Research Square

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

When Sources Disagree: Accuracy-Preserving Batch Verification and the Limits of Source-Age Screening in Retrieval-Augmented Generation

Kholofelo JONASE KGATLA

kgatlakholofelo@gmail.com

Tshwane University of Technology https://orcid.org/0000-0001-6067-4929

Nthabiseng Michael Moeti

Tshwane University of Technology https://orcid.org/0000-0003-3783-7130

Research Article

Keywords: retrieval-augmented generation, citation verification, contradiction detection, source-age screening, natural language inference, batch inference

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11678157/v1

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

Additional Declarations: The authors declare no competing interests.

When Sources Disagree: Accuracy-Preserving Batch Verification and the Limits of Source-Age Screening in Retrieval-Augmented Generation

Kholofelo Jonase Kgatla[1,*] (ORCID: 0000-0001-6067-4929)
Nthabiseng Michael Moeti[2] (ORCID: 0000-0003-3783-7130)

[1]Department of Computer Science, Tshwane University of Technology, Polokwane, South Africa.
[2]Department of Computer Science, Tshwane University of Technology, Pretoria, South Africa.
[*]Corresponding author: kgatlakholofelo@gmail.com
Co-author email: Moetimn@tut.ac.za

HIGHLIGHTS

- MiniLM-NLI was the most accurate of six contradiction-detection approaches on a 200-case balanced SciFact pilot (72.0% accuracy, macro-F1 0.714) and reached 68.8% accuracy on the full 773-case set.
- A negation-first shortcut cut median latency by 26.9% but lowered pilot accuracy to 66.5% (62.5% on the 773-case set); it is reported as a negative result and not as a replacement for the NLI model.
- Length-sorted batching of four reduced MiniLM-NLI total processing time by 10.55% with identical predictions on all 200 pilot cases.
- A 90-day source-age rule flagged all five manually corroborated FreshQA cases, including the unchanged control, so it works as a review trigger and not as a staleness detector.
- Findings are limited by a non-random pilot, a five-case temporal set and a single scientific domain; the components were evaluated separately and not as a deployed RAG pipeline.

Abstract

Retrieval-augmented generation (RAG) can attach citations that contradict a generated claim or no longer reflect current facts. We revisit an experimentally evaluated citation-verification design using SciFact for semantic contradiction and a small, independently reviewed FreshQA case study for temporal change. On 200 balanced SciFact SUPPORT/CONTRADICT cases, a cross-encoder MiniLM-NLI baseline achieved 72.0% accuracy and 0.7144 macro-F1. The originally proposed negation-first shortcut reduced accuracy to 66.5%, despite reducing model invocations; consequently, we do not recommend that shortcut as an accuracy-preserving gate. In a separate five-repeat CPU throughput benchmark, length-sorted batch inference (batch size four) reduced total processing time from 51.779 s to 46.317 s (10.55%) while reproducing 100% of baseline predictions. Historical-source review identified four genuine FreshQA answer changes and one unchanged control. Applied to these five 148-day-old sources, the fixed 90-day age rule flagged all five: it alerted on all four corroborated changes and also on the one unchanged control (four true alerts, one false alert). Source age therefore indicates a need for re-verification, not proof of factual obsolescence. The findings support accuracy-preserving throughput optimization and motivate content-aware temporal checks; neither online latency improvement nor generalizable staleness classification is established.

Keywords: retrieval-augmented generation; citation verification; contradiction detection; source-age screening; natural language inference; batch inference

1.0 INTRODUCTION

Retrieval-augmented generation (RAG) systems answer questions by conditioning a language model's output on passages retrieved from an external corpus, with the aim of anchoring generated text in verifiable evidence rather than relying solely on parametric memory (Lewis et al., 2020). This grounding has become the dominant strategy for reducing hallucination in open-domain question answering and knowledge-intensive generation, and it underlies most production question-answering and search-assistant systems currently deployed. A citation attached to a generated answer, however, is a claim of support, not a guarantee of it: the system has retrieved a passage, but nothing in the retrieval step confirms that the passage actually entails the answer it is cited alongside.

This gap between retrieval and verification manifests in at least two distinct failure modes. The first is semantic contradiction: the retrieved evidence may directly conflict with the claim it is meant to support, either because the retriever surfaced a topically related but factually opposed passage, or because the generation step misread the evidence's polarity, for example, inverting a negated statement. Benchmarks such as FEVER (Thorne et al., 2018) and SciFact (Wadden et al., 2020) were built specifically to test this failure mode, pairing claims with SUPPORT, CONTRADICT, or NOT ENOUGH INFO labels grounded in Wikipedia and biomedical abstracts, respectively, and framing citation checking as a natural language inference (NLI) problem.

The second failure mode is temporal: a retrieved passage may have been accurate at the time it was indexed, but has since been superseded by a change in the underlying fact. Sports standings, officeholders, prices, and ongoing events are all subject to this kind of drift, and a RAG system with no awareness of source recency will cite a once-true passage as though it remains true indefinitely. FreshQA (Vu et al., 2024) was constructed to probe this behaviour directly, distinguishing fast-changing, slow-changing, and never-changing questions and re-annotating answers against different evaluation dates. Contradiction detection and staleness detection are usually studied separately, yet both are properties a production citation-verification layer needs to check before an answer is surfaced to a user.

A further practical constraint shapes how such a verification layer can be built. NLI models capable of the accuracy needed for reliable contradiction detection are typically cross-encoder transformers that must score every claim-evidence pair independently, at a cost proportional to the number of citations being checked. Running such a model on every retrieved passage in a high-throughput RAG deployment imposes a latency and compute cost that may be difficult to justify when the majority of citations are, in fact, unproblematic. This motivates efficiency-oriented verification. Two families of approach are natural: a cascade, in which a cheap heuristic disposes of clear-cut cases and the expensive model is reserved for the remainder, an approach explored elsewhere in NLP for cost-efficient inference; and systems-level optimisation of how the expensive model is executed, which leaves its decisions unchanged. We evaluate one instance of each and report the accuracy cost of the former explicitly.

The central question is how to improve the computational efficiency of claim–evidence verification without degrading its semantic decisions, while separately assessing what source timestamps can and cannot establish about factual change. We evaluate a previously proposed lexical shortcut as a negative result and a batch-inference implementation as an accuracy-preserving alternative.

The aim is to (i) evaluate the accuracy cost of lightweight lexical gates against MiniLM-NLI, (ii) test whether length-sorted batch inference preserves predictions while reducing total CPU processing

time, and (iii) examine the limitations of a source-age alert using historically corroborated changed and unchanged cases.

This study addresses the following research questions:

- RQ1: How accurately can lightweight methods detect contradictions between claims and cited evidence?
- RQ2: Can length-sorted batch inference improve MiniLM-NLI throughput while preserving its classification decisions, and how does this compare with the original negation-first shortcut?
- RQ3: Does a fixed source-age threshold distinguish genuine temporal changes from unchanged information in a manually corroborated FreshQA case study?

The contributions are threefold. First, we report a 200-case balanced SciFact contradiction evaluation, retaining the original lexical-gate ablations as diagnostic results. Second, we evaluate an accuracy-preserving length-sorted batch implementation in five repeated CPU runs and distinguish throughput from online per-request latency. Third, we document a historical-evidence review of FreshQA candidates: four corroborated answer changes, one unchanged control, and one explicitly excluded purported change, with a five-case age-screening evaluation. The five cases are illustrative and do not constitute a representative benchmark.

The remainder of this paper is organised as follows. Section 2 reviews related work on retrieval-augmented generation, automated fact verification, NLI-based contradiction detection, temporal knowledge staleness, and cascade-based verification. Section 3 describes the dataset construction, verification protocol, proposed system, and experimental setup. Section 4 presents the results for contradiction detection, the accuracy and throughput comparison, and the temporal-risk case study. Section 5 discusses the findings and their implications for RAG system design. Section 6 outlines the limitations of the study, and Section 7 concludes with directions for future work.

2.0 RELATED WORK

2.1 Retrieval-Augmented Generation and Citation Reliability

Large language models are prone to hallucination, producing fluent but factually unsupported text, a failure mode surveyed extensively by Ji et al. (2023) across natural language generation tasks more broadly. Retrieval-augmented generation combines a retriever, which selects passages from a document collection relevant to an input query, with a generator that conditions its output on those passages (Lewis et al., 2020), and is one of the principal strategies proposed to curb this failure mode. In practice, the retriever is often a dense bi-encoder such as DPR (Karpukhin et al., 2020), which embeds queries and passages independently so that passage embeddings can be precomputed and searched efficiently, a design choice this study revisits in Section 2.3 when discussing the cross-encoder/bi-encoder trade-off for the verification step itself. The approach is motivated by two related limitations of purely parametric language models: their knowledge is frozen at training time, and their outputs are not naturally traceable to a verifiable source. By surfacing the retrieved passages alongside the generated answer, RAG systems allow a user, or a downstream verification component, to check the answer against its stated evidence.

This traceability is frequently treated as equivalent to correctness in practice: if an answer is accompanied by a citation, the citation is assumed to support the answer. This assumption does not

always hold. A retriever operating on lexical or dense similarity may surface a passage that is topically related to a query without entailing the specific claim generated from it, and a generation step can misstate, invert, or overgeneralize what a retrieved passage actually says. Citation reliability, in this sense, is a property that must be checked rather than assumed, and it is separable from retrieval quality: a well-retrieved passage can still be miscited.

