PRESERVING EXPOSURE AND OUTCOME CONTEXT IN AUTOMATED CANCER EVIDENCE SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific evidence verification must preserve a claim's chemical, exposure setting, species, cancer site, endpoint, and evidence type. We study whether a deterministic toxicology scope gate transfers to general biomedical claim verification. Motivated by solvent-cancer synthesis, we combine a retrospective retrieval audit with a frozen experiment on 662 SciFact claim-abstract pairs from the public development split and an original-test citation subset. A lexical field extractor gates a TF-IDF verifier; linked source families are excluded from training, and policies receive identical complete stored abstracts. Strict compatibility retains no substantive answers because the extractor recognizes no chemical values. Mismatch-only filtering leaves development support decisions unchanged and removes two correct supports from one source family in the secondary subset, without removing any reference support-errors. At 89 accepted supports, confidence filtering retains 34 correct decisions versus 32 for mismatch filtering. Separately, we reanalyze published TCE-metabolite transcript estimates across 22 matched-dose pairs in five human cell models. At 62.5 micromolar, both metabolites induce large renal NRF2-associated transcript responses, whereas two neuronal models respond much more strongly to the cysteine conjugate, consistent with the source study. We release frozen protocols, predictions, probe-level estimates, and sensitivity analyses. The results locate the observed verification failures in missing fields and erroneous extraction; the biological analysis supports hypothesis development without measuring malignant transformation or establishing a new cancer mechanism.

1 INTRODUCTION

An answer to whether solvents in cleaning products contribute to cancer must preserve several distinctions. A chemical's carcinogenic hazard does not specify the risk associated with a particular product, dose, route, duration, or population. Occupational exposure estimates cannot simply be assigned to household use. A cellular stress response and a cancer diagnosis are different outcomes. These distinctions also create a concrete problem for automated scientific synthesis: a relevant citation may still support a different claim from the one a system presents.

Our motivating comparison spans kidney cancer, bladder cancer, breast cancer, and glioma. The National Cancer Institute identifies kidney cancer as an outcome of prolonged or repeated trichloroethylene exposure (National Cancer Institute, 2024). An occupational solvent analysis in INTEROCC did not find a consistent association with glioma (Benke et al., 2017), whereas a newer benzene meta-analysis reports a positive but uncertain glioma subgroup estimate (Mangiaterra et al., 2025). These results concern different exposure definitions and do not resolve risk from an unspecified household cleaning mixture. We ask whether a reproducible evidence workflow can make such changes in scope explicit before a generated answer is accepted.

The present work combines a retrospective pipeline audit with a frozen external claim-verification experiment. The archived artifacts originated in glioma-focused and mechanism-focused searches. We identify failures in provenance, counting, and interpretation, then evaluate the compatibility check as a gate on a simple learned text verifier using published relation labels. The transfer experiment yields a negative result: strict completeness prevents substantive answers, while a mismatch-only policy fails to improve support decisions. A separate reanalysis of published human-cell mea-

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.

Table 1: Selected evidence contexts. Estimates are from the cited studies, not from the retrieval corpus. OR: odds ratio; meta-RR: meta-analysis relative risk. Parentheses give 95% confidence intervals.

Cancer site	Exposure and source	Interpretation boundary
Kidney	Prolonged or repeated TCE exposure; NCI assessment (National Cancer Institute, 2024).	Established hazard does not supply a product-specific household dose or risk.
Bladder	Dry-cleaning employment: meta-RR 1.47 (1.16–1.85); PCE-specific analysis: 1.08 (0.82–1.42) (Vlaanderen et al., 2014).	The analyses use different study sets and exposure definitions. Job title is not measured PCE exposure.
Breast	Highest versus lowest self-reported cleaning-product use quartile: OR 2.1 (1.4–3.3) (Zota et al., 2010).	Retrospective reporting and recall bias limit interpretation; no individual causal solvent is identified.
Glioma	Occupational solvents: OR 0.91 (0.74–1.11), INTEROCC (Benke et al., 2017); benzene meta-analysis subgroup: RR 3.88 (1.33–11.31) (Mangiaterra et al., 2025).	The benzene subgroup has only two risk estimates; the meta-analysis reports publication-bias concerns. Neither study establishes household-mixture risk.

surements grounds the biological follow-up in observed responses, while distinguishing those responses from hypotheses about cancer formation.

The contributions are (1) a replayable audit of two archived retrieval collections; (2) processing corrections and a structured compatibility check with explicit missingness and provenance; (3) an external experiment with a frozen text adapter, shared verifier, source-family exclusions, confidence-matched controls, and case-level predictions; and (4) a descriptive TCE-metabolite comparison with a probe-level ledger and explicit limits on mechanistic interpretation. The experiment distinguishes software conformance from useful verification and identifies concrete limitations of the current adapter. Scope equality does not assess study quality or infer causality.

2 Scientific context and related work

Scientific claim verification already distinguishes support, refutation, and insufficient evidence. SciFact pairs expert-written claims with evidence-containing abstracts and rationales (Wadden et al., 2020). Scientific literature agents such as PaperQA2 study retrieval and synthesis over source documents (Skarlinski et al., 2024). SciFact-Open also identifies cases in which evidence supports only a restricted instance of a broader claim (Wadden et al., 2022a). CliVER uses clinical claims encoded by population, intervention, comparison, and outcome (PICO), and identifies errors involving missing claim–evidence alignment (Liu et al., 2024). Scientific-text pretraining in SciBERT (Beltagy et al., 2019) and the joint label/rationale predictions of MultiVerS (Wadden et al., 2022b) offer stronger learned alternatives to a lexical verifier. SciBERT is an encoder requiring task adaptation; MultiVerS directly addresses scientific verification. Neither is evaluated here. These precedents already address context in scientific verification. We study a toxicology-specific compatibility baseline with explicit exposure setting, species, cancer site, and evidence type; we test its portability with an explicit text adapter and report the limits of that experiment.

Table 1 gives selected reference examples. It is not a systematic review or a ranking of cancer risks. The studies use different exposures, populations, and designs; their effect estimates must not be pooled across rows. They motivate the schema and will require independent review before becoming evaluation labels.

Positive associations retain their exposure definitions. The benzene meta-analysis does not support a uniformly null account of solvent–glioma evidence, but its authors do not interpret the association as causal (Mangiaterra et al., 2025). A newer population-based case–control study of occupational benzene, toluene, and xylene (BTX) reported OR 2.23 (95% CI 1.35–3.69) for the highest cumulative hybrid-exposure quartile versus unexposed participants (Xie et al., 2024). Correlated ex-

posures prevent assigning the mixture estimate to each chemical separately. These published results motivate context-preserving verification; they are not new biological findings of this project.

