error.

2 Related Work

Scientific information extraction. LLMs have broadened scientific extraction from specialized sequence labeling toward instruction- and schema-driven records. Prior studies evaluate structured extraction across materials tasks, compare language models with specialized entity and relation systems, reconstruct materials datasets from full papers, and examine missing information and verification (Dagdelen et al., 2024; Foppiano et al., 2024; Ghosh et al., 2024; Polak & Morgan, 2024). Their principal target is the correctness of entities, relations, or populated records after extraction. Our target is earlier: whether a source contains evidence adequate for a downstream schema without review, only with review, or not at all.

Paper-grounded QA and evidence attribution. Scientific QA already conditions an answer on a paper and an information need. QASPER provides answerability and supporting spans; ResearchQA evaluates citation-grounded single-paper QA; LitTraceQA separates paper identification, evidence localization, and final answers (Dasigi et al., 2021; Imran & Solanky, 2026; Liu et al., 2026). These are important pairwise precedents, but they optimize answer and citation outcomes rather than prospective structured-extraction source readiness. Evidence-grounded generation similarly separates answer quality, citation correctness, retrieval relevance, correctness, and faithfulness (Gao et al., 2023; Es et al., 2024; Adlakha et al., 2024); RAG and Self-RAG connect retrieval with generation or critique (Lewis et al., 2020; Asai et al., 2024). Our unit is instead a specified paper–task pair before accepting a structured record.

Evidence sufficiency, answerability, and abstention. Unanswerable reading comprehension, Abstain-QA, abstention surveys, selective prediction, and learning to defer establish that rejection and expert routing are not new (Rajpurkar et al., 2018; Madhusudhan et al., 2025; Wen et al., 2025; Xin et al., 2021; Mozannar & Sontag, 2020). Selective clinical-text extraction is an especially relevant applied precedent (Swaminathan et al., 2024). Our labels instead encode whether a scientific source supports a task-valid structured record. The closest benchmark is ESB, which fixes a QA question and varies external evidence across complete, incomplete, irrelevant, absent, and conflicting contexts (Zhang & Wu, 2026); related science-QA work perturbs context and observes abstention (Wen et al., 2024). The intervention axes are related but distinct: those studies keep the question fixed and alter evidence, whereas we keep the paper fixed and alter the extraction requirement. Our target is three-way source readiness and preservation of within-paper transitions, a narrow benchmark claim rather than a claim to have introduced sufficiency or abstention.

Controlled evaluation and scientific agents. Contrast sets reveal local boundaries through small label-changing perturbations (Gardner et al., 2020). We use the same high-control intuition while holding the paper fixed and varying the task. Scientific agents such as PaperQA2 and benchmarks such as ScienceAgentBench motivate readiness checks because downstream agents can propagate

Table 1: Controlled task definitions. The paper and all requirements except material subtype are fixed.

Task	Base material	Subtype	Decision rule
V2-A1	Ti–6Al–4V	Non-ELI	Grade 5/non-ELI is compatible; Grade 23/explicit ELI is a hard conflict; plain Ti–6Al–4V is unresolved.
V2-A2	Ti–6Al–4V	ELI	Grade 23/explicit ELI is compatible; Grade 5/non-ELI is a hard conflict; plain Ti–6Al–4V is unresolved.
V2-A3	Ti–6Al–4V	None	Only base Ti–6Al–4V identity is required.

unsupported records (Skarlinski et al., 2024; Chen et al., 2025); they do not supply this source-level target.

3 Problem Formulation

Let $p \in \mathcal{P}$ be a scientific paper, $t \in \mathcal{T}$ a structured-extraction task specifying target scope and evidence requirements, and $E(p)$ the permitted evidence from p. We define readiness over the pair:

$$R(p,t) \in \{\texttt{READY}, \texttt{REVIEW_REQUIRED}, \texttt{ABSTAIN}\}. \tag{1}$$

The decision concerns whether $E(p)$ supports the record required by t; it does not assess paper merit, predict an unreported property, or determine truth outside the supplied evidence.

`READY` requires every task-required assertion and a continuous chain for a task-valid specimen or explicitly documented lineage, with no unresolved hard scope conflict. Here the chain links material and manufacturing route to specimen identity and orientation, post-build heat treatment and cooling, tensile-test conditions, and a UTS value with statistic interpretation. `REVIEW_REQUIRED` means relevant evidence exists but material incompleteness, ambiguity, or linkage uncertainty prevents unreviewed extraction. Human adjudication, clarified policy, or more source material may resolve it. `ABSTAIN` is the operational action under the frozen task policy for a hard-incompatible or unsupported paper–task pair that cannot support the requested extraction without speculation. These are task-policy actions rather than absolute scientific gold classes; the three-way label set itself is not a novelty claim.