Recent work has begun to treat citation quality as a first-class evaluation target rather than an afterthought. Gao et al. (2023a) introduce ALCE. This benchmark scores generated text along fluency, correctness, and citation quality, and report that even strong LLMs fail to fully support their claims with cited evidence a substantial fraction of the time. Gao et al. (2023b) propose RARR, a framework that retrospectively checks and edits a language model's output against retrieved evidence, treating attribution repair as a distinct post-hoc step rather than something the generation process guarantees on its own. Niu et al. (2024) provide direct evidence that RAG does not eliminate this problem: their RAGTruth corpus of manually annotated LLM outputs shows that models using RAG still produce claims that are unsupported by, or directly contradict, their own retrieved context, across multiple LLMs and task types. More recent work reaches similar conclusions for deployed systems and evaluation practice. Liu et al. (2023) audited commercial generative search engines and found that their responses often contain statements that are not supported by the accompanying citations, and citations that do not support the statements they are attached to. Reference-free evaluation frameworks such as RAGAs (Es et al., 2024) score the faithfulness of generated answers to retrieved context at the pipeline level, rather than checking individual claim–citation pairs. Together, this body of work motivates the present study's premise: verifying that a citation supports its claim, and remains current, is a separate problem from retrieval and generation, and needs its own evaluation and tooling.

2.2 Automated Fact and Claim Verification

Automated fact verification frames citation checking as a supervised classification problem over claim-evidence pairs, typically with SUPPORT, REFUTE/CONTRADICT, and NOT ENOUGH INFO labels. FEVER (Thorne et al., 2018) established this paradigm at scale using Wikipedia as the evidence source, requiring systems to retrieve relevant sentences and then classify the claim's relationship to that evidence. SciFact (Wadden et al., 2020) adapted the same task to the biomedical literature, pairing scientific claims with abstracts from PubMed-indexed papers and expert-annotated SUPPORT or CONTRADICT labels together with the specific rationale sentences. SciFact's domain, dense with numerical findings, comparative statements, and negated results, makes it a demanding testbed for contradiction detection specifically, which is why it is used as the contradiction-detection benchmark in this study. Later systems have continued to use SciFact: MultiVerS (Wadden et al., 2022b) predicts abstract-level labels and rationale sentences jointly, using full-document context and weak supervision, and SciFact-Open (Wadden et al., 2022a) re-evaluates scientific claim-verification systems against a much larger corpus and reports a substantial drop in performance relative to the original benchmark, a reminder that results on SciFact alone should not be extrapolated.

A closely related line of work treats evidence change itself as the object of study. Schuster et al. (2021) introduce VitaminC, a benchmark built from over 100,000 Wikipedia revisions in which claim-evidence pairs are constructed to be nearly identical except for the specific factual detail that was revised, and show that models trained on such contrastive pairs become substantially more sensitive to subtle factual edits than models trained on FEVER-style data alone. VitaminC sits conceptually between the contradiction and staleness problems addressed separately in this study: it captures cases where an

evidence source's own content has changed over time, which is closer to the staleness scenario this study evaluates via FreshQA, but frames the resulting task as a semantic classification problem over evidence pairs, in the manner of SciFact and FEVER, rather than as a metadata-based timestamp check. This reinforces the research gap identified in Section 2.6: few evaluation efforts jointly treat within-text semantic contradiction and external temporal staleness as distinct signals requiring different detection strategies.

2.3 Natural Language Inference for Contradiction Detection

Natural language inference (NLI) casts the relationship between a premise and a hypothesis as one of entailment, contradiction, or neutrality, and pretrained cross-encoder NLI models have become a common off-the-shelf tool for claim verification tasks, since a claim-evidence pair can be treated directly as a hypothesis-premise pair. Such models are almost universally built on the Transformer architecture (Vaswani et al., 2017), whose self-attention mechanism allows every token to attend to every other token in the input; pretrained encoders derived from this architecture, most notably BERT (Devlin et al., 2019), provided the first large-scale demonstration that a single pretrained model could be fine-tuned to strong performance across entailment, question-answering, and classification tasks alike, and remain the base architecture from which the cross-encoder NLI models used in claim verification, including MiniLM in this study, are derived. Cross-encoder architectures, which jointly encode both text spans before producing a relationship score, generally outperform embedding-similarity approaches on entailment-style tasks because they can model fine-grained interactions between the two texts, such as negation scope or numerical comparison, that similarity metrics computed from independently encoded vectors cannot capture. This accuracy advantage comes at a computational cost: unlike a bi-encoder, whose passage embeddings can be precomputed and cached, a cross-encoder must run a full forward pass for every claim-evidence pair at inference time, which is the central efficiency constraint this study addresses.

Large-scale human-annotated NLI corpora such as SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) established entailment/contradiction/neutral classification as a well-studied three-way task and remain the standard training data for general-purpose NLI models, including cross-encoder models of the kind used in this study. Reimers and Gurevych (2019) formalised the cross-encoder/bi-encoder distinction explicitly in the context of sentence-pair tasks, showing that bi-encoders trade some accuracy for the ability to precompute and cache embeddings, while cross-encoders retain higher accuracy by jointly attending over both texts at the cost of a full forward pass per pair; this trade-off motivates asking whether a cheaper filter, or a more efficient way of executing the cross-encoder, can reduce its cost without discarding it in favour of a bi-encoder outright. The MiniLM architecture used as the cross-encoder in this study (Wang et al., 2020) was itself introduced as a distillation method for compressing large pre-trained transformers while retaining most of their task accuracy, which is consistent with its use here as the more capable but more expensive verifier. Beyond SciFact-style verification, NLI-based models are also used to check the factual consistency of generated text. Laban et al. (2022) show that NLI models detect inconsistencies in summaries more reliably when applied at sentence rather than document granularity; Honovich et al. (2022) meta-evaluate factual-consistency metrics across a diverse set of datasets and find that large-scale NLI-based methods perform strongly; and Zha et al. (2023) train a single alignment model on data from many tasks that is competitive with much larger metrics. These results support the use of an NLI model as the semantic verifier in this study.

Negation is a specific linguistic phenomenon of direct relevance to this study's gate design. Hossain et al. (2020) show, through a systematic analysis of major NLI benchmarks, that negation is underrepresented in standard training data and that state-of-the-art transformer NLI models perform substantially worse on inference pairs constructed specifically to require correct negation handling than on the benchmarks' original test sets. This finding cuts both ways for the present study's design: it supports treating a negation mismatch as a strong, largely model-independent signal of likely contradiction, since human-verified negation asymmetries are a reliable cue by construction, but it also means the MiniLM-NLI model used as the cascade's fallback cannot be assumed to handle the negation cases it does see especially well, a caveat relevant to interpreting the error analysis in Section 4.3.

2.4 Temporal Knowledge and Source Staleness

A separate line of work addresses the temporal validity of retrieved or memorised knowledge rather than its semantic consistency. Facts that are true at one point in time, such as an officeholder, a champion, or a company's leadership, can become false without any change in phrasing, meaning a citation can be internally consistent with a claim while still being outdated. FreshQA (Vu et al., 2024) operationalises this distinction by categorising questions according to how frequently their correct answer changes and by re-evaluating the same questions at different points in time, which is the design this study draws on when constructing its staleness dataset. Unlike contradiction, which can in principle be detected from the text of the claim and evidence alone, staleness typically requires some notion of the evidence's timestamp relative to the query time, which motivates a metadata-based rather than a purely semantic detection strategy.

FreshQA is one of several benchmarks constructed to probe this failure mode. RealTimeQA (Kasai et al., 2022) takes a complementary approach, running a continuously updated weekly quiz drawn from current news so that systems must be evaluated against the state of the world at the time a question is asked rather than a fixed historical snapshot; the authors report that models often return outdated answers when retrieved evidence does not clearly update their prior belief, a failure mode closely related to the staleness problem this study addresses. Dhingra et al. (2022) approach the same underlying issue from the model-internals side, introducing TempLAMA and a time-conditioned language-modelling objective to probe and improve how models represent facts that expire or change, which reinforces the view that source recency is a property distinct from a fact's semantic content and needs to be modelled or checked explicitly rather than inferred from surface text alone. Luu et al. (2022) show that language-model performance degrades when training and evaluation data come from different time periods, and that the degradation depends on the task and domain. In the retrieval setting, Ouyang et al. (2025) introduce HoH, a dynamic question-answering benchmark built from Wikipedia snapshots, and find that outdated passages substantially reduce answer accuracy even when current information is also retrieved. Together these results motivate checking the recency of cited evidence, but they do not show that source age alone identifies outdated content, which is the question examined in Section 4.4.