Environmental chemistry is a separate evidence layer. HOMEChem measurements characterize indoor chlorinated and nitrogenated compounds during bleach cleaning (Mattila et al., 2020). They motivate exposure assessment, but they do not measure malignant transformation or cancer incidence. A list of radical-related terms and carcinogen-related terms across papers is also not a demonstrated sequence of chemical reactions. Establishing a connected mechanism requires evidence for each transformation and its conditions, followed by an assessment of biological relevance.

LUCIR supplies a direct household example: personal air monitoring in 50 women before and after product substitution found decreases in 17 measured chemicals and increases in three fragrance compounds (Harley et al., 2021). The measured endpoint was breathing-zone chemical concentration. This study did not measure a change in cancer incidence or cell transformation, and its product substitution should not be described as a tested cancer-prevention intervention.

An out-of-domain warning example. A supplied screenshot displays a Proposition 65 warning about acrylamide and furfuryl alcohol in baked goods. We retain only a paraphrase as a development example; seller, formulation, concentrations, and warning date are unverified. The official chemical entries list acrylamide for cancer, developmental toxicity, and male reproductive toxicity, whereas the furfuryl-alcohol entry lists cancer (Office of Environmental Health Hazard Assessment, 2026a;b). Thus a combined warning cannot assign every named hazard to each chemical. The screenshot is evidence that a warning was displayed; it is neither a primary cancer study nor evidence about cleaning-product exposure. We make no determination of the warning's current legal compliance.

3 Materials and audit design

Archived collections. The glioma collection contains 68 retrieved paper records, their historical grades, query traces, and six inspection responses. Its recorded retrieval date is August 24, 2026. The mechanism expansion comprises two September 1, 2026 searches: eight queries on solvent-related reactive oxygen species and oxidative-stress mechanisms, followed by ten queries targeting indoor bleach chemistry, TCE mitochondrial/renal injury, and cleaner-related cancer or biomarker studies. The first retained 95 derived records; the second retained 146 paper records. These selected searches are separate from the glioma collection. The first expansion did not preserve raw abstracts; the second saved 145 nonempty abstract fields, all at most 280 characters long, with ellipses in 106. Its saving code imposed an additional 900-character cap, which none of these fields reached. The audit analyzes these stored artifacts. A separate bounded Firecrawl access attempt returned HTTP 403 before source discovery, and supplied no new papers or passages. A live index may return different documents, and replay does not establish completeness of the literature.

Provenance and identity. We retain the original JSON files and compute SHA-256 digests for the seven audit inputs. A frozen copy of the historical tagging function reproduces all 68 archived grades. New outputs are stored separately. For the expansion, records are merged transitively if they share a normalized identifier or an exact normalized title. Title normalization lowercases text and removes all characters outside ASCII letters and digits; it is a disclosed fallback, not fuzzy bibliographic resolution. We keep aliases, source-query identifiers, source-file identifiers, and the longest available abstract. This rule can still merge different works with identical titles and cannot identify duplicate cohorts without further review.

Source availability. An inspection request is counted separately from returned source passages. Only explicit text entries in a successful provider passage response are attached to the matching paper. A generated answer, search question, or error-body record cannot supply passage evidence. Provider passages have a source identifier and a locator when available; their attribution and completeness still require review. We mark missing and visibly abbreviated abstracts rather than treating them as full-text reads.

Analysis scope. Counts describe a selected retrieval archive, not a random sample of publications or people. Historical contingency-table tests of retrieval labels therefore do not estimate a chemical–cancer association. The archive contains no independent gold labels, complete effect-estimate extraction, or registered prospective evaluation. We make no claim that a fixed random seed or a subsequent analysis plan establishes preregistration.

4 A WORKFLOW THAT PRESERVES CLAIM CONTEXT

4.1 RETRIEVAL AND CONSERVATIVE TAGGING

The optional live pipeline searches an index, selects seeds, requests source passages, expands similar papers and citations, merges records, and exports citations with provenance. The revised query plan separately targets the four cancer sites and environmental chemistry. It was introduced after the archived collections and did not produce their reported counts. New runs use fresh directories, retain returned paper records, and distinguish errors, passage requests, and attached passages.

The tagger reads titles, available abstracts, and attached source passages. It excludes query text and generated answers. Explicit cancer terms are recorded separately from tissue injury. Result-cue rules check negation before positive cues and exclude recognizable objective sentences; an odds-ratio mention alone does not establish a positive result. Multiple recognized directions produce a mixed label, while absent cues produce an unclear label. These are lexical cues, not validated study-level conclusions. They can misassign a result to the wrong exposure or subgroup and can miss negation beyond the matching span. Study quality is unassessed.

4.2 STRUCTURED SCOPE COMPATIBILITY

Let a claim c and a source card e contain fields

$$F = \{\text{chemical, exposure setting, species, cancer site, endpoint, evidence type}\}.$$

Additional constraints on direction, route, dose, and duration are checked when the claim supplies them. This is an operational baseline choice, not a judgment that these exposure dimensions are scientifically optional. A structured claim card that specifies a dose or route must match it; a claim that omits them cannot support a dose- or route-specific conclusion. A card also carries a source URL or locator, a paraphrased result, optional effect estimates, limitations, and review status. Missing information is represented explicitly; "not applicable" is distinct from "unknown."

For each required field, the implementation uses equality between canonical scalar values. It returns *scope mismatch* if any known value differs, *insufficient metadata* if no mismatch is known but a required value or provenance is missing, and *compatible scope* otherwise. A mismatch takes precedence over missingness, and both remain visible in the diagnostic output. Compatibility is a necessary screening condition, not proof that the claim is true. Synonym normalization, hierarchical relations between cancer sites, quantitative dose comparisons, evidence quality, and causal identification are outside this baseline.

For example, a PCE-specific occupational card cannot support a household-mixture statement merely because both concern solvents. An indoor-air measurement cannot support a cancer-incidence statement merely because it names a chemical with carcinogenic potential. Where the relevant fields match, the source still needs a valid result, appropriate design, and a defensible interpretation. The system therefore records causal assessment as *not performed*, rather than treating a preset negative conclusion as an evaluated refusal.