For diagnostic interpretation, let $S(p,t)$ conceptually represent task-conditioned evidence states across required fields and constraints, retaining source locations, specimen linkage, and constraint hardness. Components can describe supported, missing, ambiguous, conflicting, or unsupported assertions. In particular, unsupported completion asserts more than the source supplies. This vocabulary is not a newly annotated, mutually exclusive state taxonomy. We write

$$R(p,t) = \pi(S(p,t), t), \tag{2}$$

where π denotes the evidence-to-action policy. The task remains explicit because the consequence of a gap or conflict depends on its requirements. A recorded hard conflict followed by review can suggest an action-mapping inconsistency under that policy, even without establishing correctness of the entire state.

This decomposition is conceptual and diagnostic. The current experiment does not provide independent gold evidence-state labels for every method output and therefore does not separately estimate evidence-state recognition accuracy and policy-mapping accuracy. The evaluated configurations were not run as independently validated stages. The notation organizes the existing error analysis and does not add a method, score, or causal attribution.

The paired intervention permits $R(p,t_1) \neq R(p,t_2)$ with the paper fixed. Table 1 records the three existing task variants.

The shared requirements are a compatible manufacturing route, build orientation, post-build heat-treatment temperature, holding time, atmosphere and cooling, tensile-test temperature, UTS, statistic interpretation, and supported specimen/group continuity. A reference-label change is therefore attributable to subtype compatibility and its interaction with remaining evidence gaps.

4 Diagnostic Set Construction

4.1 Candidate selection and full-text gate

The paired diagnostic set was assembled through stages separating document relevance from readiness labeling. Metadata discovery identified papers potentially concerning Ti–6Al–4V, laser powder bed fusion (L-PBF), post-build heat treatment, and tensile outcomes. A candidate screen assessed whether acquisition was warranted but did not assign readiness. Acquired documents passed a full-text gate requiring the material, route, post-build heat treatment, and tensile endpoint. Acquisition failure remained operational and was never converted into scientific `ABSTAIN`. The set contains R01, R03, R05, R07, R08, R09, R10, and R13: eight independent papers, each crossed with A1–A3, giving 24 correlated evaluations. Papers outside the gate are absent from the denominator. This small controlled benchmark supports paired diagnosis, not population estimation.

4.2 Frozen author reference

Reference records contain task-required field states, evidence identifiers, hard mismatches, specimen/group continuity, rationales, and one readiness label per unit. Construction used evidence-grounded annotation, focused review for R13, and a deterministic freeze. Its authority is an AI-assisted author reference, not independent expert gold: no independent expert consensus, external adjudication, or inter-annotator agreement study was performed.

The reference was frozen before formal execution at commit `1964e62b10c961a16dabc081215d2ec28c807cef`. Generation components could not read labels or review artifacts; method outputs could not modify the reference; and the scorer accessed it only after all 96 scheduled runs reached terminal status. This temporal separation supports comparison to the frozen reference without elevating it to expert gold.

5 Evaluated Configurations

The four frozen prompting and control configurations differ in degrees and types of schema, evidence, and decision constraints; they are not a strictly nested capability hierarchy. All receive the same paper–task input and provider/model configuration and share a logical output envelope with slots for extracted fields, evidence links, unresolved items, and one readiness decision. A shared envelope does not make evidence binding comparable by itself: only Evidence-grounded and Full require comparable field-linked bindings in their output contracts.

Direct LLM. Direct extracts requested fields and issues a decision without an added fixed field list, field-level evidence protocol, sample-chain rule, or specialized abstention guidance beyond the common output contract.

Schema-only. Schema-only adds a fixed schema covering alloy and subtype, manufacturing route and parameters, orientation, heat treatment, tensile-test temperature, UTS, and statistic interpretation. It has no explicit evidence-validation, sample-chain, or readiness-reasoning protocol, isolating output structure from grounding rules.

Evidence-grounded. Evidence-grounded requires each reported field and rationale to link to a supplied evidence location. Insufficient evidence remains unresolved rather than being completed from task context or parametric knowledge. It lacks Full's sample-chain and decision-policy instructions.

Full. The full structured configuration combines the schema with task conditioning, per-field evidence checks, specimen/sample-chain verification, contradiction and missing-link recording, and explicit readiness rules. It is directed to select `READY` only when required evidence is supported. Full is an evaluated configuration; no advantage is presumed.

Table 2: Formal readiness-label results relative to the frozen AI-assisted author reference.

Method	Coverage	E2E exact	Conditional	Common support
Direct LLM	23/24 (95.8%)	16/24 (66.7%)	16/23 (69.6%)	11/14 (78.6%)
Schema-only	21/24 (87.5%)	13/24 (54.2%)	13/21 (61.9%)	8/14 (57.1%)
Evidence-grounded	18/24 (75.0%)	13/24 (54.2%)	13/18 (72.2%)	10/14 (71.4%)
Full	23/24 (95.8%)	14/24 (58.3%)	14/23 (60.9%)	8/14 (57.1%)

6 Experimental Protocol

The formal evaluation crosses eight papers, three tasks, and four methods for 96 scheduled calls. All methods use identical paper–task inputs, a common logical output schema, provider configuration, and frozen scorer. Attempt 003 requested the provider model identifier `deepseek-v4-pro`, high reasoning, a 10,240-token maximum output budget, one completion, JSON-object response format, no tools or web access, and zero SDK retries. The provider response did not expose an independently verifiable snapshot identifier, so no model snapshot is claimed. Temperature and top-p were not sent in reasoning mode. Attempt 003 returned 85 valid outputs and retained 11 execution failures. Two pre-scoring preflights identified output-budget exhaustion; increasing the output budget was the only configuration change leading to the final source. Earlier attempts remain provenance only and never backfill results.