2.5 Lightweight and Cascade-Based Verification

Cascade or early-exit architectures address the cost of running an expensive model on every input by first applying a cheaper filter that resolves a subset of cases with high confidence, escalating only the remainder to the expensive model. Within NLP specifically, Xin et al. (2020) introduce DeeBERT, which attaches early classifiers to intermediate transformer layers so that high-confidence inputs can exit

before reaching the final layer, reporting up to approximately 40% inference-time savings with minimal accuracy loss on GLUE-style classification tasks; this is architecturally different from the present study's gate, which operates on hand-engineered lexical features ahead of a separate model rather than on intermediate representations within a single model, but the underlying motivation, that many inputs do not need a model's full capacity to classify correctly, is the same. This strategy also has precedent, more recently, in routing strategies for large language model inference. Chen et al. (2024) introduce FrugalGPT, which learns which combination of cheaper and more expensive LLM APIs to call for a given query and report that a learned cascade can match the accuracy of the best individual model at up to 98% lower cost, or improve accuracy over the best individual model at equal cost; their result depends on a learned, query-level stopping rule rather than a fixed heuristic gate. The present study evaluates a fixed, interpretable negation-mismatch gate rather than a learned one, which offers transparency about exactly which cases are resolved without the expensive model, and reports explicitly (Sections 4.2 and 5.3) that the gate did not preserve accuracy, rather than reporting only its cost savings. Most directly relevant to this study's aims, Tang et al. (2024) introduce MiniCheck, a small fact-checking model distilled from GPT-4-generated synthetic training data that matches GPT-4-level grounding-verification accuracy at reported cost reductions of roughly two orders of magnitude, evaluated across a unified benchmark spanning eleven fact-checking and grounding datasets. MiniCheck pursues efficiency by training a single small model to replace an expensive one outright, whereas this study pursues it by gating a fixed-size expensive model with a cheaper filter; MiniCheck's results suggest that a learned, purpose-trained lightweight verifier may be a stronger foundation for a future cascade gate than the hand-engineered lexical features used here, a direction discussed further in Sections 5.7 and 7. Ding et al. (2024) extend this idea to learned query routing: their hybrid system sends each query to a small or a large model according to its predicted difficulty and a quality target that can be adjusted at test time, and reports up to 40% fewer calls to the large model with no loss in response quality. Such learned, tunable routers are a natural successor to the fixed lexical gate evaluated here. Applied to citation verification specifically, a cascade design raises a question this study investigates directly: which cheap signal should serve as the gate, since gates that trigger too rarely leave most cases to the expensive model with little efficiency gain, while gates that trigger too often can override cases the expensive model would have classified correctly.

2.6 Research Gap

Contradiction detection, temporal staleness detection, and inference-cost-aware cascade design have each been studied, but largely in isolation from one another. Fact-verification research reports accuracy without addressing inference cost; efficiency-oriented cascade research is rarely evaluated on a genuine contradiction-detection benchmark with an explicit false-alarm and missed-contradiction breakdown; and temporal staleness detection is rarely paired with a semantic contradiction pipeline in the same evaluation. Few studies, to our knowledge, jointly examine contradiction detection, staleness detection, and computational efficiency within a single, transparently reported pipeline. This study takes a bounded step toward that gap: it reports contradiction-detection accuracy, the accuracy cost of a lexical shortcut, an accuracy-preserving throughput optimisation, and a small historical-evidence check of what source age can and cannot establish. These components are evaluated separately, not as an integrated pipeline.

3.0 MATERIALS AND METHODS

3.1 Research Design

This quantitative study combines a SciFact contradiction benchmark, an independent CPU batching benchmark, and a small historical-evidence case study based on FreshQA questions. The experiments are separate: no end-to-end RAG system or jointly calibrated contradiction-and-staleness classifier is evaluated.

3.2 Dataset Construction

3.2.1 SciFact Contradiction Dataset

The SciFact corpus (Wadden et al., 2020) was obtained from its official repository and comprises 5,183 corpus documents drawn from PubMed abstracts, together with claim files spanning training, development, and test splits. Claim-evidence annotations were extracted from the training and development splits, since the test split does not release gold labels. After grouping annotations by claim and cited document and combining multiple evidence sentences belonging to the same claim-document-label combination into a single citation-verification case, 773 grouped claim-document cases were obtained, spanning 693 unique claims, comprising 508 SUPPORT and 265 CONTRADICT cases with no missing evidence text. This 773-case set constitutes the full-scale SciFact evaluation set used in Section 4.5. The 200-case pilot used in Sections 4.1 to 4.3 was drawn deterministically, not at random: after sorting the annotated cases by label, claim identifier and document identifier, the first 100 SUPPORT and the first 100 CONTRADICT cases were taken. The pilot is therefore a contiguous block of the sorted data, is drawn from the same annotated pool as the 773-case set, and is not independent of it.

3.2.2 FreshQA Staleness Dataset

Two 600-question FreshQA answer collections, associated with November 2025 and April 2026 comparison snapshots, were matched by question text. Of 574 matched questions, 89 had different answer strings; 51 fell below a 0.60 answer-similarity screening threshold, and 50 were selected as candidate changes. Answer differences and dated URLs were treated as candidates rather than proof of a historical factual change. The comparison interval was 148 days. The two collections should not be described as verified archived source-page snapshots.

3.2.3 Pilot Dataset

The experimental worksheet expanded the 50 candidate changes into 100 provisional CURRENT/STALE records, of which 76 (38 pairs) had been designated as verified in an earlier screening. Those 76 records are not used for any performance claim in this paper: their labels follow from the snapshot design (CURRENT at 0 days, STALE at 148 days), and the source-level verification of each pair could not be independently re-established for this submission. The FreshQA analysis therefore uses five individually checked cases: four changed answers (FRESHQA-003, -004, -005 and -008) and one unchanged control (FRESHQA-CONTROL-005). The SciFact evaluation remains a separate 200-case balanced sample.

3.3 Data Verification and Inclusion Criteria

The historical review checked whether an old and a later answer were supported by time-specific evidence rather than merely differing in wording. Four candidate changes were checked against archived Wikipedia revisions and, where available, authoritative dated records. The reviewer notes documenting these checks were reconstructed from the review record on 1 November 2026 rather than written contemporaneously, and the archived evidence URLs for all five cases are listed in Appendix A (Table A1) so that each check can be repeated. The four changes concern the next Walt

Disney Animation Studios film, ATP men's singles number one, the most recent Academy Award for Best Animated Feature, and TIME's Entertainer of the Year. A fifth apparent change concerning the largest single-tournament tennis payday was excluded because the purported new event predated the old snapshot. Four additional unchanged-control candidates were not accepted because the retrieved excerpts did not establish the relevant historical claim.

One unchanged control, the status of the P versus NP problem, was supported by archived Clay Mathematics Institute pages captured on 16 November 2025 and 21 April 2026, each explicitly identifying the problem as unsolved. The first capture precedes the nominal old snapshot by eight days. The resulting five-case subset is manually corroborated but highly selected (four changes, one unchanged), and cannot support generalizable accuracy estimates or an age-sensitivity curve.

For reproducibility, the case-study records preserve question, old and current answers, reference labels, archived evidence URLs, and review provenance. Review decisions should not be confused with independent prospective model predictions; the source-age rule is evaluated separately below.

3.4 Proposed Verification System

Figure 1 depicts the architecture evaluated in this study. The contradiction branch scores every claim–evidence pair with MiniLM-NLI and forces a binary SUPPORT/CONTRADICT decision; in the throughput experiment, inputs are sorted by length and processed in batches of four. The original design, in which a negation mismatch bypassed MiniLM and directly returned CONTRADICT, was empirically less accurate than MiniLM alone and is reported as an ablation (Section 3.5.6 and Table 8). The temporal branch is an age-based REVIEW alert, not a factual STALE verdict. The two branches were evaluated separately and were not tested as an integrated deployed RAG pipeline.

