Practical Advances in Artificial Intelligence and Computing
Frontier Research Publishing • https://www.frpubs.org
Vol. 4, No. 4, 2025

Typed Evidence Checks Distinguish Citation Presence from Support in a Controlled Scientific Corpus

Tavielle Ardentis[1]*, Nerovan Sablewick[1], Ossian Mirevale[1]
[1]University of Guelph, Guelph, Canada
*Corresponding author: taviellеardentis827@outlook.com

Abstract: This illustrative numerical study separates source existence, lexical relevance and proposition support in a controlled scientific style corpus. A seeded generator creates 240 fictional arithmetic records for eighty material lots and 560 claims spanning supported forms and changes to number, unit, condition, negation or entity. All supported evidence is retrieved first by a specified weighted cosine retriever. A citation presence rule accepts all 400 unsupported claims; a prescribed lexical threshold accepts 328, giving precision 32.8 percent. A six field parser and exact tuple check accept every supported claim and reject every unsupported claim within the declared closed grammar. Retrieving three passages adds no benefit over one. Removing source response content while retaining citation identity reduces supported recall to zero, although rejection of the majority class gives 71.4 percent accuracy. This is a symbolic attribution proxy with no language model, real scientific measurements or open language verification claim.

Keywords: citation verification; retrieval augmented generation; evidence binding; scientific claims; quantity checking; controlled corpus

1 A reference can exist without supporting a claim

A scientific sentence accompanied by a plausible reference presents at least two different questions. One asks whether the cited source exists and can be retrieved. The other asks whether the source actually supports the sentence's entity, numerical result, conditions and qualification. Retrieval augmented systems can answer the first question while leaving the second unresolved. This study separates those operations through a controlled corpus of fictional arithmetic records. We construct claims that preserve much of a source's vocabulary while changing one material field, then compare citation presence, lexical similarity and an exact typed evidence check. The purpose is to make an attribution mechanism inspectable, not to estimate the accuracy of an unexecuted language model.

Retrieval augmented generation combines parametric generation with external documents [1]. Dense passage retrieval [2], contextualized late interaction [3], and passage based generative question answering [4] offer different ways to connect queries and retrieved information. REALM additionally incorporates retrieval into language model pretraining [5]. These developments establish the importance of retrieval but do not make relevance identical to support. A passage can be highly relevant because it discusses the same experiment while disagreeing about the reported result. Our numerical example focuses on that narrow gap and deliberately uses a simple retriever so that the source of an accepted claim can be traced.

Fact verification benchmarks distinguish evidence identification from a support judgment. FEVER provides a large scale verification setting [6], SciFact focuses on scientific claims [7], and SciFact Open broadens the retrieval setting [8]. Our corpus is not drawn from those benchmarks. It contains no real scientific result, public abstract or human annotation. Its records are generated from seeded integer fields and

fixed sentence patterns. That artificiality is useful for an exact attribution experiment because every field can be checked, but it makes the language and evidence substantially easier than in actual scientific literature.

The central comparison asks whether changing an entity, number, unit, condition or negation can pass a relevance oriented check. We include two supported sentence forms and five unsupported transformations for each fictional material lot. The typed verifier parses six fields and accepts only when one retrieved record matches them all. Its perfect result in this closed language is expected and should be read as an implementation and representation demonstration. It is not a claim of general natural language inference, scientific truth assessment or state of the art citation verification. The useful contrast is the measured behavior of weaker checks on exactly the same claims and documents.

Recent citation research reinforces the need for an explicit evaluation target. CiteEval studies fine grained citation assessment [9], while work on imperfect factuality metrics documents the importance of evaluating the evaluators themselves [10]. Our experiment adopts a particularly narrow target: support within a fully specified fictional corpus. It does not judge source authority, experimental validity, causal identification, publication quality or whether a claim is true outside the corpus. Keeping those questions separate makes the numerical success informative without allowing it to stand in for a broader scholarly review.

2 A corpus whose support relations are known by construction

The generator creates eighty fictional material identifiers, from lot000 through lot079. Each receives a base temperature drawn from twenty, thirty or forty degrees on the displayed C scale, an integer dose from one through four milligrams, and an integer response from two through eighteen. These fields are arbitrary symbolic values. They are not measurements of a material, a chemical response or a laboratory instrument. The word yield supplies a scientific style endpoint for the sentence pattern, but it carries no empirical scientific claim. Every document visibly states that it is a fictional arithmetic record.

For each lot the corpus contains three records. The first uses its base temperature and response with a response unit of percent. The second raises temperature by ten and response by three while retaining percent. The third raises temperature by twenty and response by six and uses milligrams as the response unit. The dose remains the same across the three records. This creates 240 paragraphs with related vocabulary and deliberately different conditions. A verifier must therefore link a response to its condition and unit rather than find the lot identifier alone. The records have stable document identifiers used as citation targets in the experiment.

The documents follow one controlled sentence pattern. Claims use either a closely related pattern or a rearranged paraphrase with the response described as a yield shift. Both forms remain within the parser's specified grammar. The use of the word paraphrase should not be mistaken for unconstrained natural language variation: it means exactly the second generated sentence pattern. There are no synonyms outside that pattern, pronouns, anaphora, tables, mathematical derivations or statements spanning several paragraphs. Table 1 lists the corpus dimensions and transformations so that the difficulty of the test is stated before its scores are interpreted.

Table 1. Executed fictional corpus and claim generation. The relation being evaluated is exact support within this closed corpus, not truth about a material.

Component	Count or transformation	Interpretation
Material identifiers	80 fictional lots	No real measured material

Component	Count or transformation	Interpretation
Source paragraphs	3 per lot; 240 total	Condition and unit variants
Supported claims	2 per lot; 160 total	Canonical and rearranged pattern
Unsupported claims	5 per lot; 400 total	Number unit condition negation entity
Attribution methods	5 methods; 2,800 decisions	No language model calls
Retriever	TF IDF cosine; fixed vocabulary	Source fields excluded from ranking

Each lot generates seven claims. The supported claim reproduces the base fields in a reordered sentence. The supported paraphrase uses the second grammar. The number transformation adds seven to the base response. The unit transformation substitutes milligrams for percent while retaining the base condition. The condition transformation raises temperature by ten while retaining the base response, which differs from the response in that lot's second document. The negation transformation inserts did not change. The absent entity transformation replaces the lot identifier with one not present in the corpus. Thus 160 claims are supported and 400 are unsupported under the stated exact relation.