5 AUDIT OBSERVATIONS

Association labels and missing source text. The historical tagger marked 26 of the 68 records as reporting an association. Its triggers included generic association language and odds-ratio terminology. One labelled title explicitly reports a lack of association between glioma and selected occupational agents. The revised cue rules yield five no-clear-association labels and 63 unclear labels. An additional regression check removed a positive cue originating in an abbreviated abstract headed `OBJECTIVES:`. That heading had escaped the initial singular-only exclusion rule. This

Table 2: Deterministic observations from archived files. Tag counts are algorithm outputs; they do not measure scientific truth, cancer risk, or benchmark accuracy.

Observation	Count
Original glioma-focused records	68
Historical glioma / household / joint tags	24 / 6 / 0
Original abstracts containing ellipses	41
Inspection requests / distinct seeds	6 / 3
Attached source passages from those requests	0
Mechanism expansion source rows	241
Records after identity and title merging	238
Duplicate rows removed	3
Merged records with / without an abstract	145 / 93
Joint indoor–radical–product lexical category	0

change demonstrates sensitivity to implementation choices, not improved accuracy: there are no independent annotations against which to score either version. It would be incorrect to interpret these counts as five verified null studies or evidence that positive findings do not exist.

All six archived inspection responses contain empty passage arrays. Moreover, 41 stored abstracts contain ellipses, which signal possible abbreviation but do not establish exactly how much source text was lost. Thus the archive does not substantiate a claim that passage-level verification was completed. New source-backed reference cards were assembled separately during this revision and must not be retroactively attributed to the old retrieval loop.

Identity and mechanistic interpretation. Merging the two expansion files collapses three pairs of records sharing an explicit PMID: 10566590, 18441786, and 26649641. The result is 238 records, not 241 unique publications. An identifier-only sensitivity check produces the identical partition: all three pairs share an explicit PMID, and title matching introduces no additional links. No component has more than two records or conflicting identifiers within an observed namespace. This checks these archived groups, not external bibliographic ground truth; title-based over-merging remains a general risk. Under the current mutually exclusive lexical buckets, 72 concern general key-characteristic or radical terminology, 20 indoor radical terminology, 30 selected product-class terminology, and 116 fall outside those categories. No record enters the joint indoor–radical–product category. These classifications are noisy descriptions of available text. A zero category count neither disproves the underlying chemistry nor establishes that no relevant publication exists; it prevents this particular output from being presented as recovery of a connected mechanism.

Development conformance. Ten source-linked cards generate 24 cases authored during implementation: ten compatible-scope examples, twelve scope changes, and two missing-metadata examples. The deterministic checker matches all expected statuses. Because the same development process supplied the rules and examples, this is software conformance, not held-out scientific accuracy. The collection contains no independently adjudicated labels and evaluates no language model. Appendix B reports the simple comparisons to make their restricted meaning explicit.

6 External verification experiment

Task, data, and frozen methods. We evaluate fixed claim–abstract pairs using SciFact's original relation labels, not invented six-field compatibility labels (Wadden et al., 2020). The first set contains all 300 public-development claims and their 339 cited-source pairs. The second contains 279 original-test claims recovered through SciFact-Open and their 323 cited-source pairs, using historical citation-provenance labels (Wadden et al., 2022a). All secondary claim texts match the original test release; 21 unrecovered test claims are excluded by this source availability, not by performance. These are neither official full-test scores nor open-domain retrieval evaluations. Seventy-four source/claim families span both sets, so they are not independent replications.

One TF-IDF logistic verifier is fitted on 384 training claims (435 pairs) after excluding 425 of 809 training claims whose transitive source/claim families overlap either evaluation set. Features, vocab-

Table 3: External support-selection results. Development has 138 reference SUPPORT pairs; the secondary subset has 107. Risk and coverage are percentages. A dash denotes undefined risk at zero acceptance. Strict and always-abstain have the same support counts; their other output labels differ.

Set	Policy	Accepted	Errors	Risk	Coverage
Dev	Base	100	44	44.00	40.58
Dev	Mismatch-only	100	44	44.00	40.58
Dev	Confidence at 100	100	44	44.00	40.58
Dev	Strict / always-abstain	0	0	—	0.00
Test subset	Base	91	57	62.64	31.78
Test subset	Mismatch-only	89	57	64.04	29.91
Test subset	Confidence at 89	89	55	61.80	31.78
Test subset	Strict / always-abstain	0	0	—	0.00

ulary, and inverse-document frequencies use retained training inputs only. All policies receive the same claim and complete stored source title/abstract; no gold rationale spans or extracted gold fields are supplied. The classifier, narrow lexical text adapter, protocol, and input hashes were committed before first inference. Public schemas and one reference-label discrepancy were inspected before that freeze; this is an auditable ordering record, not preregistration or assurance against public-data exposure during AI pretraining.

Text adapter and component scope. The adapter applies case-insensitive regular-expression dictionaries independently to each claim and full source title/abstract. A field receives a scalar value only when exactly one category is recognized; zero or multiple categories produce unknown. Its chemical vocabulary contains TCE, PCE, benzene, toluene, xylene, formaldehyde, acrylamide, and furfuryl alcohol. It performs no learned entity recognition or claim-to-source field copying. The adapter emits only the six core fields; extraction of the conditional exposure dimensions is not evaluated in this transfer experiment. The documented "adult"-to-human rule remains in the frozen run. Appendix D audits coverage of every field. Policy comparisons hold extracted fields and verifier outputs fixed; they do not measure the gate's utility under accurately extracted metadata.

The baseline predicts SUPPORT, CONTRADICT, or NOT_ENOUGH_INFO. The strict gate retains a substantive prediction only when all required scope fields are compatible; the separately named mismatch-only policy rejects detected mismatches but permits unknown fields. Confidence controls retain the highest-scored base SUPPORT predictions at the gate's support-acceptance count. This matching uses no reference labels and evaluates support selection only. An always-abstain policy supplies the zero-coverage comparison.

Measured outcomes. Table 3 reports support decisions. Reference-error risk is accepted non-SUPPORT reference labels divided by accepted SUPPORT outputs; supported coverage is retained correct SUPPORT outputs divided by all reference SUPPORT pairs. Errors refer to benchmark annotations, not definitive scientific falsehood. The base verifier's three-way macro-F1 is 0.4547 on development and 0.4159 on the secondary set, so this modest baseline cannot stand in for contemporary language agents.

The strict gate retains zero substantive answers in both sets; it preserves base NOT_ENOUGH_INFO predictions rather than literally returning ABSTAIN for every input. The eight-category chemical extractor recognizes no chemical value in any claim or source, making full compatibility impossible in this transfer setting regardless of the underlying verifier. This is zero recognition, not measured zero extraction recall: independent field labels and an annotated positive denominator are unavailable. The mismatch-only gate leaves development SUPPORT decisions unchanged. In the secondary set it removes two correct supports and no reference support-errors. Both lost pairs share one source: the adapter incorrectly maps "adult" in claims about tissue-resident macrophages to human species, while the source describes murine evidence. This post-run extraction failure is preserved in the frozen experiment.