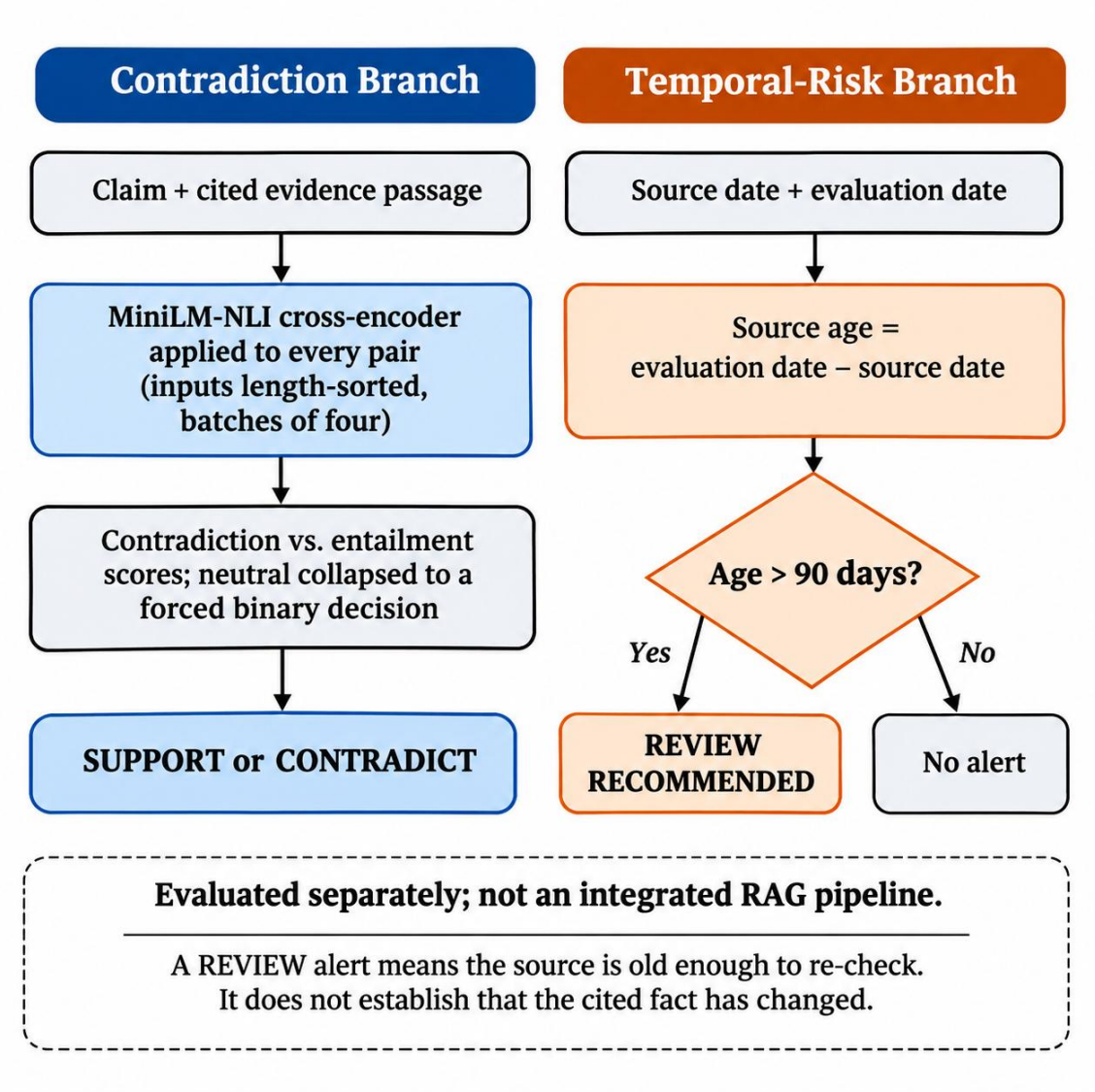


Figure 1: Verification architecture evaluated in this study. The MiniLM-NLI cross-encoder evaluates every claim–evidence pair, with length-sorted batches of four used in the throughput experiment. A separate source-age-based screening rule generates a REVIEW alert when the source age exceeds 90 days. The two branches were evaluated independently and were not integrated into or deployed as a complete retrieval-augmented generation (RAG) pipeline. The original negation-first bypass is included only as an ablation study (Table 8).

3.5 Contradiction-Detection Approaches

Six contradiction-detection approaches were implemented and compared on the SciFact pilot, evaluated using five-fold, claim-grouped, stratified cross-validation (Section 3.7).

3.5.1 Majority Baseline

A grouped majority-class baseline predicts the more frequent class in the training fold for every test case, providing a floor against which more sophisticated methods are compared.

3.5.2 TF-IDF and Logistic Regression

Claim and evidence text were concatenated ("CLAIM: ... EVIDENCE: ...") and represented using a TF-IDF vectoriser (unigrams and bigrams, minimum document frequency of 2, maximum 10,000 features, sublinear term-frequency scaling), on which a class-weighted logistic regression classifier was trained.

3.5.3 Eight-Feature Relationship Model

An interpretable feature-based classifier was built from eight hand-engineered features describing the claim-evidence relationship: cosine similarity between TF-IDF vectors, Jaccard token overlap, claim-token coverage by evidence tokens, a binary negation-mismatch indicator (whether exactly one of claim or evidence contains a negation term from a fixed lexicon of ten common negation words), a binary numerical-mismatch indicator (whether claim and evidence both contain numbers but share none in common), claim token count, evidence token count, and the absolute difference between the two token counts. Features were standardised and passed to a class-weighted logistic regression classifier.

3.5.4 Compact Mismatch Model

A further reduced model retains only the two mismatch features, negation mismatch and numerical mismatch, from the eight-feature set, trained as a standalone class-weighted logistic regression classifier. This model is deliberately minimal: it requires no learned text embeddings or similarity computation at inference time beyond simple lexical lookups, and its serialised size (scaler, classifier, and negation lexicon) is 1,670 bytes (0.0016 MB).

3.5.5 MiniLM-NLI

A pretrained cross-encoder NLI model (cross-encoder/nli-MiniLM2-L6-H768) was applied to each evidence-claim pair, producing contradiction, entailment, and neutral scores. Two readings of this model's output were used: a raw three-way prediction (used to characterise the model's neutral/abstention rate) and a forced binary prediction comparing only the contradiction and entailment scores, used for direct comparison against the other binary classifiers and against the CONTRADICT/SUPPORT-only gold labels in this study.

3.5.6 Negation-First Cascade

The original negation-first cascade directly classified a claim–evidence pair as CONTRADICT whenever a fixed negation lexicon differed between claim and evidence; other pairs were passed to MiniLM-NLI. This is retained as an ablation and negative finding, not as the recommended production configuration. The revised efficiency approach instead retains the same MiniLM scoring and forced-binary decision for every pair, sorts inputs by approximate length, and processes them in batches of four, restoring outputs to their original order.

3.6 Source-Age Alert Rule

For a source captured at date t_s and evaluated at t_e, the age-only alert is A = 1[(t_e − t_s) > 90 days]. A = 1 means REVIEW RECOMMENDED; it does not mean the source has been proven factually obsolete. The original implementation used the output labels STALE and CURRENT; those labels are retained only when describing its historical code.

$$\text{STALE} = \begin{cases} 1, & \text{if source age} > 90 \text{ days} \\ 0, & \text{otherwise} \end{cases}$$

The 90-day threshold was used in the original notebook and was not re-optimized on the five corroborated cases. The original threshold sweep on provisionally labelled 0-day and 148-day records was structurally separable by age and is not evidence of factual-change detection. In the revised evaluation, all five historical sources are 148 days old, including the unchanged control.

3.7 Experimental Setup

The trained lightweight classifiers were evaluated with five-fold StratifiedGroupKFold cross-validation grouped by SciFact claim identifier (random seed 42); 95% confidence intervals were obtained by claim-grouped bootstrap resampling (5,000 iterations). MiniLM-NLI and the negation-first cascade involve no training and were applied directly to every case. The new batch-throughput experiment processed the same 200 SciFact pairs with a warmed CPU MiniLM implementation in five repeated sequential and five repeated length-sorted batch-four runs. We report mean total elapsed time, standard deviation and exact prediction agreement. These measurements quantify offline workload throughput, not per-query interactive latency. Historical source review used archived pages and documented reviewer decisions.

3.8 Evaluation Metrics

Contradiction performance is reported as accuracy, macro-F1 and class-specific recall. The earlier gate latency and invocation-rate measurements are retained as historical ablations. The revised efficiency comparison reports total elapsed time over 200 pairs, repetition-level variation and exact prediction agreement. The FreshQA case study reports case-level alerts and a confusion matrix only; given n = 5 and the 4:1 label imbalance, no summary accuracy metric and no inferential or generalization claim is made.

3.9 Ethical Considerations

This study used only publicly available benchmark datasets, namely SciFact and FreshQA. It did not involve human participants, personal information, private records, biological samples, or direct interaction with individuals. Therefore, formal ethical clearance was not required. The study was conducted in accordance with accepted principles of research integrity, responsible data use, and the applicable terms and licences of the original datasets.

4.0 RESULTS

4.1 Pilot Contradiction-Detection Performance

Figure 2 compares the six contradiction-detection approaches by macro-F1 on the 200-case balanced pilot. MiniLM-NLI has the highest macro-F1, and the cascade and the trained lightweight classifiers fall below it.

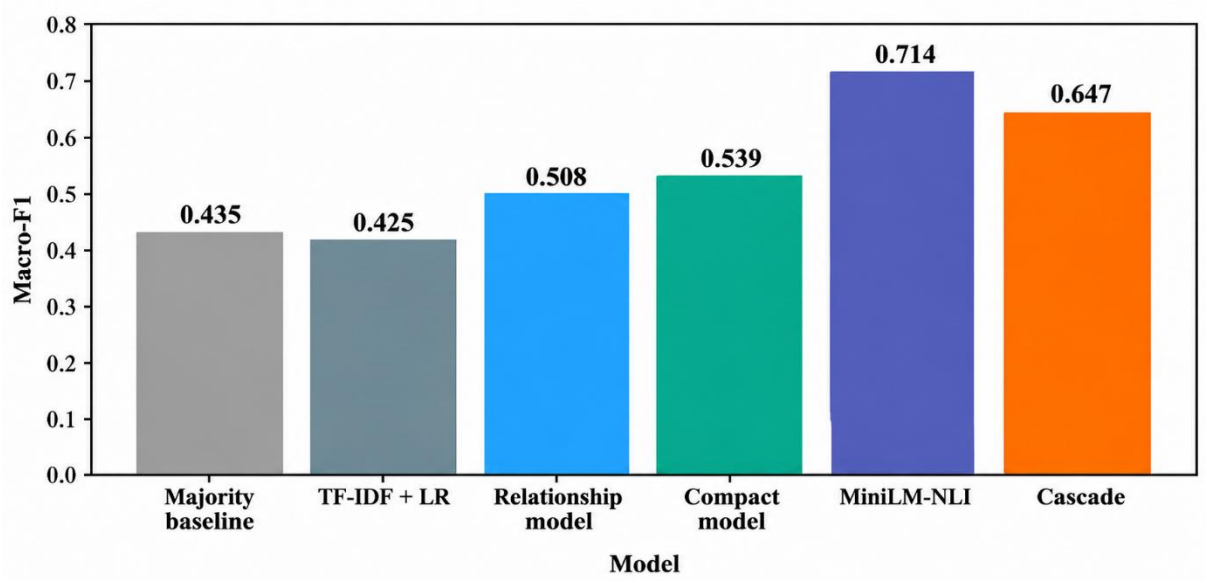


Figure 2: Macro-F1 comparison across the six contradiction-detection approaches on the 200-case balanced SciFact pilot.