No unsupported label is assigned solely because of a verbal intuition. For the number, unit and condition cases, no corpus record has the same complete tuple as the transformed claim. For negation, all generated source records are affirmative, so the negated claim lacks a matching supporting record. For absent entities, no document contains the requested identifier. We use the single label unsupported for these outcomes rather than conflate absence with contradiction. A real verification task may need several labels and reasoning about inconsistent studies; this binary experiment deliberately avoids that additional ambiguity.

Figure 1 presents a detailed fictional document plate with an actual generated evidence sentence, a supported claim and a number changed claim. The plate is drawn directly from the saved corpus fields and contains no downloaded paper image. It is both illustrative and data linked: its layout is a communication device, while its displayed sentences come from the executed generator. The figure does not imply that a scientific publication exists for the fictional lot. Real literature citations elsewhere in this manuscript provide methodological context and are separate from the synthetic document identifiers used by the attribution experiment.

3 Retrieval is intentionally simple and fully specified

The retriever tokenizes lowercase text into alphabetic strings with optional trailing digits and standalone digit strings. Lot identifiers remain single tokens. A term count vector is calculated for every document. The inverse document frequency is the logarithm of the ratio between the corpus size plus one and the document frequency plus one, followed by addition of one. Multiplying counts by those weights gives a conventional term weighting representation. Every document vector is normalized to unit Euclidean length. This makes the retrieval score a cosine similarity without a trained encoder or external embedding service.

$$idf(t)=log[(D+1)/(df(t)+1)]+1 \quad (1)$$

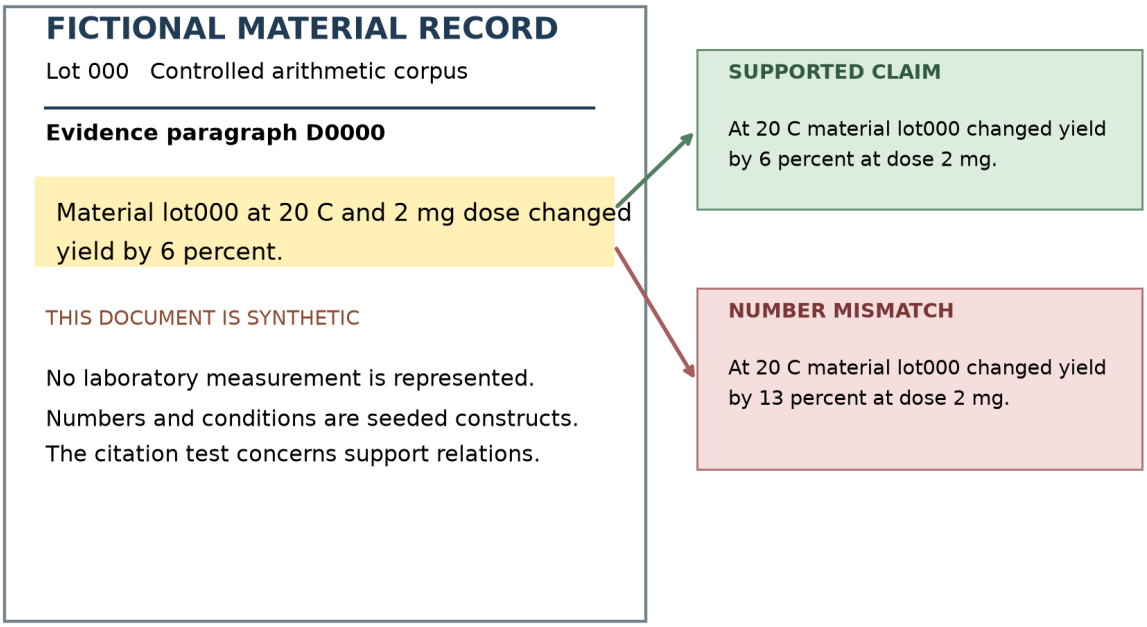


Figure 1. Code drawn fictional scientific document plate using the actual generated first lot evidence sentence and claims. The page is synthetic and represents no real publication or laboratory measurement. Green annotation is supported under the controlled tuple relation; red annotation changes the response value.

The query uses the same tokenization, count weighting and unit normalization. Tokens absent from the document vocabulary are ignored. That design is common in a fixed vocabulary bag of terms model, but it has an important consequence for absent entities: an unknown identifier may contribute no explicit mismatch penalty. The remaining words can still produce a high similarity with an existing document. The experiment measures that behavior rather than assuming that a missing query token automatically causes rejection. It provides one concrete way a relevance score can look reassuring even when a source is about another fictional lot.

$$v_{dt} = tf_{dt} idf(t), \tilde{v}_d = v_d / \| v_d \|_2 \tag{2}$$

Term specificity has long been central to information retrieval [11], and Okapi experiments provide a primary precedent for probabilistic ranking methods [12]. We do not implement BM25 or compare tuned retrieval systems. The simple weighted cosine model is chosen because every contribution can be reconstructed from the corpus and query. Its limitations are part of the study. A stronger retriever could improve relevance, but a relevant document with the wrong number or condition would still need an attribution judgment. The numerical experiment isolates that judgment rather than claiming a new retrieval algorithm.

$$s(q,d) = \tilde{v}_q^{\,T} \tilde{v}_d, R_k(q) = top k_d s(q,d) \tag{3}$$

Documents are ranked by decreasing score, and the top one, two, three and five positions are examined for supported claim evidence recall. The typed attribution methods use one or three retrieved paragraphs. The presence and lexical baselines use the first retrieved paragraph. There is no document filtering by the hidden truth label and no use of the generated structured fields in ranking. The fields remain available for evaluating

support after retrieval. Because each supported claim includes a distinctive lot identifier and exact numerical information, the retrieval task is deliberately easy; the observed recall confirms how easy it is.

Figure 2 shows top three scores for the seven claim forms associated with the first lot and the aggregate supported evidence recall by retrieval budget. The complete support is retrieved first for every supported claim in this executed corpus. Increasing the budget from one to three or five therefore provides no additional supported evidence recall. This is a useful negative result. It prevents a later improvement in attribution from being incorrectly explained as a benefit of retrieving more passages. In the present construction the difference comes from the acceptance rule, because the relevant supporting record was already available.