At 89 accepted secondary supports, confidence selection retains 34 correct answers versus 32 for mismatch-only. The latter's risk is 2.25 percentage points higher (paired 95% interval 0.00 to 5.43); supported coverage is 1.87 points lower (interval -6.00 to 0.00). Intervals use 1,000 bootstrap

Table 4: NRF2-associated transcripts at 62.5 micromolar, 24 hours. Arm entries are median gene log2 fold changes versus medium control; the last column is the median of paired gene differences, which need not equal the difference of arm medians. No new significance tests are performed.

Model	Genes	DCVC	DCVG	Paired difference
HUVEC/TERT2	6	0.909	0.452	0.506
HepaRG	6	0.812	0.628	0.326
RPTEC/TERT1	6	2.860	2.802	0.157
LUHMES	5	2.811	0.062	2.881
BCC42	6	2.353	0.126	2.564

resamples of connected source/claim families, with policy decisions fixed and shared resamples for comparisons. Their inclusion of zero and the single-family failure prevent a claim of statistically established superiority. These observed results provide no evidence of a benefit from the present gate and adapter.

Post-feedback component diagnostic. A subsequent descriptive audit of the unchanged prediction ledger finds zero complete pairs even after omitting the chemical requirement: 0/339 development and 0/323 secondary pairs. Missingness in other fields therefore remains a sufficient obstacle. With the retained fields fixed, any verifier's substantive acceptance is bounded by the number of compatible pairs, which is zero. This does not evaluate a stronger extractor, supply oracle fields, or establish the gate's utility under accurate metadata. Appendix D reports every field's coverage.

Reference-version sensitivity and scope. One cited pair has a newer human pooled SUPPORT annotation but no historical citation annotation. A predeclared sensitivity changes only that reference label, with all predictions fixed. Support-error risks are unchanged; the secondary SUPPORT denominator increases from 107 to 108, giving coverage 31.48% for base/confidence and 29.63% for mismatch-only. Pooled model-generated rationale highlights never enter inference. This illustrates that a reference NOT_ENOUGH_INFO label need not imply absence of scientific support. The experiment supplies independent published relation labels, not independent toxicology-field annotations, a systematic cancer review, or proof that structured verification in general fails.

7 Biological measurements and discriminating hypotheses

Published human-cell data. We reanalyze the complete processed expression tables of Capinha et al. (2023), covering 63 available 1,2-DCVC/1,2-DCVG compound–dose conditions in six human cell models. DCVC is the cysteine conjugate and DCVG its upstream glutathione conjugate. The source study exposed cells for 24 hours and selected transcriptomics concentrations for at most 25% viability loss. We freeze two selected six-gene panels and the calculation before deriving new summaries, after reading the source study and schemas. This is exploratory secondary analysis, not preregistration or independent replication. The tables provide precomputed DESeq2 estimates, not raw replicate counts.

Within-model dose matching with 1% relative tolerance gives 22 pairs across five models; UKN5 lacks DCVG. Eligible probes have finite log2 fold changes and baseMean > 35 in both arms. We intersect probe IDs, take within-gene medians, then summarize the paired gene differences. Table 4 shows the only concentration exactly shared across all five paired models. Both renal arm responses are large: a small DCVC–DCVG difference does not mean little renal response. In LUHMES and BCC42, DCVG responses are much weaker than DCVC responses. This is consistent with the source authors' tissue-dependent metabolite-activation findings, not a newly discovered mechanism. Culture conditions, intracellular exposure, remaining viability, and gene coverage differ; these contrasts do not rank organ cancer susceptibility. Appendix C reports every trajectory and the sensitivities without abundance filtering and with exact dose matching.

Selection of pre-existing mutant cells. A separate, more specific cancer-promotion hypothesis concerns benzene metabolites and blood stem cells. Guo et al. (2019) found that TET2 knockdown reduced benzoquinone-induced AIM2/CASP1-pathway activation and toxic cell death in immortal-

ized human B cells. We hypothesize that this pathway difference could give pre-existing TET2-deficient primary hematopoietic stem cells an exposure-dependent fitness advantage. The B-cell experiment did not test this competition. Human occupational evidence is suggestive but limited: He et al. (2026) report clonal hematopoiesis in 12/103 asymptomatic exposed workers versus 2/44 controls (unadjusted OR 2.3, 95% CI 0.66–12.28); the cross-sectional design cannot resolve selection versus mutation induction. Clonal hematopoiesis is not a cancer diagnosis.

A discriminating test must measure baseline genotype, absolute surviving stem-cell numbers, subsequent lineage contribution, and pyroptosis. Preferential wild-type death is selection even without absolute mutant expansion. Genotype-independent stochastic attrition is a competing explanation: endogenous aldehyde damage can produce clonal dominance through stem-cell loss (Kamimae-Lanning et al., 2026). An exposure-dependent genotype advantage linked to the death pathway, and removed by correcting that pathway difference, would support the proposed mechanism. Benzene-associated competitive contribution in preleukemic Mll–Af9 mice is already reported (Zhou et al., 2025); novelty of the TET2-specific hypothesis is unestablished.

Temporary impairment of DNA repair. Kong et al. (2024) demonstrate that endogenous methylglyoxal can transiently deplete BRCA2 protein and produce episodic mutations without a permanent second gene hit. This is published prior work. A TCE-metabolite $\rightarrow$ carbonyl burden $\rightarrow$ BRCA2 impairment bridge is an additional, more speculative hypothesis. It predicts carbonyl accumulation before protein loss and repair failure, with persistent mutations in surviving lineages after repair recovers. Reducing carbonyl burden or preserving BRCA2 function should reduce downstream defects if this mediation is correct. Direct genotoxicity and mitochondrial injury are alternatives. Our transcript tables lack identifiable BRCA2 probes and do not measure carbonyls, repair function, mutant-clone fitness, or transformation; they cannot establish either proposed causal chain.

8 Research needed to test the proposal

Independent machine-learning evaluation. A useful next study would test whether context-preserving synthesis reduces unsupported transfers while retaining supported answers. Before accessing a held-out test set, a protocol should fix the source selection, annotation rules, models, prompts, retrieval budgets, primary metrics, and exclusions. Two independent domain-qualified annotators should extract source spans and fields, label claim compatibility and support separately, and adjudicate disagreements. Positive, null, inverse, mixed, and insufficient evidence should be represented across sites and exposure settings. Split at the source and cohort family level; reserve chemical families to test transfer. Development examples in this paper must stay out of the test set.