Table 1 gives the corresponding accuracy and macro-F1 values for each approach, so that the differences visible in Figure 2 can be read as numbers.

Table 1: Contradiction-detection performance on the 200-case balanced SciFact pilot.

Model	Accuracy	Macro-F1	Contradict recall	Support recall
Grouped majority baseline	0.440	0.435	—	—
TF-IDF + Logistic Regression	0.425	0.425	0.430	0.420
Eight-feature relationship model	0.510	0.508	0.440	0.580
Compact mismatch model	0.555	0.539	0.370	0.740
MiniLM-NLI (forced binary)	0.720	0.714	0.860	0.580
Negation-first cascade	0.665	0.647	0.890	0.440

MiniLM-NLI achieved the highest accuracy and macro-F1 of the six approaches (72.0% accuracy, macro-F1 0.714), with a 95% bootstrap confidence interval of [0.657, 0.778] for accuracy and [0.651, 0.773] for macro-F1. Its improvement over the compact mismatch model was significant, with a 95% CI on the accuracy difference of [0.067, 0.259] and on the macro-F1 difference of [0.081, 0.268]. Among the non-NLI approaches, the compact two-feature mismatch model (55.5% accuracy) outperformed both the eight-feature relationship model (51.0%) and TF-IDF with logistic regression

(42.5%), which performed close to, and in fact marginally below, the majority baseline (44.0%). A feature-set ablation of the eight-feature model (Table 2) confirms that the two mismatch features carry most of the model's discriminative power.

Table 2 reports how the relationship model's performance changes with the feature set it is given, to show which features carry the signal.

Table 2: Feature-set ablation of the relationship model on the SciFact pilot.

Feature configuration	Accuracy	Macro-F1
Mismatch features only	0.555	0.539
All eight features	0.510	0.508
Lexical similarity only	0.505	0.501
Without negation mismatch	0.485	0.485
Length features only	0.480	0.480

Notably, the mismatch-only feature subset outperforms the full eight-feature model, suggesting that the lexical-similarity and length features add noise rather than signal on this pilot, and that negation mismatch in particular is the most useful single feature (removing it reduces accuracy from 55.5% to 48.5%).

4.2 Accuracy–Latency Trade-off

Figure 3 plots accuracy against latency for the four approaches whose inference time was measured. It shows what each speed-up costs in accuracy, and is read together with Table 3, which lists the underlying latency and model-size figures.

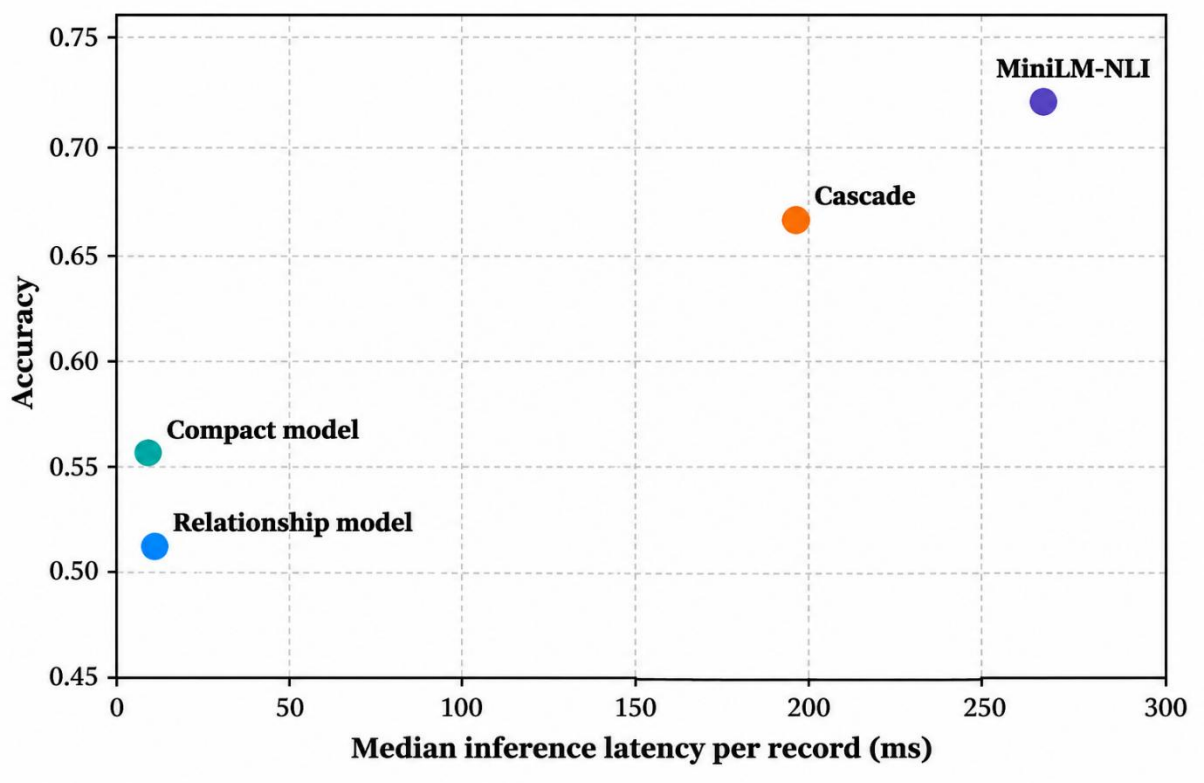


Figure 3: Accuracy–latency trade-off across the four timed approaches.

In the original latency ablation, the negation-first cascade reduced measured median per-record time from 268.57 ms to 196.31 ms (26.9%) but also lowered SciFact accuracy from 72.0% to 66.5%. This was an unfavorable accuracy–latency trade-off, not a successful accuracy-preserving cascade. In a separate five-repeat CPU benchmark on the same 200 cases, sequential MiniLM processing required a mean 51.779 s (SD 0.653 s), whereas length-sorted batch-four processing required 46.317 s (SD 1.379 s): a 10.55% reduction in total elapsed time. Every batch prediction agreed with its sequential counterpart (200/200), preserving 72.0% accuracy. These throughput figures are not directly interchangeable with the earlier per-record latency measurements.

Table 3 gives the measured median latency and model size for each approach.

Table 3: Latency and model size by approach (200-case SciFact pilot).

Model	Median latency (ms)	Model size (MB)
Compact mismatch model	5.24	0.0016
Eight-feature relationship model	7.83	0.168
MiniLM-NLI (forced binary, warmed)	268.57	313.27
Negation-first cascade	196.31	313.27

The compact model has a much smaller parameter footprint than MiniLM, but the lexical shortcut did not preserve MiniLM accuracy. The batch-four implementation keeps the MiniLM model and its storage requirements unchanged; its demonstrated benefit is reduced total processing time for the tested offline workload.

4.2.1 Accuracy-Preserving Batch-Throughput Experiment

Table 4 reports the five-repeat comparison. Across five repeated CPU measurements, sequential MiniLM inference processed the 200 SciFact cases in 51.779 ± 0.653 s (mean ± SD). Length-sorted batch-four inference required 46.317 ± 1.379 s, reducing mean total elapsed time by 10.55% (5.462 s). The two methods agreed on all 200 predictions, and both achieved 72.0% accuracy. The amortized processing times were 258.897 ms and 231.583 ms per case, respectively; these are workload-average throughput measures, not single-request latency estimates. The comparison does not imply fewer MiniLM evaluations or a smaller model.

Table 4: Batch-throughput comparison for MiniLM-NLI on the 200 SciFact pilot pairs (five repeated CPU runs per configuration). Total elapsed time is workload throughput, not per-request latency.

Metric	Sequential (batch size 1)	Length-sorted (batch size 4)
Repeated CPU runs	5	5
Mean total time, 200 pairs (s)	51.779	46.317
Standard deviation (s)	0.653	1.379
Amortised time per case (ms)	258.897	231.583
Mean total-time reduction	—	10.55% (5.462 s)
Accuracy (200 cases)	72.0%	72.0%
Predictions identical to sequential	reference	200 / 200

4.3 MiniLM-NLI and Cascade Error Patterns

Figure 4 shows the confusion matrices of MiniLM-NLI and of the negation-first cascade, so that the kinds of error each makes can be compared; the same counts appear in Table 5.