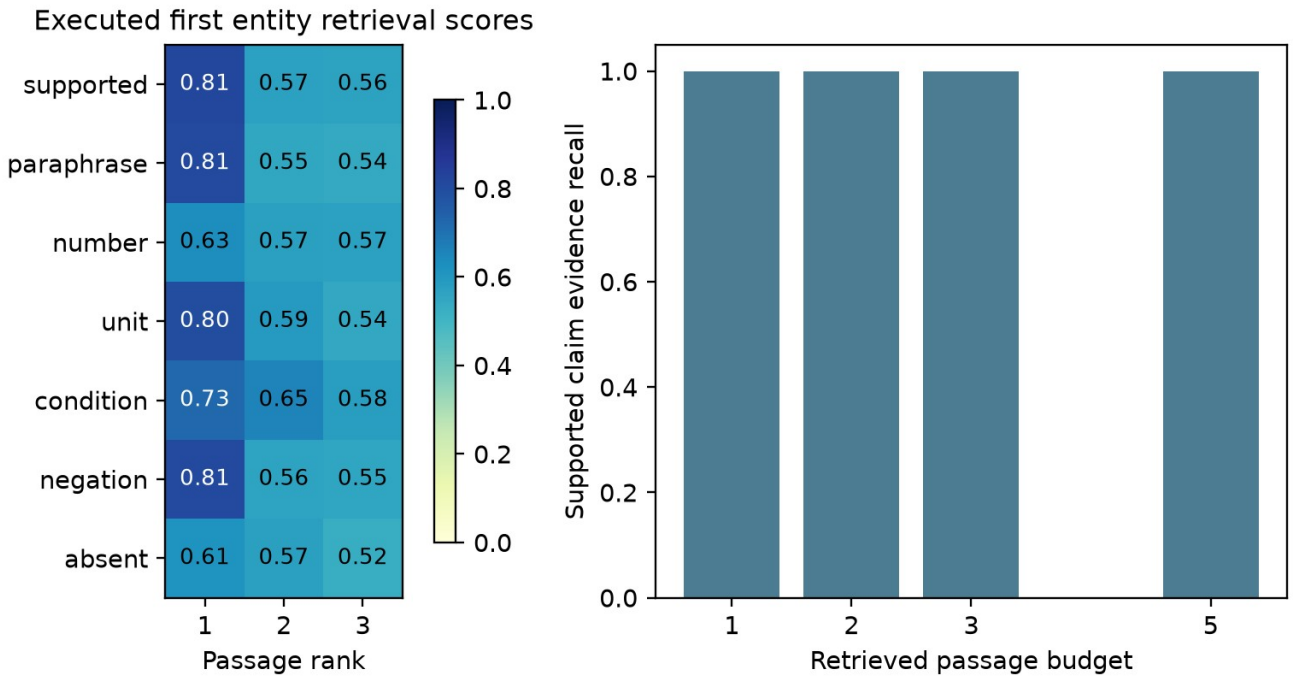


Figure 2. Executed retrieval diagnostics. Left gives top three weighted cosine scores for seven generated claim transformations associated with the first lot. Right gives evidence recall for 160 supported claims; their complete supporting paragraph is already first, so larger budgets add no recall.

4 Three acceptance concepts and one evidence ablation

The citation presence baseline accepts every claim because the retrieval procedure always returns an existing document identifier. This is intentionally weak. It represents the mistaken inference that a resolvable reference is sufficient evidence for the attached sentence. We do not attribute this behavior to a particular product, researcher or language model. Its role is to establish the maximum recall and false acceptance pattern of a pure existence check under a fixed class prevalence. Since every returned identifier belongs to the synthetic corpus, the baseline has perfect document existence even when the entity or scientific field in the claim is wrong.

The lexical baseline accepts a claim if the top cosine score is at least 0.60. This threshold is prescribed rather than tuned on the generated test labels. It transforms a retrieval relevance measure into an attribution decision, which is precisely the heuristic being tested. A claim can retain nearly all words while changing a crucial unit or negation, so the score need not respond in the way the support label requires. We report its

actual threshold and complete confusion counts. There is no assertion that 0.60 is an optimal threshold for attribution or that another threshold would have the same quantitative behavior.

$$a_{lex}(q)=1\{s(q,d_1)\geq 0.60\} \tag{4}$$

The typed method extracts entity, temperature, dose, response value, response unit and a negation flag from the claim and each retrieved paragraph. The parser uses explicit regular expressions for the two claim forms and the source form. It recognizes the first lot identifier, the temperature followed by C, the dose pattern and a response introduced by by or was. A missing required field yields no tuple. Negation is detected through the generated not change phrase. This is exact pattern parsing within a small declared grammar. It is not an entailment model or a language model judge.

$$\tau(q)=(e,T,d,r,u,n) \tag{5}$$

A claim is accepted when its tuple matches at least one complete retrieved source tuple. Every field must match exactly, including the response unit and negation flag. No numerical tolerance, conversion between units or combination across sources is performed. This strict rule is suitable for the artificial exact integer corpus but would be incomplete for realistic scientific prose. For example, two equivalent units could be rejected, and a claim requiring subtraction of two reported means would lack a direct tuple match. The paper treats those cases as limitations rather than silently granting the verifier broader reasoning abilities.

$$a_{typed,k}(q)=1\{\exists d\in R_k(q):\tau(d)=\tau(q)\neq\varnothing\} \tag{6}$$

The evidence ablation retains the same retrieved document identifiers but deletes the source response value and unit from its result phrase before parsing. It does not alter the claim, retrieval ranking or displayed citation identity. The typed method then has no complete evidence tuple and rejects the claim. This is a deliberate fail closed behavior: inability to parse the necessary support is not converted into acceptance merely because a document exists. The ablation separates evidence content from citation metadata. It also creates a class imbalance caution, because rejecting everything can have a seemingly respectable overall accuracy while providing no supported claim recall.