Compare unrestricted synthesis, citation-based retrieval, structured extraction with the compatibility check, and an always-abstain baseline. For a fixed-evidence comparison, systems must receive the same documents; evaluate retrieval quality separately. Report source-attribution accuracy, field extraction, unsupported claims, supported-answer coverage, and abstention together. A system that rejects every claim cannot demonstrate useful synthesis. Use uncertainty intervals clustered by source family and inspect failure types, rather than testing independence of tags in a search-selected corpus. The external experiment above implements a narrower fixed-evidence comparison with a simple verifier; the proposed domain-specific and contemporary-agent comparisons remain separate research.

9 Limitations and conclusion

This study audits selected archives and evaluates a particular scope gate with a narrow extractor and simple learned verifier. Its external relation labels address the previous reliance on constructed cases, but do not provide independent toxicology-field labels or contemporary-agent comparisons. The secondary set is incomplete and correlated with development through shared sources; document identities do not resolve every shared clinical cohort. Search bias, incomplete abstracts, title-based merging, extraction errors, and AI-assisted source-card construction limit interpretation. The work is not a systematic review or new epidemiological analysis. The secondary expression analysis uses previously processed estimates with no new biological replication; correlated genes and doses cannot supply independent samples. The proposed selection and repair hypotheses are not observed cancer

outcomes. Null estimates do not establish safety, and mechanistic relevance does not establish a clinical causal effect.

The measured conclusion is that the implemented completeness rule and narrow lexical adapter do not improve external support selection: strict checking eliminates substantive answers, while the relaxed policy can discard supported claims because of extraction errors. A useful successor must preserve missingness without confusing it with contradiction, measure extraction separately, and outperform confidence-based selection at useful coverage. The released protocols, predictions, and uncertainty estimates make that claim testable. The biological reanalysis makes metabolite/model differences auditable while identifying missing measurements needed to test mutant-cell selection and temporary repair impairment. Neither the expression differences nor the literature connections establish a new cancer mechanism or household-exposure risk.

AI USE STATEMENT

Generative AI tools, including Cursor, OpenAI Codex, and Google's Paper Assistant Tool, assisted with research framing, literature search, methodological critique, implementation, data reformatting, interpretation of audit outputs, development-example construction, and manuscript drafting. Development source cards were prepared with AI assistance. The 24 constructed development cases and their expected labels are AI-assisted synthetic examples grounded in these cards; they have not undergone independent expert adjudication. External verification reference labels originate in published human annotations; AI assistance was used for the adapter, experiment implementation, methodological review, and interpretation. Automated regression and deterministic replay checks verify specified software behavior; they do not validate scientific conclusions. A second AI agent independently checked every extracted expression record and all primary and sensitivity panel calculations against the source workbooks; this is computational verification, not independent biological or human expert validation. Automated pre-submission feedback informed additional diagnostic checks and presentation corrections. Independent scientific review and final author verification remain pending in this draft. The external experiment evaluates the specified TF-IDF verifier and gate, not the performance of Cursor or Codex as research agents.

ETHICS STATEMENT

The reported work uses archived bibliographic records, published sources and processed cell-model expression estimates, and a paraphrased warning example. It includes no new participant recruitment, patient-level analysis, or biological experiments. Exposure context and uncertainty are retained to reduce unsupported health claims. No product-specific risk estimate or personal medical advice is provided. The identifying interface around the supplied warning is omitted.

REPRODUCIBILITY STATEMENT

The accompanying repository contains frozen input artifacts, input digests, the historical tagger, revised processing code, source-linked development cards, expected cases, and deterministic audit outputs. Appendix A describes offline replay. The external benchmark additionally includes frozen method/data hashes, case-level predictions, reference-version sensitivity, and clustered uncertainty. A full refit reproduced the saved predictions and report byte-for-byte in the recorded environment. Live retrieval is not required for replay; original public benchmark inputs can be downloaded with checksum verification. The biological analysis includes its pre-calculation method hashes, publisher URLs and checksums, 1,386 extracted probe estimates, full probe inventories, and all descriptive comparisons. A post-freeze gene-alias correction is disclosed in Appendix C.

REFERENCES

Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of EMNLP-IJCNLP*, pp. 3615–3620, 2019. doi: 10.18653/v1/D19-1371. URL https://aclanthology.org/D19-1371/.

Geza Benke, Michelle C Turner, Sarah Fleming, Jordi Figuerola, Laurel Kincl, Lesley Richardson, Maria Blettner, Martine Hours, Daniel Krewski, David McLean, Marie-Elise Parent, Siegal Sadetzki, Klaus Schlaefer, Brigitte Schlehofer, Jack Siemiatycki, Martie van Tongeren, and Elisabeth Cardis. Occupational solvent exposure and risk of glioma in the INTEROCC study. *British Journal of Cancer*, 117(8):1246–1254, 2017. doi: 10.1038/bjc.2017.285.

Liliana Capinha, Yaran Zhang, Anna-Katharina Holzer, Anna-Katharina Ückert, Melinda Zana, Giada Carta, Cormac Murphy, Jenna Baldovini, Zahra Mazidi, Johannes Grillari, Andras Dinnyes, Bob van de Water, Marcel Leist, Jan N. M. Commandeur, and Paul Jennings. Transcriptomic-based evaluation of trichloroethylene glutathione and cysteine conjugates demonstrate phenotype-dependent stress responses in a panel of human in vitro models. *Archives of Toxicology*, 97(2): 523–545, 2023. doi: 10.1007/s00204-022-03436-6.

Xiaoli Guo, Wen Zhong, Yujiao Chen, Wei Zhang, Jing Ren, and Ai Gao. Benzene metabolites trigger pyroptosis and contribute to haematotoxicity via TET2 directly regulating the Aim2/Casp1 pathway. *EBioMedicine*, 47:578–589, 2019. doi: 10.1016/j.ebiom.2019.08.056.

Kim G. Harley, Lucia Calderon, James E. S. Nolan, Randy Maddalena, Marion Russell, Katherine Roman, Stephanie Mayo-Burgos, Jessica Cabrera, Norma Morga, and Asa Bradman. Changes in latina women's exposure to cleaning chemicals associated with switching from conventional to "green" household cleaning products: The LUCIR intervention study. *Environmental Health Perspectives*, 129(9):097001, 2021. doi: 10.1289/EHP8831.