Coverage is valid outputs divided by 24. End-to-end (E2E) exact readiness success is exact valid predictions divided by all 24, treating failure as operationally unsuccessful but not as a scientific label. Conditional exact agreement uses method-valid outputs. Common-support agreement uses the 14 units valid for every method. Safety reporting separates false `READY`, missed `READY`, and failure. Transition behavior is evaluated per paper only when all A1–A3 outputs are valid. Since units are dependent within eight paper clusters, comparisons are descriptive and support no statistical-superiority claim.

7 Results

7.1 Task-conditioned readiness

Six of eight papers change frozen author-reference readiness when only material-subtype requirement changes. The paper and every other extraction requirement stay fixed, so the observed 6/8 establishes the existence of task-conditioned transitions in this controlled set, not their prevalence in scientific literature. R09 moves from `ABSTAIN` under non-ELI to `REVIEW_REQUIRED` under ELI and unconstrained tasks. R05 moves from review in both constrained tasks to `READY` when subtype is removed. Both invariant papers remain in the denominator.

7.2 End-to-end decision comparison

Table 2 reports the frozen comparison. Direct has 23/24 coverage and 16/24 E2E exact agreement; Full has the same coverage and 14/24 agreement. Schema-only and Evidence-grounded reach 13/24 with 21/24 and 18/24 coverage. Evidence-grounded's 13/18 conditional result excludes six failures and must be read with coverage. On 14 common-valid units, Direct obtains 11/14. Additional structure did not consistently improve end-to-end decisions in this testbed. These observations do not establish a general ranking or isolate the effect of adding each instruction.

Pairwise common-valid counts support the same restraint. For Direct versus Full, both are correct on 12 units, Direct alone on three, Full alone on one, and both wrong on six. For Evidence-grounded versus Full, both are correct on nine, Evidence-grounded alone on three, Full alone on one, and both wrong on four. These descriptive comparisons do not support Full-over-baseline superiority.

7.3 Safety, coverage, and execution tradeoff

Evidence-grounded produces zero false `READY` decisions among 18 valid outputs, but misses one reference-`READY` case and has six execution failures among 24 scheduled evaluations. Full produces

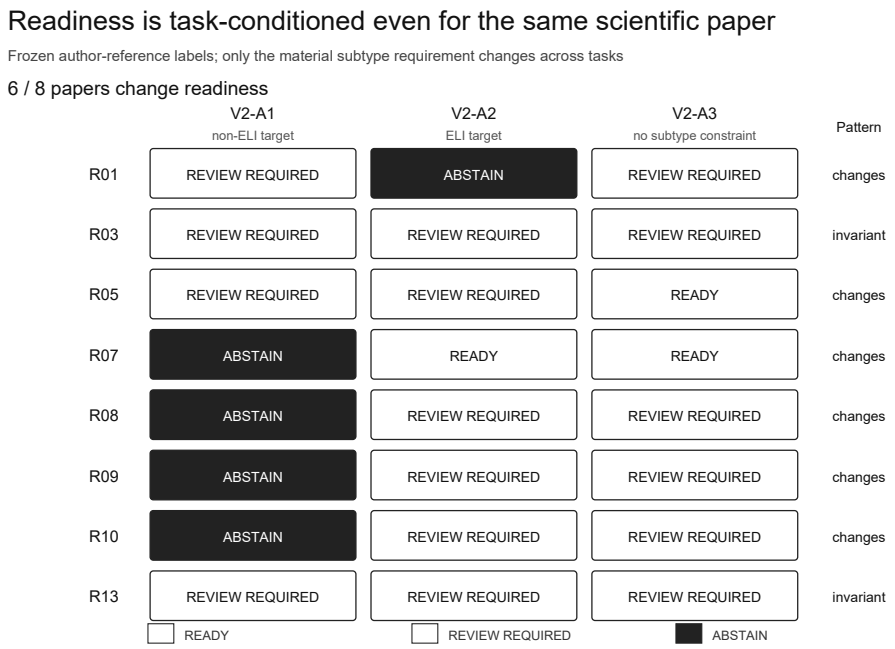

Figure 1: Paired same-paper evaluation: frozen AI-assisted author-reference labels under A1, A2, and A3. Six of eight papers change label; each source and all non-subtype requirements are fixed.