5 Executed attribution results

Table 2 contains the full confusion counts over 560 claims. Citation presence accepts all 160 supported and all 400 unsupported claims. Its precision is therefore 28.6 percent, exactly the supported claim prevalence, and its recall is 100 percent. This is not a discovery about retrieval quality. It is the predictable consequence of using source existence as the acceptance rule. Reporting both precision and recall makes that consequence visible and avoids interpreting complete acceptance of supported claims as a sufficient sign of reliability. A system can have perfect recall while attaching support to every unsupported statement as well.

$$P=TP/(TP+FP),R=TP/(TP+FN) \tag{7}$$

The lexical threshold accepts every supported claim but also 328 unsupported claims. It rejects 72 unsupported claims and no supported claims, giving precision 32.8 percent and accuracy 41.4 percent. Its modest improvement over citation presence comes from rejecting some changed numbers and absent entities. It does not reliably detect changed units, conditions or negation at the stated threshold. Figure 3 displays these counts beside the typed decision matrix. The numerical contrast is large because the unsupported transformations preserve most of the original vocabulary while altering fields that the lexical score does not interpret.

Table 2. Executed attribution confusion counts and metrics over 560 claims. Precision is undefined when no claim is accepted. The typed result is conditional on the declared grammar and exact field relation.

Method	TP / FP	FN / TN	Precision	Recall
citation_present	160 / 400	0 / 0	28.6%	100.0%
lexical	160 / 328	0 / 72	32.8%	100.0%
typed1	160 / 0	0 / 400	100.0%	100.0%
typed3	160 / 0	0 / 400	100.0%	100.0%
typed3_masked	0 / 0	160 / 400	Undefined	0.0%

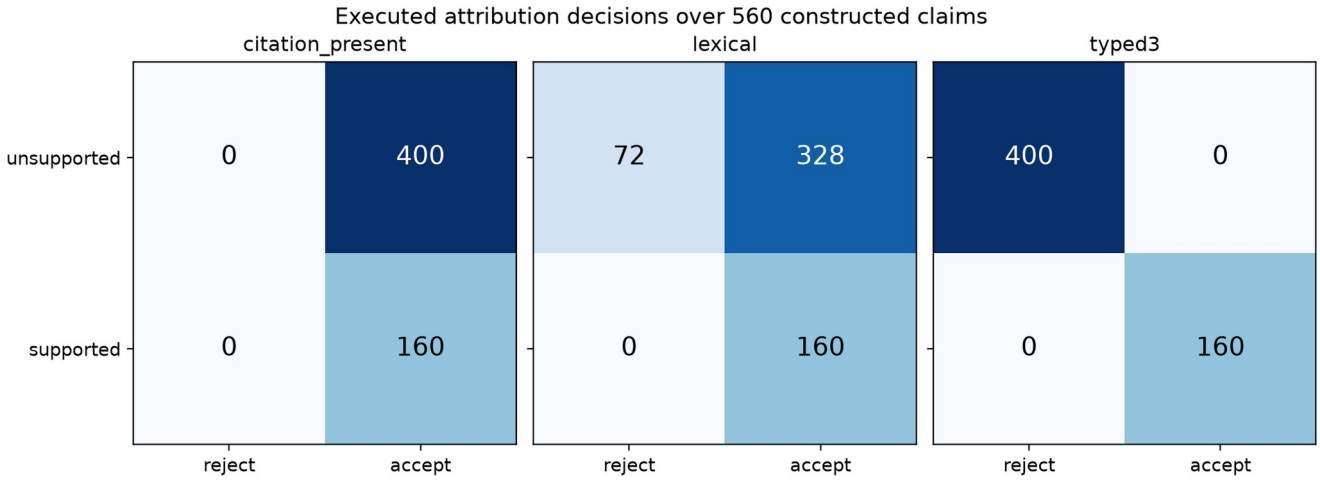


Figure 3. Executed attribution confusion matrices over 560 constructed claims. Rows are corpus support labels and columns are accept or reject decisions. All source identifiers exist; citation presence and lexical relevance still accept unsupported statements.

Both typed top one and typed top three accept all 160 supported claims and reject all 400 unsupported claims. This perfect score is an exact closed language result. It follows from the parser covering the generated sentence forms and the equality rule matching the relation used to define support. We do not present it as evidence of broad semantic understanding. Its scientific value is to show that explicit field preservation resolves the particular errors that survive a relevance threshold, once the source is available. The equality between top one and top three is equally important: extra retrieval supplies no benefit in this controlled case.

After numeric evidence masking, the typed method accepts no claims. Supported recall becomes zero, while overall accuracy is 71.4 percent because all 400 unsupported claims are rejected. Precision is undefined when no claim is accepted; the result should not be described as high precision. This ablation demonstrates why overall accuracy alone can be misleading under an imbalanced claim set. A verifier that refuses all evidence can look better by accuracy than an overaccepting baseline while performing no useful positive attribution. Table 3 reports acceptance by transformation so that these very different behaviors cannot be conflated.

$$Acc = (TP + TN) / (TP + TN + FP + FN) \tag{8}$$

Table 3. Executed acceptance in percent by transformation. There are eighty claims in every column category. Top one and top three typed checks coincide because supporting evidence is already first.

Claim kind	Presence	Lexical	Typed top 1	Typed top 3
supported	100	100	100	100
paraphrase	100	100	100	100
number	100	50	0	0
unit	100	100	0	0
condition	100	100	0	0
negation	100	100	0	0
absent	100	60	0	0

Figure 4 presents the same transformation level acceptance as a heatmap. The lexical method accepts 50 percent of number changes, every unit change, every condition change, every negation and 60 percent of absent entities. These are executed proportions from eighty claims per transformation. They are not estimates of how a neural verifier behaves on scientific papers. The design shows why a relevance score can fail on a controlled set where the correct distinction resides in a few small fields. A system that uses lexical retrieval should therefore attach a separate support check before treating the retrieved identifier as attribution.

6 Units and conditions are part of the proposition

The unit transformation is especially informative because it changes only a word while retaining the numerical response and all acquisition conditions. In the corpus, percent and milligrams describe different response quantities. The typed verifier rejects the mismatch, while the lexical threshold accepts all eighty changed unit claims. This is not a demonstration of dimensional analysis: no conversion or physical equation is evaluated. It is a demonstration that a proposition's unit field cannot be replaced by another unit merely because the surrounding sentence remains similar. Real numerical verification would need richer quantity representations, conversion rules and awareness of what the measured endpoint means.