Jinxu He, Lin Zhang, Jianxin Zhao, Xiaohan Yang, Tongshuai Wang, Shiwei Cui, Xiurong Cheng, Xuejing Yang, Jin Zhou, Ming Li, Xue Wang, Yan Zhang, Ying Li, Jun Shi, Jiaping You, Qiang Jia, Zhao-lin Xia, Caihong Xing, and Qian-Fei Wang. Accelerated clonal hematopoiesis and premature hematopoietic aging in benzene-exposed populations. *American Journal of Hematology*, 101(7):1589–1592, 2026. doi: 10.1002/ajh.70324.

Ashley N. Kamimae-Lanning, Jill M. Brown, Matthias Günther, Franziska Esau, Holly Russell, Lise Larcher, Frédéric Langevin, Tomoya Isobe, Nicola K. Wilson, Felix A. Dingler, Rebecca L. Cordell, Meng Wang, Christopher L. Millington, Nina Claudino, Ewa Gogola, Matthew Nicholls, Verena Körber, Berthold Göttgens, Marella F.T.R. de Bruijn, Juan I. Garaycoechea, Jean Soulier, Thomas Höfer, and Ketan J. Patel. Metabolite-induced DNA damage drives stochastic stem cell loss and clonal hematopoiesis. *Cell Stem Cell*, 33(4):642–659.e11, 2026. doi: 10.1016/j.stem.2026.02.011.

Li Ren Kong, Komal Gupta, Andy Jialun Wu, David Perera, Roland Ivanyi-Nagy, Syed Moiz Ahmed, Tuan Zea Tan, Shawn Lu-Wen Tan, Alessandra Fuddin, Elayanambi Sundaramoorthy, Grace Shiqing Goh, Regina Tong Xin Wong, Ana S.H. Costa, Callum Oddy, Hannan Wong, C. Pawan K. Patro, Yun Suen Kho, Xiao Zi Huang, Joan Choo, Mona Shehata, Soo Chin Lee, Boon Cher Goh, Christian Frezza, Jason J. Pitt, and Ashok R. Venkitaraman. A glycolytic metabolite bypasses "two-hit" tumor suppression by BRCA2. *Cell*, 187(9):2269–2287.e16, 2024. doi: 10.1016/j.cell.2024.03.006.

Hao Liu, Ali Soroush, Jordan G. Nestor, Elizabeth Park, Betina Idnay, Yilu Fang, Jane Pan, Stan Liao, Marguerite Bernard, Yifan Peng, and Chunhua Weng. Retrieval augmented scientific claim verification. *JAMIA Open*, 7(1):ooae021, 2024. doi: 10.1093/jamiaopen/ooae021. URL `https://doi.org/10.1093/jamiaopen/ooae021`.

Silvia Mangiaterra, Paolo Boffetta, and Monireh Sadat Seyyedsalehi. Occupational benzene exposure and risk of nervous system cancers: A systematic review and meta-analysis. *Cancer Epidemiology*, 95:102779, 2025. doi: 10.1016/j.canep.2025.102779.

James M Mattila, Pascale S J Lakey, Manabu Shiraiwa, Chen Wang, Jonathan P D Abbatt, Caleb Arata, Allen H Goldstein, Laura Ampollini, Erin F Katz, Peter F DeCarlo, Shan Zhou, Tara F Kahan, Felipe J Cardoso-Saldaña, Lea Hildebrandt Ruiz, Andrew Abeleira, Erin K Boedicker, Marina E Vance, and Delphine K Farmer. Multiphase chemistry controls inorganic chlorinated and nitrogenated compounds in indoor air during bleach cleaning. *Environmental Science & Technology*, 54(3):1730–1739, 2020. doi: 10.1021/acs.est.9b05767.

National Cancer Institute. Trichloroethylene, 2024. URL https://www.cancer.gov/about-cancer/causes-prevention/risk/substances/trichloroethylene. Accessed September 5, 2026.

Office of Environmental Health Hazard Assessment. Acrylamide: Chemical listing, 2026a. URL https://oehha.ca.gov/chemicals/acrylamide. Chemical entry accessed September 5, 2026; year denotes access, not listing date.

Office of Environmental Health Hazard Assessment. Furfuryl alcohol: Chemical listing, 2026b. URL https://oehha.ca.gov/chemicals/furfuryl-alcohol. Chemical entry accessed September 5, 2026; year denotes access, not listing date.

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. *arXiv preprint arXiv:2409.13740*, 2024.

Jelle Vlaanderen, Kurt Straif, Avima Ruder, Aaron Blair, Johnni Hansen, Elsebeth Lynge, Barbara Charbotel, Dana Loomis, Timo Kauppinen, Pentti Kyyronen, Eero Pukkala, Elisabete Weiderpass, and Neela Guha. Tetrachloroethylene exposure and bladder cancer risk: a meta-analysis of dry-cleaning-worker studies. *Environmental Health Perspectives*, 122(7):661–666, 2014. doi: 10.1289/ehp.1307055.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of EMNLP*, pp. 7534–7550, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. SciFact-Open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4719–4734, 2022a. doi: 10.18653/v1/2022.findings-emnlp.347. URL https://aclanthology.org/2022.findings-emnlp.347/.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 61–76, 2022b. doi: 10.18653/v1/2022.findings-naacl.6. URL https://aclanthology.org/2022.findings-naacl.6/.

Shuai Xie, Melissa C. Friesen, Dalsu Baris, Molly Schwenn, Nathaniel Rothman, Alison Johnson, Margaret R. Karagas, Debra T. Silverman, and Stella Koutros. Occupational exposure to organic solvents and risk of bladder cancer. *Journal of Exposure Science & Environmental Epidemiology*, 34(3):546–553, 2024. doi: 10.1038/s41370-024-00651-4.

Jin Zhou, Pinpin Sui, Jianxin Zhao, Xiurong Cheng, Tao Yu, Shiwei Cui, Xiangrong Song, and Caihong Xing. Benzene-induced hematotoxicity enhances the self-renewal ability of HSPCs in Mll-Af9 mice. *Toxicology*, 511:154061, 2025. doi: 10.1016/j.tox.2025.154061.

Ami R. Zota, Ann Aschengrau, Ruthann A. Rudel, and Julia Green Brody. Self-reported chemicals exposure, beliefs about disease causation, and risk of breast cancer in the Cape Cod Breast Cancer and Environment Study: a case-control study. *Environmental Health*, 9:40, 2010. doi: 10.1186/1476-069X-9-40.