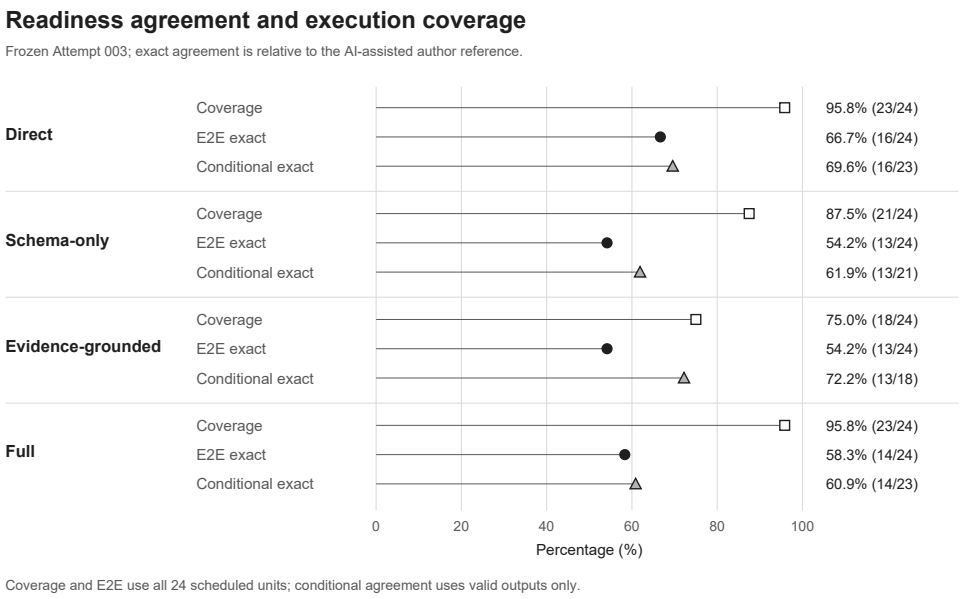

Figure 2: Coverage and readiness accuracy. E2E uses all 24 units; conditional agreement uses valid denominators 23, 21, 18, and 23.

two false and two missed `READY` decisions and fails once. Direct produces three false `READY` decisions, misses none among valid outputs, and fails once; Schema-only produces three, zero, and three respectively. Evidence-grounded thus produces fewer false `READY` decisions among valid

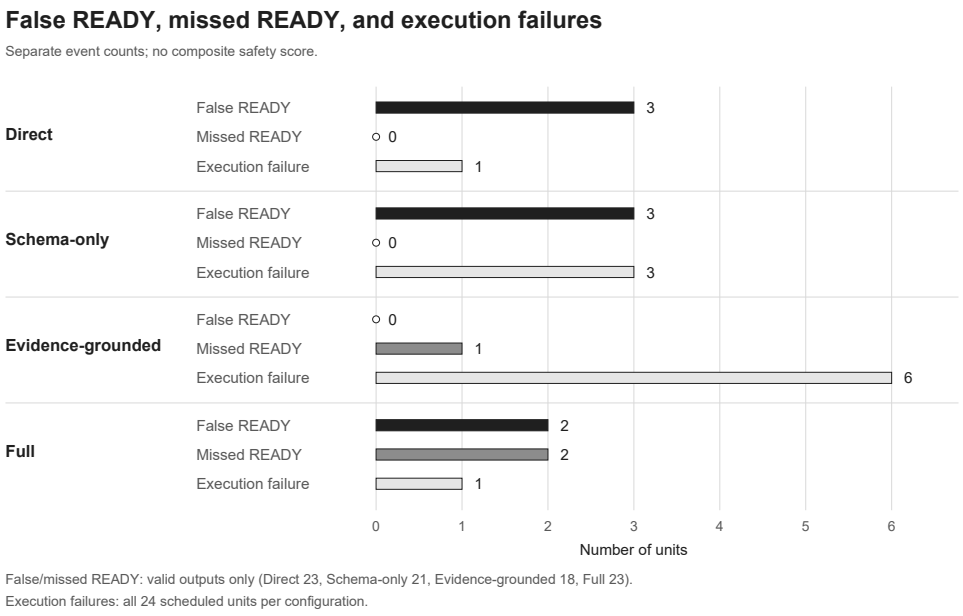

Figure 3: False READY, missed READY, and execution failure remain separate. Evidence-grounded has zero false READY but six failures.

outputs while exhibiting substantially lower execution coverage. Appendix Table 3 preserves the separate counts; the observed zero does not establish overall safety superiority.

Evidence-grounded has nonempty evidence bindings on 18/18 valid outputs and Full on 23/23. Direct and Schema-only lack a formally comparable binding interface. This supports inspectability, not field-level binding correctness or superiority: traceability and readiness correctness are distinct.