The condition transformation also preserves most vocabulary. Raising the stated temperature by ten points to another record for the same lot, but that record has a different response. A verifier that checks entity and a nearby number independently could combine fields from incompatible records. Our equality rule requires one source tuple to match the complete claim. It therefore blocks that cross condition assembly. The experiment is a simple instance of a common evidence binding problem: supporting the components of a sentence separately does not necessarily support their conjunction under one experiment or setting.

Negation changes the logical status of the response without changing its entity, condition, dose, number or unit. The lexical baseline's complete acceptance of these cases shows that the selected cosine threshold treats the added words as a small relevance change. The typed method's negation flag makes the distinction explicit. Again, this is a limited parser rule, not general logical reasoning. It detects exactly the generated phrase and does not understand every way a source might deny, qualify or fail to establish an effect. The experiment identifies the importance of the distinction while keeping the implemented recognition mechanism narrow.

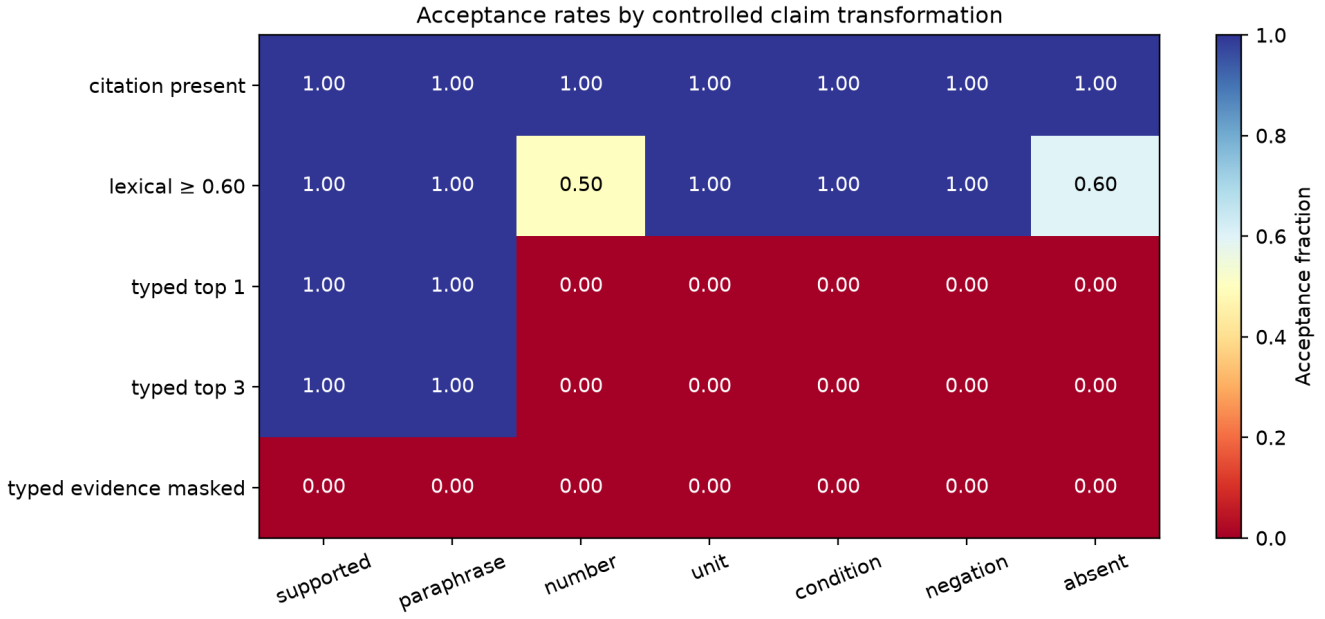


Figure 4. Executed acceptance fraction for eighty claims in each transformation category. Typed matching covers the declared closed sentence grammar. Numeric evidence masking preserves citation identity but removes the response fields required for a complete match.

The absent entity cases expose a separate retrieval artifact. Unknown lot identifiers are absent from the fixed document vocabulary and are ignored when constructing query vectors. Shared scientific words and numerical tokens can still match existing records strongly enough to exceed the lexical threshold. Consequently, 48 of eighty claims about absent lots are accepted by that heuristic. The typed verifier rejects them because the parsed identifier must match a retrieved record. This result illustrates how a vocabulary design decision can affect an attribution decision when a relevance score is used without an entity binding check.

Figure 5 displays field differences between the seven first lot claims and its primary record, alongside the evidence masking recall result. The field plate is a diagnostic visualization of the exact tuples, not a learned explanation generated after prediction. It makes the locus of each perturbation inspectable. The masked evidence experiment then removes the response content while keeping citation identity fixed. Together, these controls distinguish three things that can otherwise be collapsed in a generated answer: a related document, a matching numerical statement and a complete source bound proposition.

7 Attribution differs from scientific truth

Support within the corpus does not establish that any material actually changes yield. The records are invented and explicitly labeled fictional. An accepted claim only matches an affirmative source tuple in that corpus. This distinction is central to scientific citation checking. A published source can support a sentence while containing a flawed experiment, an invalid inference or a result that does not replicate. Conversely, a scientifically true statement may lack support in the particular retrieved collection. Our binary label measures attribution under an exact source relation, not truth in the world, evidential quality or consensus across independent studies.