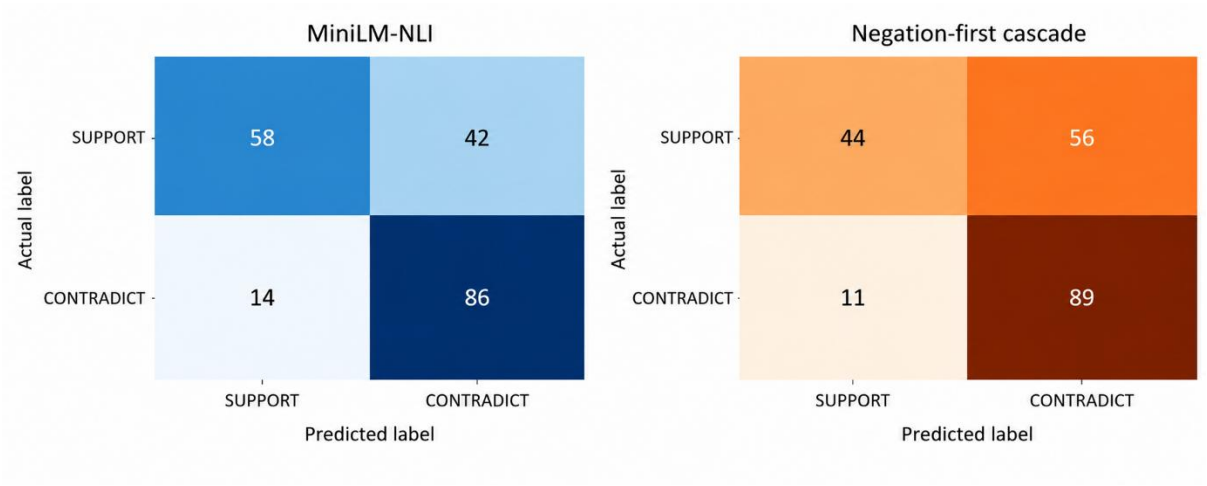


Figure 4: Confusion matrices for MiniLM-NLI (forced binary) and the negation-first cascade on the 200-case SciFact pilot.

MiniLM-NLI's confusion matrix on the 200-case pilot (Table 5) shows a bias toward predicting CONTRADICT: of 100 true SUPPORT cases, 58 were correctly predicted and 42 misclassified as CONTRADICT, while of 100 true CONTRADICT cases, 86 were correctly predicted and 14 misclassified as SUPPORT. This asymmetry is consistent with the model's neutral rate under its raw three-way prediction: 50.5% of cases were assigned a raw NEUTRAL label before the forced-binary collapse was applied, meaning roughly half of the pilot's cases sat in a region of genuine model uncertainty that the forced-binary comparison resolves toward whichever of contradiction or entailment scored marginally higher, rather than reflecting confident discrimination. The negation-first cascade correctly classifies 44 of 100 true SUPPORT cases and 89 of 100 true CONTRADICT cases ((44+89)/200 = 0.665 accuracy), reflecting the same higher-contradiction-recall, lower-support-recall pattern discussed in Section 5.4.

Table 5: Confusion matrices, 200-case SciFact pilot.

MiniLM-NLI (forced binary)

	Predicted SUPPORT	**Predicted CONTRADICT**
Actual SUPPORT	58	42
Actual CONTRADICT	14	86

Negation-first cascade

	Predicted SUPPORT	**Predicted CONTRADICT**
Actual SUPPORT	44	56
Actual CONTRADICT	11	89

An error-stage analysis of an earlier any-mismatch cascade variant (reported here because its per-case decision stage was logged in full during experimentation) classified its 70 misclassifications on the pilot into 57 false contradiction alarms (a true SUPPORT case predicted CONTRADICT) and 13 missed

contradictions (a true CONTRADICT case predicted SUPPORT). Of these errors, 28 originated at the compact-gate stage and 42 at the MiniLM-fallback stage. The majority of false contradiction alarms occurring at the MiniLM-fallback stage (24 of 31) corresponded to cases where the raw NLI prediction was NEUTRAL before being forced toward CONTRADICT or SUPPORT, reinforcing that much of the cascade's residual error traces back to genuinely ambiguous cases for the NLI model rather than to the gate's own mistakes. This error-stage breakdown corresponds to the any-mismatch variant rather than the final negation-only cascade shown in Table 5, since only the former had its per-case decision stage logged; this distinction is noted for transparency.

4.4 Source-Age Alert Performance

Figure 5 shows the age of the source in each of the five reviewed FreshQA cases against the 90-day alert threshold. Every case falls above the threshold, so every case is flagged for review, including the unchanged control.

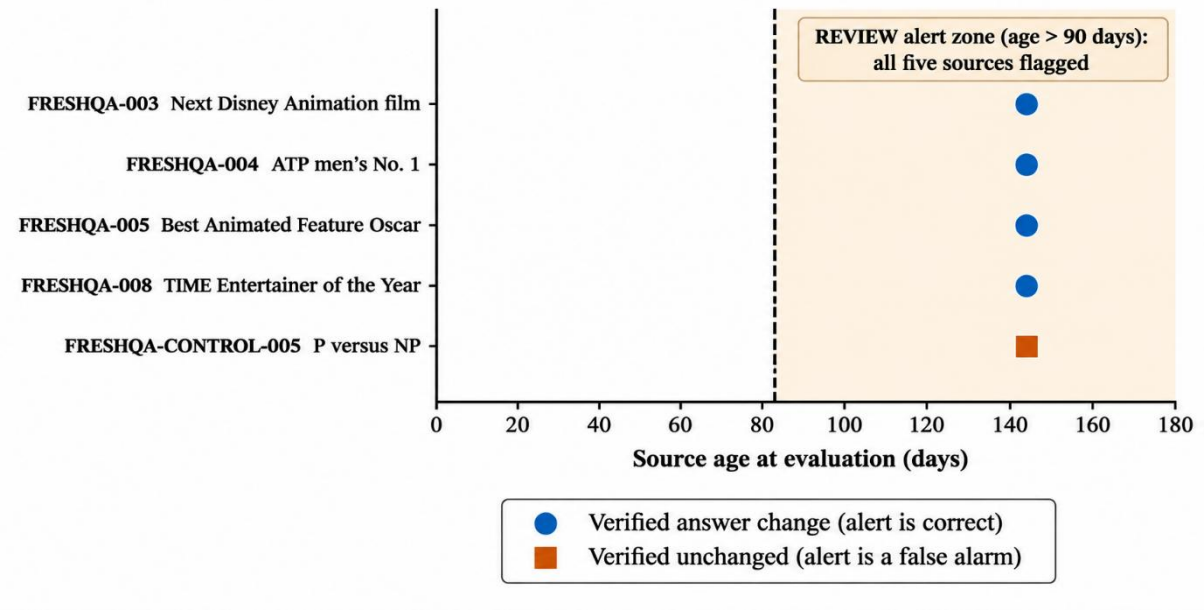


Figure 5: Source-age alert outcomes for the five manually corroborated FreshQA cases. All five sources were 148 days old, above the 90-day threshold, so every case received a REVIEW alert, including the unchanged P versus NP control (Table 6).

Table 6 summarises the same alert outcomes by case type, separating the four verified changes from the one unchanged control.

Table 6: Age-only alert outcomes for five manually corroborated FreshQA cases (four changes, one unchanged control).

Metric	Observed	Interpretation	Scope
Verified changed cases	4	All flagged for review	Five selected cases
Verified unchanged controls	1	Also flagged for review	P versus NP
Alert threshold	90 days	Fixed in original implementation	Not re-tuned
Source age	148 days	Identical for all cases	No age variation

Confusion matrix	TN=0; FP=1; FN=0; TP=4	Age-only alert treated as change proxy	Not factual verification

For the five independently reviewed historical-source cases, the 90-day rule issued a REVIEW alert for all five, because each source was assigned the same 148-day age (Figure 5, Table 6). Four alerts coincided with corroborated factual changes; one was a false alert on the unchanged P versus NP control. Because every case receives the same alert, the rule carries no information that separates changed from unchanged facts of identical age; the 4-of-5 agreement reflects only the 4:1 mix of the selected cases and is not a performance estimate. The result illustrates that a timestamp threshold cannot discriminate between changed and unchanged facts and is not a validated temporal fact-checker. The five cases are too few and too selected for reliable performance estimation.

4.5 Full SciFact Evaluation

To check whether the pilot-scale pattern recurs on a larger set, MiniLM-NLI and the negation-first cascade were additionally evaluated on the full, class-imbalanced 773-case SciFact set (508 SUPPORT, 265 CONTRADICT). This set shares its annotated pool with the pilot and is therefore not an independent held-out sample.

Table 7: Full-scale SciFact evaluation (773 cases).

Model	Accuracy	Balanced accuracy	Macro-F1	Support recall	Contradict recall
MiniLM-NLI (forced binary)	0.688	0.738	0.686	0.581	0.894
Negation-first cascade	0.625	0.694	0.625	0.474	0.913

The negation gate activated on 202 of the 773 cases, so MiniLM-NLI was invoked on the remaining 73.9% of cases (an NLI invocation rate of 0.739, or 26.1% of NLI calls avoided). The raw NLI neutral rate on this larger set was 58.3%, close to the pilot's 50.5%, suggesting that the high neutral rate was not confined to the pilot sample. The cascade's accuracy was 6.3 percentage points lower than full MiniLM (95% bootstrap CI on the accuracy difference: [-0.083, -0.045]) and its macro-F1 was correspondingly lower (95% CI: [-0.081, -0.044]), while its contradiction recall was 1.9 percentage points higher (91.3% versus 89.4%), replicating the pilot-scale pattern in which the cascade trades some overall accuracy, driven mainly by lower support recall, for higher contradiction recall.

4.6 Gate-Design Ablation

Table 8 reports the four candidate cascade-gate designs evaluated on the pilot (Section 3.5.6), sorted by macro-F1.

Table 8: Gate-design ablation, 200-case SciFact pilot.