A OFFLINE REPLAY AND PROVENANCE

The original glioma artifacts are `papers.json`, `grades.json`, `passages.json`, and `run_summary.json`. Expansion inputs are `expansion_ros.json`, `expansion_ros_breakthrough.json`, and `radical_layer_summary.json`. They remain in `data/`. Derived files are in `data/derived/`, with input digests in `audit_summary.json`. The original 68-record grades must exactly match the frozen implementation before the audit proceeds.

From the repository root, the following commands reproduce or verify the outputs:

```
python scripts/analyze_results.py
python scripts/evaluate_alignment.py
python scripts/analyze_results.py --check
python scripts/evaluate_alignment.py --check
python scripts/benchmark_scifact.py --check
python scripts/analyze_tce_response.py --check
python scripts/diagnose_scope_gate.py --check
python -m unittest discover -s tests -v
```

The audit, development-case, and saved-benchmark scoring commands use Python's standard library. Full benchmark refitting additionally requires the pinned numerical libraries and downloaded public inputs; see the benchmark README and protocol. The pipeline integration test uses the optional retrieval dependencies with a fake client; it makes no network calls. Regression cases cover negation, objective-only text, query leakage, tissue injury versus cancer, transitive bibliographic merges, missing provenance, mismatched endpoints, and passage attachment. Successful tests establish only the specified software behavior.

Live retrieval writes to a separate directory; it never replaces archived inputs. The historical retrieval-label significance figure is excluded because it does not estimate a cancer effect.

B DEVELOPMENT EXAMPLES AND COMPARISON DETAILS

The cards cover NCI TCE and formaldehyde assessments, the INTEROCC glioma study, separate dry-cleaning and PCE analyses from one bladder-cancer meta-analysis, a household cleaning-product breast-cancer study, HOMEChem indoor-air measurements, two OEHHA chemical listings, and the supplied warning paraphrase. Two cards from the same meta-analysis are not independent studies. Each entry includes source provenance, limitations, and a review-pending marker.

Table 5: Agreement with development-authored expected statuses. These figures are conformance checks, not generalization or clinical accuracy. No language model was evaluated.

Rule	Matching expected / cases	Abstentions
Full structured compatibility	24 / 24	2
Chemical identity only	13 / 24	1
Always insufficient metadata	2 / 24	24

The chemical-only comparator checks chemical equality and ignores setting, species, site, endpoint, and evidence type. Cases were deliberately constructed to expose those omissions. Consequently the difference in table entries is expected by construction, is not statistically tested, and provides no estimate of improvement on independently sampled claims. A future study must use expert-annotated sources, balanced answerable and unanswerable claims, and held-out source families.

Prospective evaluation record. The September 5 toxicology protocol in `docs/evaluation_protocol.md` proposes 120 claims subject to annotation resources. It is not preregistered or executed. The completed SciFact transfer experiment is separately specified in `benchmarks/scifact/protocol.md`; domain-specific annotation and biological validation remain pending.

C HUMAN-CELL EXPRESSION REANALYSIS

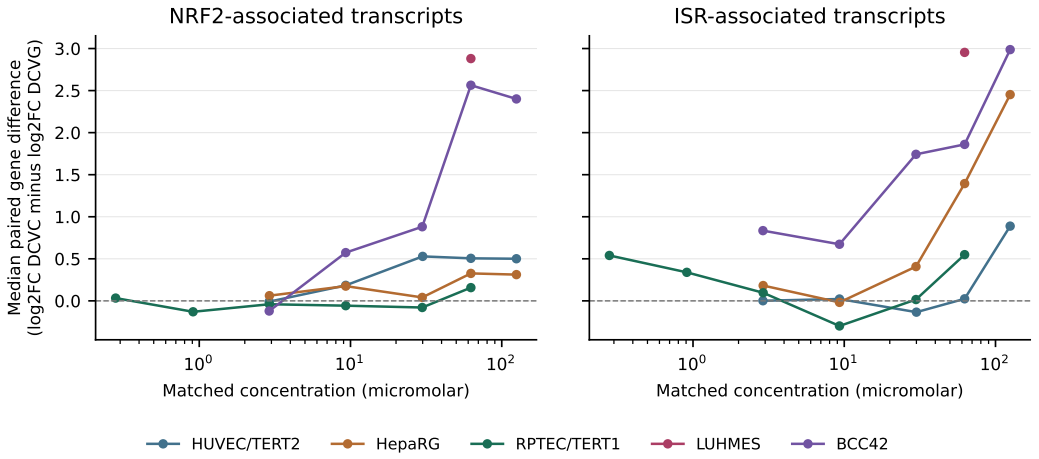

Figure 1: All 22 primary matched-dose pairs, displayed separately for the two selected transcript panels. Positive values indicate larger responses to DCVC than DCVG. Lines connect measured points to guide the eye; they are not fitted curves. There are no new uncertainty intervals or significance tests. The horizontal position uses the DCVC concentration; three pairs differ by less than 1% between arms. All exposures lasted 24 hours. UKN5 lacks DCVG and is absent. Model-specific media, residual viability, and eligible gene sets limit comparisons between curves. A near-zero contrast can accompany strong responses in both arms.

Inputs, panels, and missingness. The six complete processed workbooks are supplementary files 5–10 of Capinha et al. (2023), released under CC BY 4.0. They contain 4,041 probe rows per analyzed condition. Models are endothelial HUVEC/TERT2, tumor-derived hepatic HepaRG, renal RPTEC/TERT1, dopaminergic LUHMES, peripheral-neuronal UKN5, and cortical-neuronal/glial BCC42. We analyze only 1,2-conjugate sheets, excluding the 2,2 isomers. The source selected low-cytotoxicity doses, but exact residual viability, raw replicate counts, and post-QC sample numbers are unavailable in these matrices.

The NRF2-associated panel contains HMOX1, NQO1, GCLC, GCLM, SRXN1, and TXNRD1. The integrated-stress-response-associated panel contains ATF4, DDIT3, ASNS, TRIB3, CHAC1, and PPP1R15A. These are selected transcript panels, not comprehensive pathways or validated protein-activity scores. Terminal numeric probe suffixes are removed for symbol grouping. At least three eligible genes are required; primary panels contain five or six. LUHMES lacks eligible SRXN1 at 62.5 micromolar. Missing numbers remain null, not zero; original probe-level significance fields are retained but not used for selection or treated as tests of the new paired contrasts.