7.4 Error analysis

The existing transition summaries provide context for the qualitative cases. Direct has seven complete papers, matches 1/7 patterns, and detects sensitivity in 4/6 eligible sensitive papers. Schema-only has five complete papers, matches 0/5, and detects 3/5. Evidence-grounded has four complete papers, matches 2/4, and detects 1/2. Full has seven complete papers, matches 1/7, detects 2/5, and produces three false-invariance and one false-sensitivity errors. The all-method complete-paper intersection is empty ($N = 0$). These descriptive diagnostics retain their method-specific eligibility rules; Appendix Tables 4 and 5 preserve the full summaries.

R09: a recognized conflict followed by an inconsistent action. The reference pattern is ABSTAIN, REVIEW_REQUIRED, REVIEW_REQUIRED: Grade 23/ELI conflicts with A1's non-ELI requirement, while unsupported tensile-test temperature retains review for A2/A3. Direct reproduces the pattern. Full explicitly records the A1 subtype hard conflict yet returns review for all three tasks. The decisive conflict is visible in the output, so this particular mismatch cannot be explained solely by failure to retrieve, localize, or notice it. Under the frozen task policy the conflict requires abstention, but the emitted action is review. This case is consistent with an evidence-to-action mapping failure; it does not establish correctness of every field in S, reveal internal computation, or identify the dominant error source across the set.

R05 and R13: errors in opposite directions. For R05-A3, Full promotes atmosphere, cooling, and applicability concerns beyond the evaluated task boundary and downgrades reference READY to review. For R13-A3, Full accepts a highest reported UTS as resolving statistic semantics and upgrades reference review to READY. Relative to the frozen author policy, these cases show an over-conservative downgrade and an unsupported upgrade. The selected errors are consistent with inconsistent evidence-to-action mapping rather than a uniform conservative or permissive bias. Ev-

idence interpretation and policy application may interact, so the cases do not constitute a controlled separation of the two stages.

8 DISCUSSION

Readiness is relational. The relevant unit is paper $\times$ task. A change in scope can alter compatibility even when the source evidence is unchanged. Scholarly QA already demonstrates dependence on an information need, and selective prediction supplies defer/reject actions. Here the three labels specify scientific extraction use under a frozen task policy. `REVIEW_REQUIRED` denotes relevant evidence that cannot support unreviewed extraction, rather than a generic confidence band or a claim that review is a new action.

Paired evaluation controls document variation. ESB and related context-perturbation work fix a question and vary evidence; this paired diagnostic fixes the paper and changes a downstream requirement. It isolates observed task sensitivity from changing document content and makes within-paper transitions inspectable. Eight paper clusters provide case-level control, not population estimation. The observed 6/8 transitions describe this set and frozen reference only.

Traceability does not guarantee the final action. Evidence bindings make stated assessments available for inspection, but their presence does not certify field correctness or policy execution. R09 supplies a strong qualitative example: an explicit hard conflict and an inconsistent review action coexist in one output. The frozen comparison also retains Direct 16/24 versus Full 14/24 and Evidence-grounded's zero false `READY` among 18 valid outputs alongside six failures. Additional structure did not consistently improve decisions in this testbed; this supports neither universal Direct superiority nor the claim that grounding is useless.

A diagnostic perspective for future controlled evaluation. The errors motivate separating evidence-state recognition from evidence-to-action policy execution. The current study cannot strictly distinguish evidence localization, field extraction, specimen-linkage, evidence-state classification, and policy-mapping errors across all outputs. R09 makes one local inconsistency visible, but independent evidence-state annotations and controlled state inputs are absent. We therefore do not attribute all errors to policy mapping or estimate its prevalence. The conceptual decomposition identifies a question for future evaluation, not an empirically validated two-stage mechanism.

9 LIMITATIONS

The paired diagnostic contains only eight independent paper clusters and 24 correlated evaluations. It establishes the existence of transitions under this controlled set and reference, but provides no population-prevalence estimate or statistical-superiority claim. The reference is AI-assisted author annotation, not independent materials-expert gold, consensus, or independently adjudicated ground truth. There was no independent materials-expert annotation, external adjudication, or inter-annotator agreement study. All agreement and qualitative correctness judgments depend on the frozen author reference; freezing does not substitute for independent validation.

Attempt 003 has 11/96 execution failures, six in Evidence-grounded. Failures are not imputed, and the common complete-paper intersection across all configurations is $N = 0$. The domain is narrow—Ti–6Al–4V, L-PBF, post-build heat treatment, and tensile outcomes—and subtype is the only controlled intervention dimension. Findings may not transfer to other materials, tasks, or evidence policies. All configurations share one provider/model configuration without an independently verified snapshot, so base-model, prompt, context, schema, and controller effects cannot be isolated. Evidence-binding correctness was not deterministically comparable across all configurations.

The stage-wise evidence-to-action interpretation is diagnostic rather than separately validated: the study lacks independent gold evidence-state annotations for model outputs. We cannot separately estimate state-recognition and policy-mapping accuracy, demonstrate a dominant error source, or infer internal reasoning from an emitted rationale. R05 and R13 may involve coupled evidence and policy errors.