Gate	Gated cases	NLI call rate	Accuracy	Macro-F1	Contradict precision	Contradict recall	Support recall
Both mismatches	1	0.995	0.720	0.714	0.672	0.86	0.58

Numerical mismatch only	9	0.955	0.705	0.700	0.661	0.84	0.57
Negation mismatch only	57	0.715	0.665	0.647	0.614	0.89	0.44
Any mismatch	65	0.675	0.650	0.632	0.604	0.87	0.43

The “both mismatches” gate activates on only a single case in the pilot (an NLI call rate of 0.995, essentially equivalent to running MiniLM on every case), so its accuracy and macro-F1 are, by construction, nearly identical to full MiniLM. The “negation mismatch only” gate, used as the original primary cascade, trades a further 5.5-point accuracy reduction relative to full MiniLM for a substantially lower NLI call rate (0.715) than the “both mismatches” or “numerical mismatch only” gates. This ablation makes explicit a design choice rather than a free efficiency gain: broader gates (any mismatch, negation only) reduce NLI invocation more but cost more accuracy, chiefly through a drop in support recall (correctly recognising true SUPPORT cases), while narrower gates (numerical only, both mismatches) preserve accuracy closely but barely reduce the number of MiniLM calls.

5.0 DISCUSSION

5.1 Principal Findings

The primary revised finding is that full MiniLM-NLI remains the more accurate tested contradiction verifier on the 200-case SciFact sample (72.0% accuracy, macro-F1 0.7144), whereas a negation-first bypass degrades accuracy to 66.5%. Length-sorted batch-four inference reduces mean total CPU processing time by 10.55% without changing predictions. In the FreshQA historical case study, source age flags four genuinely changed cases but also one unchanged fact, demonstrating the difference between temporal-risk screening and factual-change verification.

5.2 Contradiction Detection

MiniLM-NLI's advantage over the lexical and feature-based baselines is consistent with the broader NLI literature: a cross-encoder that jointly attends over the claim and evidence text can capture negation scope, quantitative comparisons, and other interactions that similarity-based or hand-engineered features approximate only coarsely. This is visible directly in the feature ablation (Table 2), where the negation-mismatch feature alone accounts for most of the compact model's performance, yet the compact model still falls 16.5 accuracy points short of MiniLM, indicating that a binary negation flag captures only part of what a full contextual model discriminates.

5.3 Efficiency–Performance Trade-off

The original negation-first gate saved model calls but caused false contradiction alarms and a 5.5-percentage-point accuracy loss on the balanced SciFact pilot. Accordingly, we do not present that gate as a preferred deployment option. The separate length-sorted batch-four implementation preserves every baseline prediction and reduces total CPU workload time from 51.779 s to 46.317 s on average across five runs (10.55%). This is a throughput optimization under batchable conditions; no online per-request latency, tail-latency or end-to-end RAG improvement was demonstrated.

5.4 Contradiction-Recall Trade-off

Contradiction recall was higher under the cascade than under full MiniLM on both the pilot (89.0% versus 86.0%, three additional cases out of 100, a difference too small to interpret on this sample) and the full set (91.3% versus 89.4%), because the negation gate deterministically labels any negation-mismatched case as CONTRADICT, removing the possibility that MiniLM's own uncertainty on such cases resolves toward SUPPORT. This came at the cost of support recall, which fell from 58.0% to 44.0% on the pilot and from 58.1% to 47.4% on the full set, since the same deterministic override also converts some true SUPPORT cases containing a negation mismatch into false CONTRADICT predictions. Whether this trade favourably depends on the relative cost, in a given RAG deployment, of a missed contradiction (a false citation reaching the user) against a false contradiction alarm (a correct citation being unnecessarily flagged or suppressed). Because the shortcut also lowered overall accuracy and support recall, we do not recommend it as a default.

5.5 Temporal Verification

The original FreshQA pilot assigned CURRENT to zero-day records and STALE to 148-day records by construction, so perfect threshold separation was a property of the label design rather than independent evidence of factual obsolescence. Subsequent historical review supports only four changed examples and one unchanged control. All five sources were 148 days old at the evaluation point, and the 90-day rule flagged all five. Its apparent 4-of-5 agreement reflects only the 4:1 mix of changed and unchanged cases; the false alert on P versus NP shows that source age cannot by itself establish factual change. Future temporal verification requires content-aware comparison against dated evidence and a substantially larger independently labelled dataset. Evidence from RAG benchmarks points the same way: Ouyang et al. (2025) find that outdated passages reduce answer accuracy even when current evidence is retrieved alongside them, which supports flagging old sources for review but does not by itself show that age identifies which sources are outdated.

5.6 Implications for RAG Systems

A practical citation-verification workflow should separate claim–evidence semantic consistency from temporal-risk triage. Full MiniLM can verify semantic relations, with batching used when multiple citations can be processed together. Old timestamps can trigger retrieval of newer authoritative evidence, but should not be treated as a definitive stale-fact label. The present study does not evaluate this combined workflow end-to-end.

5.7 Comparison with Previous Studies

Direct numerical comparison against the SciFact leaderboard is not straightforward, since published SciFact systems (e.g., the VeriSci baseline introduced alongside the dataset (Wadden et al., 2020) and later systems such as MultiVerS (Wadden et al., 2022b)) are typically evaluated on the full three-way SUPPORT/CONTRADICT/NOT ENOUGH INFO pipeline, including retrieval and rationale selection, whereas this study evaluates label prediction alone on a fixed, pre-retrieved, two-way SUPPORT/CONTRADICT subset; the two setups are not evaluated under the same conditions, and reporting a side-by-side number without matching those conditions would misrepresent both systems. The comparison that can be made honestly is qualitative: this study's MiniLM-NLI accuracy (72.0% pilot, 68.8% full set) sits well below the accuracy ceiling reported by fully supervised, task-fine-tuned SciFact systems in the literature, which is expected given that MiniLM-NLI is used here off-the-shelf rather than fine-tuned on SciFact training data, and is consistent with this study's framing of MiniLM-

NLI as a general-purpose, readily deployable verification component rather than a claim to state-of-the-art SciFact performance.

On efficiency, Chen et al. (2024) report that a learned LLM cascade can match the accuracy of the best individual model at up to 98% lower cost, or improve accuracy at equal cost, on their evaluated tasks. The cascade evaluated in this study achieves a substantially smaller latency reduction (26.9% pilot, comparable NLI-call savings on the full set) at a non-trivial accuracy cost, rather than FrugalGPT's reported cost reduction at no accuracy cost. This gap is consistent with the difference in gating strategy: FrugalGPT learns a query-level stopping rule from data, whereas the gate evaluated here is a single fixed heuristic (negation mismatch) chosen for interpretability rather than optimised for the accuracy-cost trade-off, which the gate-design ablation in Section 4.6 shows trades call savings against accuracy depending on the mismatch signal chosen. This suggests a learned gate, in the spirit of FrugalGPT, is a plausible direction for narrowing this gap, which we return to in Section 7.

On citation quality more broadly, Gao et al. (2023a) report with ALCE that even strong LLMs fail to fully support their generated claims with cited evidence a substantial fraction of the time, which is consistent with this study's motivating premise that citation verification cannot be treated as solved by retrieval alone. Unlike ALCE, which scores citation quality as part of an end-to-end generation pipeline, this study isolates the verification step itself and evaluates it against pre-existing claim-evidence pairs, which is a narrower but more controlled setting.

The most directly comparable prior system is MiniCheck (Tang et al., 2024), which pursues the same goal, cheap and accurate grounding verification, by a different means: rather than gating a fixed cross-encoder behind a heuristic filter, MiniCheck trains a single small model from scratch on synthetic data designed to teach it the specific error patterns an expensive model would otherwise be needed to catch, and reports GPT-4-level accuracy at roughly two orders of magnitude lower cost on a broad multi-dataset benchmark. This study's original cascade, by contrast, achieved a smaller efficiency gain (26–29% fewer expensive-model calls, not a two-orders-of-magnitude reduction) at a measurable accuracy cost; its gate is interpretable and auditable, but that transparency did not translate into an acceptable accuracy-efficiency trade-off. Neither comparison is evaluated on the other's benchmark, so no single number can be quoted across them; the qualitative lesson is that MiniCheck's results suggest a learned lightweight verifier is likely a stronger long-run direction than a hand-engineered heuristic gate, which this study returns to explicitly as future work in Section 7 rather than treating its own simpler design as a final answer.

6.0 LIMITATIONS

Several limitations qualify the findings. SciFact is a biomedical claim–evidence benchmark and does not establish performance on real RAG-generated citations or other domains. The 200-case pilot was inspected during development and, being the first 100 cases of each class in sorted order, is a contiguous rather than random sample that is not independent of the 773-case set (Section 3.2.1); additional independent evaluation is required. The earlier 773-case SciFact results are exploratory comparisons and do not establish that the lexical cascade is preferable. The revised batch benchmark uses only five repeated CPU runs on one workload and reports total processing time, not online per-request latency or statistical significance. The FreshQA dataset consists of four corroborated changed cases and one unchanged control selected after extensive inspection, with an eight-day gap between one archived capture and the nominal snapshot date. Its small size, strong class imbalance, common

148-day age and review-based selection preclude generalizable accuracy, precision, recall or F1 claims. The earlier 38-pair/76-record set was provisionally screened, not independently source-verified.