Full results and sensitivities. Of 22 NRF2-associated primary contrasts, 16 are positive and six negative; for the stress-response panel, 19 are positive and three negative. Removing the abundance cutoff preserves those sign counts but changes some magnitudes. Exact matching retains 19 dose pairs, with 14 positive NRF2-associated and 17 positive stress-response contrasts. These are correlated descriptive counts, not binomial trials. Figure 2 displays constituent gene estimates at the common dose; gene spread is not uncertainty across biological replicates. Without replicate-level information and shared-control/cross-gene covariance, we cannot establish statistical separation of the new panel summaries. All 44 primary and 82 sensitivity panel summaries, their gene constituents, both arm medians, and source coordinates are released. The 62.5-micromolar table is selected for common dose availability, not effect magnitude. Raw counts, shared-control covariance, and biological replication would be needed for new inferential comparisons.

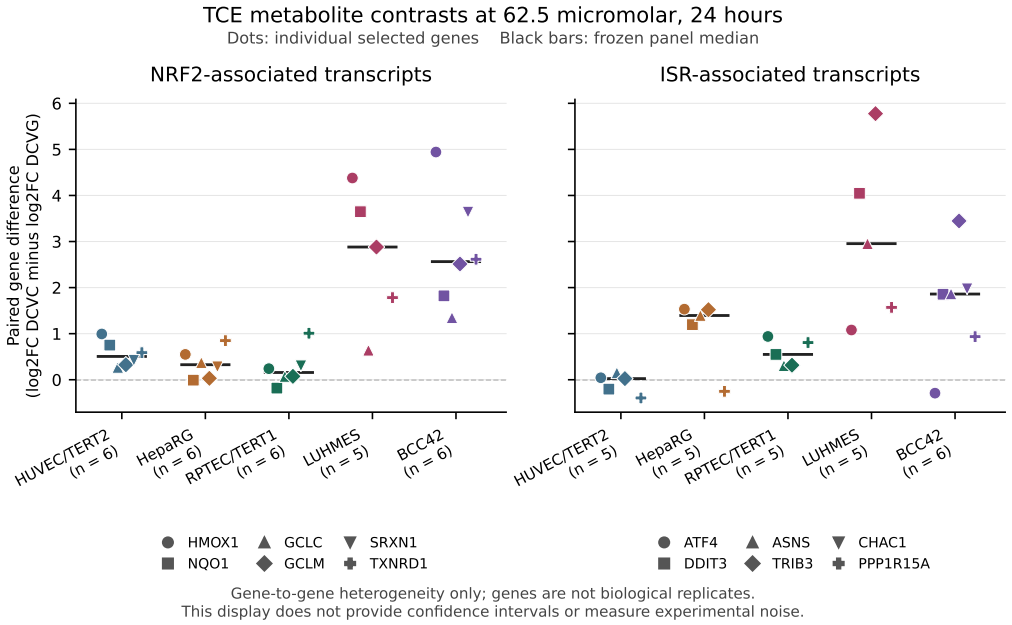

Figure 2: Constituent paired gene estimates at the common 62.5-micromolar dose. Points are the eligible gene-level log2 fold-change differences already present in the frozen report; horizontal summaries show their medians. The displayed genes are selected, correlated transcript targets, not independent biological replicates. Their spread does not provide a confidence interval, standard error, or test of statistical separation. Missing or ineligible genes are not imputed. This figure supplements the complete dose trajectories without changing the analysis.

Measurement-coverage correction. A separate exact-symbol audit queried selected DNA-repair-related transcripts. BRCA2, RAD51, FANCD2, and ATR have no identifiable probes after checking current, previous, and listed HGNC alias symbols. An initial literal lookup also missed H2AFX: it is the previous symbol of H2AX, whose two probes are present in every source workbook. The frozen ledger retains the original selection, and the alias omission is disclosed rather than silently revising the method. H2AX RNA does not measure phosphorylated H2AX foci. Transcript presence, absence, or stability cannot establish BRCA2 protein abundance or intact repair. The audit therefore identifies assay requirements, not evidence against a protein-level mechanism.

Replay. The protocol and computational review are in `docs/discovery-2026-09-08/`; inputs and derived ledgers are in `data/discovery-2026-09-08/`. Standard-library replay uses `python scripts/analyze_tce_response.py --check`. Re-extraction uses the pinned `requirements-discovery.txt` dependency and the same script with `--download --extract`; downloads are verified against recorded hashes. Raw workbooks are not bundled. Figure regeneration uses `scripts/plot_tce_response.py` and `scripts/plot_tce_gene_profiles.py` with Matplotlib 3.11.1. Biological estimates and the external verification benchmark remain separate analyses.

D POST-FEEDBACK SCOPE-INPUT DIAGNOSTICS

This descriptive component audit was specified after the original outcomes and automated pre-submission feedback. It reads the unchanged extraction and prediction ledger, checks its hash against the original run manifest, and neither refits a model nor changes a field or prediction. Counts use pair-role occurrences: a claim or abstract appearing in multiple cited-source pairs is counted repeatedly. They are not independent documents or biological replicates.

Table 6: Known scalar counts for each field. C: claim; S: source; both: both scalars known, whether equal or unequal. Denominators are 339 development pairs and 323 secondary pairs. These are extraction-coverage counts, not precision or recall.

	Development			Secondary		
Field	C	S	Both	C	S	Both
Chemical	0	0	0	0	0	0
Exposure setting	0	29	0	2	30	1
Species	45	187	30	43	183	31
Cancer site	17	33	15	10	26	8
Endpoint	15	84	5	14	80	11
Evidence type	3	92	2	3	88	1

The released JSON additionally records all emitted categories, ambiguous multiple-category matches, equality versus inequality, and every detected mismatch with the original predictions. Zero emission differs from an ambiguous match even though both become unknown scalars. Independent field annotations would be needed to estimate extraction accuracy or distinguish a missed applicable value from true absence or inapplicability.

For fixed fields and provenance, let $C(F)$ count pairs compatible on required set F, and $A(h, F)$ count substantive answers retained from verifier h. By construction, $A(h, F) \leq C(F)$. Here $C(F) = 0$ for all six fields and for the five-field set obtained by removing chemical identity. The latter is a disclosed counterfactual omission, not corrected extraction or a selected improved policy. Holding the remaining fields fixed, filling chemical identity alone cannot restore strict acceptance. This bound diagnoses a bottleneck; it says nothing about scope-gate performance with accurately extracted fields.

Replay this diagnostic separately from the original benchmark:

```
python scripts/diagnose_scope_gate.py --check
```

The report is `data/pat-review-2026-09-14/scope_diagnostics.json`. The accompanying methods note distinguishes pair-level emissions from field recall. An independently implemented identifier-only deduplication sensitivity, with its own replay and exact source offsets, is documented in `docs/pat-review-2026-09-14/`. It preserves the entire archived 238-group partition when all title keys are removed.