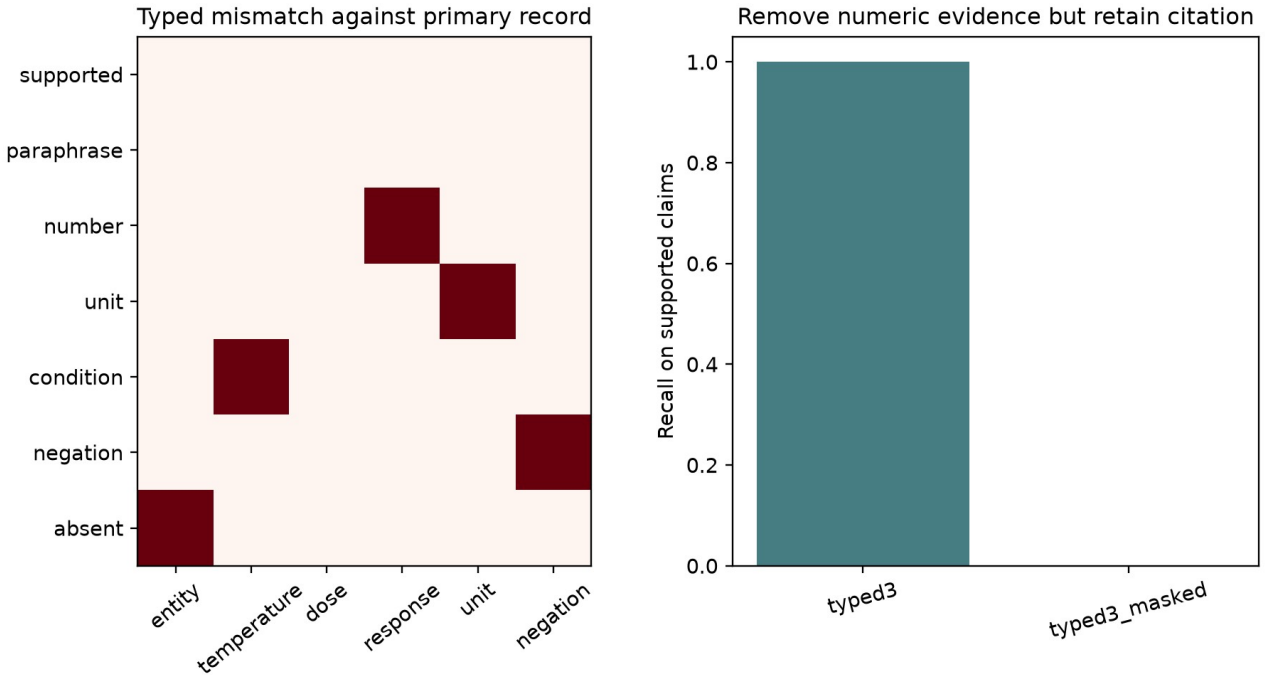


Figure 5. Executed typed mismatch and evidence ablation plate. Left identifies which of six fields changes in first lot claims. Right shows supported recall before and after masking source response content while keeping retrieval and document identity fixed.

Measuring attribution in generated language formalizes the relationship between statements and identified sources [13]. ALCE evaluates generation with citations [14], and fine grained factual precision work separates atomic claims from long form output [15]. Our experiment generates each claim atomically, so it avoids sentence decomposition as a source of error. That simplification is consequential. A real scientific sentence can combine a numerical result, a causal interpretation and a generalization beyond the studied population. A tuple matching one number would not establish support for the entire sentence. The present equality rule is appropriate only because the generated claims have a known fixed structure.

Scientific text models and corpora offer richer foundations for real language tasks. SciBERT studies domain specific representation learning [16], S2ORC provides a large research corpus [17], and QASPER anchors information seeking questions in research papers [18]. None supplies data to the current numerical study. Their relevance is to identify what the controlled grammar omits: varied discourse, section context, scientific terminology and answers distributed across a paper. A realistic extension would need to evaluate parsing and evidence binding on such material rather than assume that success on two sentence patterns transfers automatically.

MuSciClaims adds scientific figures and multimodal evidence to claim verification [19]. Our document plate contains only generated text and is not a multimodal reasoning benchmark. The parser cannot read an image, identify a plotted series or infer a value from an axis. The citation context therefore must not be expanded into a claim of figure verification. The comparison instead provides a small text only mechanism study that can serve as a baseline conceptual distinction when multimodal systems are assessed: retrieval of the right paper remains different from identification and interpretation of the actual evidence supporting a claim.

8 Implications for retrieval augmented pipelines

The experiment suggests an explicit division of responsibility. A retriever locates candidate evidence; an attribution stage checks whether the candidate contains the proposition being asserted. In our code the first stage uses weighted cosine and the second uses a tuple parser. In a richer pipeline either stage could be learned, symbolic or hybrid. What matters is that their outputs are evaluated separately. Complete evidence recall cannot excuse an acceptance rule that confuses conditions, and a precise verifier cannot support a claim whose evidence was never retrieved. The controlled corpus makes the second point dormant because recall is already perfect, which helps isolate the first.

Evidence guided generation explicitly studies the relation between retrieved passages and generated content [20], while Self RAG introduces retrieval and self critique mechanisms [21]. We do not execute either system or claim that its errors match our lexical baseline. Those methods are relevant because they make evidence use part of the model's design. Our result provides a transparent diagnostic that can complement aggregate answer quality: perturb a number, unit, condition or negation while holding much of the text fixed and measure whether attribution changes. Such tests can reveal a field binding weakness that a broad relevance score conceals.

RAGAS supplies automated evaluation components for retrieval augmented generation [22], and long context experiments show that evidence use can depend on where information appears [23]. Our corpus has one short result sentence per paragraph, and the verifier checks each paragraph independently. It therefore has no context length or position problem. This absence should be stated rather than credited to the typed checker. A real document can introduce the entity in one paragraph, define units in a table caption and qualify the result in a later section. Extending the tuple method would require reliable context assembly before exact field comparison becomes meaningful.

Natural language inference datasets such as SNLI [24] and MultiNLI [25] study broader sentence relationships than our parser handles. They do not guarantee that a verifier trained on them will bind scientific quantities or experimental conditions correctly. The present study neither trains such a verifier nor evaluates transfer from those datasets. It uses them as context for the distinction between a controlled symbolic relation and open language support. The perfect tuple result is best interpreted as a reference behavior for the declared grammar, against which more general systems could later be compared on the same controlled transformations.

Selective prediction studies the option to abstain when a model cannot make a reliable decision [26]. The numeric evidence ablation gives an elementary example of why an attribution stage may need such an option. The implemented checker rejects a claim when required evidence is absent, conflating no support found with incomplete parsing in its binary output. A practical system should expose that distinction to a user. We do not build a user interface or calibrate abstention probabilities here. The result simply shows that preserving a citation identifier while deleting its substantive result should not leave a positive support decision unchanged.

9 Limitations of a closed language experiment

The typed checker and the labels share a field representation. Its perfect performance is therefore partly a consequence of experimental design, not a surprising discovery. The meaningful comparison lies in how a purely relevance based rule behaves when a few proposition fields change. This limitation should be central to interpretation. We do not claim that regular expressions solve scientific citation verification. They solve the exact grammar supplied to them. The experiment's value is to isolate support relations and failure

mechanisms without introducing unknown model behavior, then make the assumptions available for critique and extension.