No result was repaired after observation: no output, reference, scoring rule, or method was altered, and no Attempt 004 was created. This preserves the negative result while leaving expert validation, broader coverage, and redesigned mapping to future work.

Future work. Prospective extensions include independent expert reference annotation, more independent paper clusters, multiple scientifically meaningful task axes, expert-annotated or oracle evidence-state experiments, deterministic policy mapping, and multiple model families. These would require a separately frozen protocol and do not revise or add to the completed results.

10 CONCLUSION

The same scientific paper can require different extraction actions when the downstream requirement changes. In this paired diagnostic, changing only material subtype changes the frozen AI-assisted author-reference action for 6/8 papers. Direct reaches 16/24 E2E exact decisions and Full 14/24; additional structure did not consistently improve decisions in this testbed. R09 shows that an output can record a hard conflict and still select an inconsistent action. These observations motivate a stage-wise diagnostic perspective while leaving separate validation of evidence recognition and action mapping to future controlled studies.

AI USE STATEMENT

Generative AI tools assisted research planning, code generation and revision, literature discovery and screening, scientific evidence annotation, and manuscript drafting and editing. The authors checked retained literature citations and bibliographic metadata against primary proceedings, publisher, or canonical preprint records. The benchmark reference is explicitly an AI-assisted author reference rather than independent expert gold; it was frozen before formal model evaluation. The paper–task population, task definitions, reference labels, generation configurations, and scoring policy were frozen and preserved, and no output repair, selective rerun, post-hoc metric, or Attempt 004 was used to change the reported results. The four formal model evaluations reported in the paper are separately documented experiments, not manuscript-writing assistance. The authors reviewed the AI-assisted code, annotations, analyses, and text and remain responsible for the scientific claims, source verification, and final content of this work.

REPRODUCIBILITY STATEMENT

The main paper defines the controlled A1/A2/A3 tasks, paper–task readiness labels, evaluated pipelines, and denominator-aware metrics. The appendix records the complete task requirements and taxonomy, candidate and full-text gates, frozen-reference protocol, Attempt 003 generation settings, execution-failure policy, scoring definitions, per-paper patterns, and qualitative error taxonomy. All formal outputs and provenance artifacts are retained, including the 11/96 execution failures, and no failed output was backfilled. Code, task specifications, frozen reference artifacts, model outputs, and deterministic scoring/reporting utilities are planned for release with the submission materials; no public or anonymous repository URL is claimed in this version.

REFERENCES

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699, 2024. doi: 10.1162/tacl_a_00667. URL https://aclanthology.org/2024.tacl-1.38/.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/25f7be9694d7b32d5cc670927b8091e1-Abstract-Conference.html.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/f12b4df26344f3be803c06b555252efe-Paper-Conference.pdf.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nature Communications*, 15:1418, 2024. doi: 10.1038/s41467-024-45563-x. URL https://www.nature.com/articles/s41467-024-45563-x.

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 NAACL-HLT 2021*, pp. 4599–4610, 2021. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365/.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 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, 2024. doi: 10.18653/v1/2024.eacl-demo.16. URL https://aclanthology.org/2024.eacl-demo.16/.

Luca Foppiano, Guillaume Lambard, Toshiyuki Amagasa, and Masashi Ishii. Mining experimental data from materials science literature with large language models: An evaluation study. *Science and Technology of Advanced Materials: Methods*, 4(1):2356506, 2024. doi: 10.1080/27660400.2024.2356506. URL https://arxiv.org/abs/2401.11052.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of EMNLP 2023*, pp. 6465–6488, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

Satanu Ghosh, Neal Brodnik, Carolina Frey, Collin Holgate, Tresa Pollock, Samantha Daly, and Samuel Carton. Toward reliable ad-hoc scientific information extraction: A case study on two materials dataset. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15109–15123, 2024. doi: 10.18653/v1/2024.findings-acl.897. URL https://aclanthology.org/2024.findings-acl.897/.

Saba Imran and Debanjum Singh Solanky. ResearchQA: Benchmarking citation-grounded question-answering on scientific papers, 2026. URL https://arxiv.org/abs/2607.11074.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230-Abstract.html.

Xuye Liu, Yimu Wang, Peng Shi, Bo Xue, Xiangrui Ke, Songcheng Cai, Kath Choi, Di Wu, Freda Shi, and Krzysztof Czarnecki. LitTraceQA: A benchmark for multi-stage grounding and verification in scientific question answering, 2026. URL https://arxiv.org/abs/2608.07370. Work in progress.

Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. Do LLMs know when to NOT answer? investigating abstention abilities of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9329–9345, 2025. URL https://aclanthology.org/2025.coling-main.627/.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Martin P. Polak and Dane Morgan. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15:1569, 2024. doi: 10.1038/s41467-024-45914-8. URL https://www.nature.com/articles/s41467-024-45914-8.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of ACL 2018, Volume 2: Short Papers*, pp. 784–789, 2018. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124/.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. Language agents achieve superhuman synthesis of scientific knowledge, 2024. URL https://arxiv.org/abs/2409.13740.