The negation-mismatch heuristic is brittle because it ignores negation scope and contextual meaning; the observed accuracy loss confirms the danger of using it to override NLI predictions. The forced-binary MiniLM decision also collapses neutral outcomes. Source-age thresholds do not account for how quickly different facts change and should trigger review rather than assert factual obsolescence. The batch-four optimization requires a queue of citations and may not help latency-sensitive single requests.

Finally, the contradiction and temporal components were assessed separately on static datasets. No end-to-end deployment, user-facing citation correction, retrieval-refresh mechanism or content-aware temporal-change detector was evaluated. The manuscript retains earlier exploratory plots and gate ablations only when clearly labelled as such; they should not be read as confirmatory evidence for the revised approach.

7.0 CONCLUSION AND FUTURE WORK

This study re-examined citation verification in RAG using SciFact contradiction classification and a historically corroborated FreshQA case study. A MiniLM-NLI baseline achieved 72.0% accuracy and 0.7144 macro-F1 on 200 SciFact cases; a negation-first shortcut lowered accuracy to 66.5%. In contrast, length-sorted batch-four inference preserved all baseline predictions while reducing mean total CPU processing time by 10.55% over five runs. The five-case FreshQA review showed that a 90-day source-age rule flagged four changed answers and one unchanged fact, confirming its role as a review trigger rather than a factual-staleness detector. The study therefore supports batching as a measured throughput optimization and cautions against interpreting cheap lexical or temporal heuristics as verified correctness decisions.

Future work should test batch-size and workload sensitivity, online request latency, independent SciFact and cross-domain benchmarks, and learned selective-routing methods with explicit accuracy constraints (for example, Ding et al., 2024). Temporal evaluation should expand to many independently archived changed and unchanged examples across diverse source ages and include direct comparison of dated evidence. A fully integrated RAG verification pipeline remains to be implemented and evaluated.

Declarations

Funding: This research received no external funding.

Competing interests: The authors declare that they have no competing interests.

Ethics approval and consent to participate: Not applicable. This study used only publicly available benchmark datasets (SciFact and FreshQA) and did not involve human participants, personal information, or biological samples. Therefore, formal ethical approval and consent to participate were not required.

Data and code availability: The SciFact and FreshQA datasets used in this study are publicly available. The experimental code, processed results, and historical-source verification records are available from the corresponding author upon reasonable request.

Acknowledgements: The authors have no specific acknowledgements to declare.

Author contributions: Kholofelo Jonase Kgatla contributed to the study conceptualisation, methodology, software implementation, data analysis, validation, visualisation, and preparation of the original manuscript draft. Nthabiseng Michael Moeti contributed to research supervision, methodological review, critical revision, and manuscript editing. Both authors reviewed and approved the final manuscript.

References

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 632–642). Lisbon, Portugal: Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1075

Chen, L., Zaharia, M., & Zou, J. (2024). FrugalGPT: How to use large language models while reducing cost and improving performance. Transactions on Machine Learning Research. https://openreview.net/forum?id=cSimKw5p6R

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dhingra, B., Cole, J. R., Eisenschlos, J. M., Gillick, D., Eisenstein, J., & Cohen, W. W. (2022). Time-aware language models as temporal knowledge bases. Transactions of the Association for Computational Linguistics, 10, 257–273. https://doi.org/10.1162/tacl_a_00459

Ding, D., Mallick, A., Wang, C., Sim, R., Mukherjee, S., Rühle, V., Lakshmanan, L. V. S., & Awadallah, A. H. (2024). Hybrid LLM: Cost-efficient and quality-aware query routing. In International Conference on Learning Representations (ICLR 2024).

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

Gao, T., Yen, H., Yu, J., & Chen, D. (2023a). Enabling large language models to generate text with citations. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 6465–6488). Singapore: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.398

Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., Zhao, V., Lao, N., Lee, H., Juan, D.-C., & Guu, K. (2023b). RARR: Researching and revising what language models say, using language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 16477–16508). Toronto, Canada: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.910

Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Kukliansy, D., Cohen, V., Scialom, T., Szpektor, I., Hassidim, A., & Matias, Y. (2022). TRUE: Re-evaluating factual consistency evaluation. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 3905–3920). Seattle, United States: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.287

Hossain, M. M., Kovatchev, V., Dutta, P., Kao, T., Wei, E., & Blanco, E. (2020). An analysis of natural language inference benchmarks through the lens of negation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 9106–9118). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.732

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 6769–6781). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550

Kasai, J., Sakaguchi, K., Takahashi, Y., Le Bras, R., Asai, A., Yu, X., Radev, D., Smith, N. A., Choi, Y., & Inui, K. (2022). RealTime QA: What's the answer right now? arXiv preprint arXiv:2207.13332. https://doi.org/10.48550/arXiv.2207.13332

Laban, P., Schnabel, T., Bennett, P. N., & Hearst, M. A. (2022). SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. Transactions of the Association for Computational Linguistics, 10, 163–177. https://doi.org/10.1162/tacl_a_00453

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

Liu, N., Zhang, T., & Liang, P. (2023). Evaluating verifiability in generative search engines. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 7001–7025). Singapore: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.467

Luu, K., Khashabi, D., Gururangan, S., Mandyam, K., & Smith, N. A. (2022). Time waits for no one! Analysis and challenges of temporal misalignment. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 5944–5958). Seattle, United States: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.435

Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., & Zhang, T. (2024). RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 10862–10878). Bangkok, Thailand: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.585

Ouyang, J., Pan, T., Cheng, M., Yan, R., Luo, Y., Lin, J., & Liu, Q. (2025). HoH: A dynamic benchmark for evaluating the impact of outdated information on retrieval-augmented generation. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 6036–6063). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.301

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3982–3992). Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Schuster, T., Fisch, A., & Barzilay, R. (2021). Get your Vitamin C! Robust fact verification with contrastive evidence. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (pp. 624–643). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.52

Tang, L., Laban, P., & Durrett, G. (2024). MiniCheck: Efficient fact-checking of LLMs on grounding documents. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 8818–8847). Miami, Florida: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.499

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (pp. 809–819). New Orleans, Louisiana: Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1074

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J., Wei, J., Tar, C., Sung, Y.-H., Zhou, D., Le, Q., & Luong, T. (2024). FreshLLMs: Refreshing large language models with search engine augmentation. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 13697–13720). Bangkok, Thailand: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.813

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., & Hajishirzi, H. (2020). Fact or fiction: Verifying scientific claims. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 7534–7550). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.609

Wadden, D., Lo, K., Kuehl, B., Cohan, A., Beltagy, I., Wang, L. L., & Hajishirzi, H. (2022a). SciFact-Open: Towards open-domain scientific claim verification. In Findings of the Association for Computational Linguistics: EMNLP 2022 (pp. 4719–4734). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.347

Wadden, D., Lo, K., Wang, L. L., Cohan, A., Beltagy, I., & Hajishirzi, H. (2022b). 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). Seattle, United States: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-naacl.6

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. Advances in Neural Information Processing Systems, 33. https://doi.org/10.48550/arXiv.2002.10957

Williams, A., Nangia, N., & Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (pp. 1112–1122). New Orleans, Louisiana: Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1101

Xin, J., Tang, R., Lee, J., Yu, Y., & Lin, J. (2020). DeeBERT: Dynamic early exiting for accelerating BERT inference. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 2246–2251). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.204

Zha, Y., Yang, Y., Li, R., & Hu, Z. (2023). AlignScore: Evaluating factual consistency with a unified alignment function. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 11328–11348). Toronto, Canada: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.634

Appendix A. Evidence URLs for the five FreshQA case-study records

Table A1 lists, for each case, the question, the reference label and the old and current evidence URLs, so that every check can be repeated. Wikipedia entries are permanent revision (oldid) links; the P versus NP control uses Internet Archive captures of the Clay Mathematics Institute page.

Table A1: Evidence URLs for the five FreshQA case-study records (old = November 2025 snapshot or capture; current = April 2026).

Case	Question and reference	Old evidence URL	Current evidence URL
FRESHQA-003 (change)	What is the next Walt Disney Animation Studios film? [Zootopia 2 → Hexed]	https://en.wikipedia.org/w/index.php?oldid=1323921911	https://en.wikipedia.org/w/index.php?oldid=1349554986
FRESHQA-004 (change)	Who is the current ATP top-ranked men's singles tennis player? [Carlos Alcaraz → Jannik Sinner]	https://en.wikipedia.org/w/index.php?oldid=1322937217	https://en.wikipedia.org/w/index.php?oldid=1348673443
FRESHQA-005 (change)	What won the latest Academy Award for Best Animated Feature? [Flow → KPop Demon Hunters]	https://en.wikipedia.org/w/index.php?oldid=1319608800	https://en.wikipedia.org/w/index.php?oldid=1349785785
FRESHQA-008 (change)	What country does the most recent Time Magazine's Entertainer of the Year come from? [South Korea → United States]	https://en.wikipedia.org/w/index.php?oldid=1321492770	https://en.wikipedia.org/w/index.php?oldid=1347466766
FRESHQA-CONTROL-005 (unchanged)	Why are all quickly verifiable problems also quickly solvable? [Unchanged: remains an open problem (capture dates 16 Nov 2025 and 21 Apr 2026)]	https://web.archive.org/web/20251116184700/https://www.claymath.org/millennium/p-vs-np/	https://web.archive.org/web/20260421223445/https://www.claymath.org/millennium/p-vs-np/