The corpus also lacks numerical uncertainty. Responses are integers, equality is exact and there are no confidence intervals, rounding conventions, significant figures or statistical qualifiers. A statement of approximately twelve percent and a source reporting 11.8 percent would need a tolerance policy that depends on context. A claim of a significant increase would require interpreting uncertainty and comparison groups rather than matching a response value. None of those operations is included. Extending the checker to them would require a richer proposition model and new experiments, especially because an overly generous tolerance could create false support.

There is no multi source synthesis. The acceptance rule requires one complete matching record and cannot combine evidence across documents. This is deliberate for preventing condition mixing, but it would reject legitimate claims that summarize several compatible studies or calculate a derived quantity. It also cannot reason about source disagreement, updated results or a retracted finding. Those are essential scholarly issues beyond the current proxy. The paper does not describe the corpus as a representative sample of citation practice; it is a controlled diagnostic set whose narrowness makes the comparisons exact.

The lexical threshold is only one prescribed value. We do not optimize it or present a receiver operating curve, so the reported precision and accuracy are properties of that specific rule. Another threshold could trade recall for fewer false acceptances. However, a threshold cannot by itself provide the semantic meaning of a temperature or unit field. A broader study could compare calibrated thresholds, learned entailment models and typed tools using held out grammar variations. Such an extension should assess its own evidence rather than inherit a broad performance claim from the present closed corpus.

10 A limited success with an explicit evidence boundary

The executed study separates source existence, lexical relevance and proposition support. All sources returned by retrieval have valid synthetic identifiers, but citation presence accepts 400 unsupported claims. A cosine threshold rejects some changed numbers and absent entities yet still accepts 328 unsupported claims. An exact typed check rejects those transformations and accepts every supported sentence form within its declared grammar. Increasing the retrieval budget adds no benefit because all supported evidence was already ranked first. Removing numeric evidence drives supported recall to zero even while overall accuracy reaches 71.4 percent through rejection of the majority class.

These outcomes support a specific methodological claim. A citation checking pipeline should preserve the binding among entity, result, unit, condition and qualification, and it should evaluate support separately from retrieving a plausible document. The controlled checker demonstrates that distinction without claiming open language understanding or scientific truth assessment. Its complete corpus, claims, decisions and figures are archived with the generator, making every accepted and rejected sentence inspectable. The result is a reproducible numerical example of evidence discipline: the source attached to a statement must support that statement's proposition, not merely share its vocabulary or exist at a valid identifier.

11 Error accounting makes the comparison interpretable

The lexical method's 328 false acceptances can be decomposed without a fitted explanation. Forty come from changed numbers, eighty from changed response units, eighty from changed temperatures, eighty from negation and forty eight from absent entities. These categories sum exactly to the false positive count in Table 2. The breakdown matters because two methods with the same overall precision could fail for different reasons. A threshold might reject rare identifiers well while remaining insensitive to units, or handle a

negation phrase while mixing experimental conditions. Transformation level accounting makes those differences visible and provides concrete targets for a later verifier rather than a single undifferentiated error percentage.

The source response value varies between two and eighteen, so adding seven can introduce either a familiar or less familiar standalone digit token. The cosine score responds to that lexical change through the corpus vocabulary and inverse document frequencies. Its fifty percent acceptance rate for number changes is therefore not evidence of arithmetic reasoning. The method never subtracts the claim value from the source value. It only compares weighted token vectors. The exact numerical behavior depends on the generated values and token distribution, which is why the seed and full corpus are necessary for replication.

$$FP_{lex} = 40 + 80 + 80 + 80 + 48 = 328 \tag{9}$$

The absent entity result is similarly corpus dependent. An identifier unseen by the document vocabulary is dropped from the query vector, leaving its other tokens to determine similarity. That can actually increase the relative contribution of shared words after normalization. The method does not possess an explicit unknown entity category. The typed parser does retain the identifier string, so it can require exact equality even if the string never appeared during retrieval vector construction. This is a concrete example of two stages using different representations of the same query, with a predictable consequence for attribution.

The presence baseline's precision is also prevalence dependent. With 160 supported claims among 560, accepting everything gives two sevenths precision. If the generator produced a different balance of supported and unsupported transformations, the same baseline behavior would yield a different precision without becoming more discerning. The reported percentage therefore should not be generalized to a real corpus's citation quality. It is a property of the controlled evaluation mixture. The complete confusion counts allow readers to recompute performance under alternative class weights without mistaking that reweighting for a new executed language experiment.

$$\pi = 160/560 = 2/7, P_{all} = \pi \tag{10}$$

The masking control illustrates the opposite extreme. Rejecting every claim gives five sevenths accuracy because five of the seven transformation classes are unsupported. It provides no positive evidence service at all. This is why the paper gives recall and acceptance counts beside accuracy. A practical verifier could also report abstention coverage and distinguish an explicit contradiction from a missing parse. The present binary output does not make that distinction, so its perfect rejection of unsupported masked claims should not be praised as successful verification. The ablation demonstrates dependence on evidence content and a metric pitfall at the same time.

12 Exact support is an evidence binding contract

Table 4 summarizes the contract implemented by the typed checker. A supported result needs an entity, temperature, dose, response value, response unit and polarity from the same retrieved paragraph. This same paragraph condition is important. If a claim has the base temperature but a response from the warmer record, separate matches across documents would create a false conjunction. The current implementation avoids that by comparing complete tuples one record at a time. It does not search the corpus for one matching field after another and assemble a composite source that no document actually states.

Table 4. Exact evidence binding contract and scope. Every positive decision needs one complete matching tuple from one retrieved paragraph.