Akshay Swaminathan, Ivan Lopez, William Wang, Ujwal Srivastava, Edward Tran, Aarohi Bhargava-Shah, Janet Y. Wu, Alexander L. Ren, Kaitlin Caoili, Brandon Bui, Layth Alkhani, Susan Lee, Nathan Mohit, Noel Seo, Nicholas Macedo, Winson Cheng, Charles Liu, Reena Thomas, Jonathan H. Chen, and Olivier Gevaert. Selective prediction for extracting unstructured clinical data. *Journal of the American Medical Informatics Association*, 31(1):188–197, 2024. doi: 10.1093/jamia/ocad182. URL https://academic.oup.com/jamia/article/31/1/188/7285661.

Bingbing Wen, Bill Howe, and Lucy Lu Wang. Characterizing LLM abstention behavior in science QA with context perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3437–3450, 2024. doi: 10.18653/v1/2024.findings-emnlp.197. URL https://aclanthology.org/2024.findings-emnlp.197/.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025. doi: 10.1162/tacl_a_00754. URL https://aclanthology.org/2025.tacl-1.26/.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of ACL-IJCNLP 2021*, pp. 1040–1051, 2021. doi: 10.18653/v1/2021.acl-long.84. URL https://aclanthology.org/2021.acl-long.84/.

Hantian Zhang and Wentai Wu. Do LLMs know when evidence is insufficient? an evidence sufficiency benchmark for answer-abstention calibration in retrieval-augmented generation. *Computers, Materials & Continua*, 89(1):69, 2026. doi: 10.32604/cmc.2026.086343. URL https://www.techscience.com/cmc/v89n1/68467/html.

A Full Task Specifications

Each unit pairs one fixed scientific paper with one of three extraction tasks. All tasks require Ti–6Al–4V produced by laser powder bed fusion, a post-build heat-treatment temperature and holding time, heat-treatment atmosphere and cooling, build orientation, tensile-test temperature, ultimate tensile strength (UTS), the statistic represented by the UTS value, and a supported chain connecting material, build/specimen group, heat treatment, and tensile result. Only the subtype requirement changes: A1 requires non-ELI-compatible Ti–6Al–4V, A2 requires ELI, and A3 imposes no subtype constraint. Grade 5 or explicit non-ELI evidence is compatible with A1 and conflicts with A2; Grade 23 or explicit ELI evidence is compatible with A2 and conflicts with A1; plain Ti–6Al–4V leaves a constrained subtype unresolved. A3 requires only base-alloy identity.

Task evidence must apply to the evaluated specimen or to an explicitly documented lineage. A field reported for a different material, specimen group, heat-treatment branch, orientation, or testing condition cannot complete the target chain. Task context is not source evidence: a requested value may not be copied into the record merely because it appears in the task specification.

B Readiness Taxonomy

`READY` requires supported task-required assertions and a continuous task-valid evidence chain with no unresolved hard scope conflict. `REVIEW_REQUIRED` retains a relevant paper when material incompleteness, ambiguity, or linkage uncertainty prevents unreviewed extraction but could be resolved through human adjudication, a clarified task policy, or additional source material. `ABSTAIN` applies when supplied evidence is hard-incompatible with the requested scope or cannot support the extraction without speculation. These labels describe a paper–task pair, not paper quality.

The reference distinguishes *missing* evidence (a required item is not supported), *ambiguous* evidence (multiple task-relevant interpretations remain), *conflicting* evidence (the source contradicts an active hard requirement), and *unsupported* completion (a proposed record asserts more than the source). Missing or ambiguous evidence generally motivates review; explicit subtype incompatibility motivates abstention. The current reference is a frozen task policy, not universal scientific truth.

C Candidate and Full-Text Gates

Metadata-level discovery identified papers potentially concerning Ti–6Al–4V, L-PBF, post-build heat treatment, and tensile outcomes. A metadata relevance screen determined whether acquisition was warranted but did not assign readiness. Manually acquired documents then passed a full-text gate requiring the target material family, manufacturing route, post-build heat treatment, and tensile endpoint. Acquisition failure remained operational and was never converted into `ABSTAIN`. Papers outside the full-text diagnostic scope were excluded before readiness labeling. The final diagnostic set comprises R01, R03, R05, R07, R08, R09, R10, and R13, each crossed with A1–A3.

D Formal Generation Protocol

The frozen formal design schedules $8 \times 3 \times 4 = 96$ calls. Direct, Schema-only, Evidence-grounded, and Full receive identical paper–task evidence packages and share a logical output envelope; only Evidence-grounded and Full enforce comparable field-linked evidence bindings. Attempt 003 requested the provider model identifier `deepseek-v4-pro`, high reasoning, a 10,240-token maximum output budget, one completion, JSON-object response format, no tools or web access, and zero SDK automatic retries. The provider response did not expose an independently verifiable snapshot identifier, so no model snapshot is claimed. Temperature and top-p were not sent in reasoning mode. Earlier pre-scoring attempts diagnosed truncation and are retained only for provenance; none of their predictions backfills Attempt 003. Attempt 003 contains 85 valid outputs and 11 retained execution failures.

The author reference was frozen before formal generation. Generation components could not read reference labels or review artifacts; the scorer accessed the reference only after all 96 runs reached

Table 3: Safety-oriented and operational event counts. False and missed READY are counted among valid outputs; failures are counted among 24 scheduled units.

Method	False R	Missed R	Failure
Direct LLM	3	0	1
Schema-only	3	0	3
Evidence-grounded	0	1	6
Full	2	2	1

Table 4: Complete-paper transition diagnostics. Eligibility requires valid A1, A2, and A3 outputs for a paper; denominators therefore differ by method.

Method	Complete	Exact patterns	Sens. detected	False inv.	False sens.
Direct LLM	7	1/7	4/6	2	1
Schema-only	5	0/5	3/5	2	0
Evidence-grounded	4	2/4	1/2	1	0
Full	7	1/7	2/5	3	1

terminal status. No output was repaired, no task or reference label was modified, and no Attempt 004 was created after results were observed.

E Scoring Policy

Coverage is valid formal outputs divided by 24 scheduled units per method. End-to-end exact readiness success is a valid exact label match divided by all 24 scheduled units, so execution failure is unsuccessful but is not converted into a scientific label. Conditional exact agreement divides exact matches by method-valid outputs and is always reported with coverage. Common-support exact agreement uses the 14 units valid for all four methods.

False READY is a valid READY prediction when the reference is not READY; missed READY is a valid non-READY prediction when the reference is READY. Execution failures remain separate. No composite safety score is defined. Paper-level transition scoring requires valid A1, A2, and A3 outputs for that method and paper. The 24 units are dependent within eight paper clusters; the analysis is descriptive and makes no statistical-superiority claim.

F Per-Paper Prediction Patterns

Table 5 reports A1/A2/A3 patterns; FAIL preserves execution failure. R, RR, and A abbreviate READY, REVIEW_REQUIRED, and ABSTAIN.

G Error Taxonomy and Qualitative Details

The post-results audit used overlapping, non-additive tags for subtype-conditioning errors, under-abstention, over-abstention, review-versus-ready boundary errors, review-versus-abstain boundary errors, task invariance, false sensitivity, unsupported resolution, and execution failure. Full has nine valid incorrect decisions and one execution failure. Its overlapping counts include six subtype-conditioning errors, five under-abstention errors, four review-versus-ready boundary errors, five review-versus-abstain boundary errors, and three task-invariance errors; these totals must not be summed.

R09 provides a local recognition–action inconsistency. The reference is A/RR/RR: Grade 23/ELI conflicts with A1, while unsupported tensile-test temperature retains review for A2/A3. Full explicitly records the A1 hard conflict but predicts RR/RR/RR. This output is consistent with a mapping failure under the frozen author policy; it does not establish correctness of the entire evidence state or reveal internal reasoning. R05-A3 is an over-conservative error: Full elevates atmosphere, cooling, and applicability concerns beyond the frozen task boundary and downgrades reference R to RR.

Table 5: Frozen reference and method patterns in A1/A2/A3 order.

Paper	Reference	Direct	Schema	Evidence	Full
R01	RR/A/RR	RR/RR/RR	RR/RR/RR	RR/RR/RR	RR/RR/FAIL
R03	RR/RR/RR	RR/RR/FAIL	RR/RR/FAIL	RR/RR/RR	RR/RR/RR
R05	RR/RR/R	R/RR/R	R/RR/R	FAIL/FAIL/R	R/RR/RR
R07	A/R/R	RR/R/R	RR/R/FAIL	RR/R/RR	RR/RR/R
R08	A/RR/RR	RR/RR/R	RR/R/RR	RR/RR/FAIL	RR/RR/RR
R09	A/RR/RR	A/RR/RR	RR/RR/RR	RR/RR/FAIL	RR/RR/RR
R10	A/RR/RR	RR/RR/RR	RR/RR/R	FAIL/FAIL/RR	RR/RR/RR
R13	RR/RR/RR	RR/RR/R	RR/RR/FAIL	RR/RR/RR	RR/RR/R

R13-A3 is the opposite error: Full accepts a highest reported UTS as resolving statistic semantics and upgrades reference RR to R. These selected errors do not show a uniform conservative bias. Evidence interpretation and policy application may interact, and the study does not independently evaluate their accuracies or establish a dominant error source.

H Acquisition and Exposure Policy

Formal generation used only the frozen case inputs and permitted paper evidence. Reference labels, annotation-review artifacts, and scorer outputs were prohibited from the generation context. No web or tools were available to the evaluated model. Manual acquisition and metadata screening were documented separately from readiness. The formal boundary prohibits output repair, selective reruns, Attempt 004, reference modification, scoring-policy modification, and post-hoc method tuning.