Required field	Implemented check	Outside present scope
Entity	Exact lot string	Coreference and aliases
Temperature and dose	Exact parsed integers	Unit conversion or uncertainty
Response value and unit	Exact integer and unit token	Rounding and derived quantities
Polarity	Generated negation phrase	General qualification and logic
Evidence completeness	One paragraph has all six fields	Multi source synthesis
Missing result phrase	Reject incomplete tuple	Distinguishing parse failure from no support

The evidence masking operation is confined to the result phrase in retrieved documents after ranking. The document identifier, lot identifier, temperature, dose and remaining explanatory sentence remain present. Thus retrieval scores and citation identities are held fixed by design, while the checker loses the response value and response unit required for a complete tuple. This is a counterfactual content ablation, not a claim that a real paper's evidence has been removed from the world. Its interpretation is causal within the code: changing those source fields is sufficient to eliminate every positive typed attribution in this controlled setting.

$$\tau(m(d)) = \varnothing \Longrightarrow a_{typed}(q; m(R_k)) = 0 \tag{11}$$

The parser's strictness also creates obvious false negative possibilities outside the declared grammar. A source could state a dose in micrograms, use a negative signed response, write a temperature in words or refer to a previously defined material with a pronoun. The current regular expressions would fail or produce an incomplete tuple. Those failures would be limitations of parsing rather than evidence against the claim. A practical system should preserve the difference. Our experiment does not hide it behind a language understanding claim: the accepted grammar is part of the method and the perfect score is conditional on that grammar.

A useful next experiment would generate equivalent quantities in different units while preserving their physical meaning. That would require selecting canonical units, checking dimensional compatibility and defining rounding tolerances before evaluation. Another extension could split a result and its condition across two adjacent sentences while retaining an unambiguous entity reference. That would test context assembly rather than exact sentence parsing. These are concrete falsifiable changes to the current contract. We do not predict that the present checker would handle them; its code makes clear that it generally would not without modification.

The narrowness of the current success is therefore part of its value. It provides an explicit reference behavior for a known support relation, and it exposes how citation presence and lexical overlap can deviate from that behavior even when retrieval is perfect. A more general model could be evaluated against the same transformation categories before moving to natural scientific text. If it failed a unit or condition change here, the error could be diagnosed precisely. If it succeeded, that would still be a preliminary control rather than a substitute for evaluating real prose, uncertain quantities, multimodal evidence and source quality.

References

[1] Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

[2] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020. doi:10.18653/v1/2020.emnlp-main.550.

[3] Khattab, O., and Zaharia, M. (2020). ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. SIGIR; author manuscript arXiv:2004.12832.

[4] Gautier Izacard, Edouard Grave. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021. doi:10.18653/v1/2021.eacl-main.74.

[5] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, Ming-Wei Chang. REALM: Retrieval-Augmented Language Model Pre-Training. arXiv preprint arXiv:2002.08909v1, 2020. https://arxiv.org/abs/2002.08909v1

[6] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Arpit Mittal. FEVER: a Large-scale Dataset for Fact Extraction and VERification. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018. doi:10.18653/v1/n18-1074.

[7] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, Hannaneh Hajishirzi. Fact or Fiction: Verifying Scientific Claims. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020. doi:10.18653/v1/2020.emnlp-main.609.

[8] David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, Hannaneh Hajishirzi. SciFact-Open: Towards open-domain scientific claim verification. Findings of the Association for Computational Linguistics: EMNLP 2022, 2022. doi:10.18653/v1/2022.findings-emnlp.347.

[9] Yumo Xu, Peng Qi, Jifan Chen, Kunlun Liu, Rujun Han, Lan Liu, Bonan Min, Vittorio Castelli, Arshit Gupta, Zhiguo Wang. CiteEval: Principle-Driven Citation Evaluation for Source Attribution. Proceedings of ACL, pp. 32759–32778, doi:10.18653/v1/2025.acl-long.1574, 2025. https://aclanthology.org/2025.acl-long.1574/

[10] Ameya Godbole, Robin Jia. Verify with Caution: The Pitfalls of Relying on Imperfect Factuality Metrics. Findings of ACL, pp. 22889–22912, doi:10.18653/v1/2025.findings-acl.1175, 2025. https://aclanthology.org/2025.findings-acl.1175/

[11] KAREN SPARCK JONES. A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL. Journal of Documentation, 1972. doi:10.1108/eb026526.

[12] Micheline Hancock-Beaulieu, Mike Gatford, Xiangji Huang, Stephen E. Robertson, Steve Walker, P. W. Williams. Okapi at TREC-5. , 1996. doi:10.6028/nist.sp.500-238.city.

[13] Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, David Reitter. Measuring Attribution in Natural Language Generation Models. Computational Linguistics, 2023. doi:10.1162/coli_a_00490.

[14] Tianyu Gao, Howard Yen, Jiatong Yu, Danqi Chen. Enabling Large Language Models to Generate Text with Citations. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023. doi:10.18653/v1/2023.emnlp-main.398.

[15] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, Hannaneh Hajishirzi. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023. doi:10.18653/v1/2023.emnlp-main.741.

[16] Iz Beltagy, Kyle Lo, Arman Cohan. SciBERT: A Pretrained Language Model for Scientific Text. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019. doi:10.18653/v1/d19-1371.

[17] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, Daniel Weld. S2ORC: The Semantic Scholar Open Research Corpus. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.acl-main.447.

[18] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, Matt Gardner. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021. doi:10.18653/v1/2021.naacl-main.365.

[19] Yash Kumar Lal, Manikanta Bandham, Mohammad Saqib Hasan, Apoorva Kashi, Mahnaz Koupaee, Niranjan Balasubramanian. MuSciClaims: Multimodal Scientific Claim Verification. Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, 2025. doi:10.18653/v1/2025.ijcnlp-long.175.

[20] Akari Asai, Matt Gardner, Hannaneh Hajishirzi. Evidentiality-guided Generation for Knowledge-Intensive NLP Tasks. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022. doi:10.18653/v1/2022.naacl-main.162.

[21] Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. (2023). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. arXiv:2310.11511.

[22] Shahul Es, Jithin James, Luis Espinosa Anke, Steven Schockaert. RAGAs: Automated Evaluation of Retrieval Augmented Generation. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2024. doi:10.18653/v1/2024.eacl-demo.16.

[23] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. Transactions of the Association for Computational Linguistics, 2024. doi:10.1162/tacl_a_00638.

[24] Samuel R. Bowman, Gabor Angeli, Christopher Potts, Christopher D. Manning. A large annotated corpus for learning natural language inference. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015. doi:10.18653/v1/d15-1075.

[25] Adina Williams, Nikita Nangia, Samuel Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018. doi:10.18653/v1/n18-1101.

[26] Zhengyao Gu, Mark Hopkins. On the Evaluation of Neural Selective Prediction Methods for Natural Language Processing. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023. doi:10.18653/v1/2023.acl-long.437.