The International Society of Data Scientists
Journal of Data Science and Artificial Intelligence
Vol. 5, No. 1. Online ISSN: 2831-4794

A Deterministic Extractive Framework for Evidence-Grounded Scientific Claim Verification

Bhindia Binoy
Independent Researcher

ABSTRACT

Large language models (LLMs) are increasingly applied to scientific question answering, yet their outputs often contain statements lacking explicit evidence. In high-stakes domains such as biomedicine, ensuring traceability to source documents is essential for interpretability and reliability. While retrieval-augmented generation (RAG) systems leverage external documents, most pipelines do not strictly enforce evidence dependence during answer construction.

We introduce Citation-Driven Extractive Claim Verification (CD-ECV), a deterministic, non-generative extractive framework for scientific claim verification — not a hallucination-reduction system for generative LLMs. CD-ECV retrieves biomedical literature via sparse lexical ranking (BM25), applies sentence-level lexical and semantic filtering, and constructs responses exclusively from unmodified evidence spans. Importantly, CD-ECV guarantees source traceability, not factual truth: if retrieved evidence is incorrect or outdated, outputs remain traceable but may not reflect scientific ground truth.

We evaluate CD-ECV on the SciFact benchmark (300 claims labelled SUPPORT, CONTRADICT, or NOT_ENOUGH_INFO) using a corpus of 5183 biomedical passages. Metrics include retrieval recall, evidence selection precision, label accuracy, and abstention rate. All non-abstaining outputs consist entirely of verbatim retrieved evidence spans. These results establish CD-ECV as a deterministic, non-generative extractive baseline for citation-grounded scientific claim verification, providing a transparent and reproducible reference point and enabling future integration with neural validation or generative components.

Keywords: *claim verification, biomedical NLP, retrieval-augmented generation, extractive QA, evidence grounding, BM25, SciFact, zero-shot*

1. INTRODUCTION

Large language models have rapidly advanced tasks such as scientific question answering and biomedical text analysis. Despite this progress, outputs can include statements that are not directly supported by evidence, which in scientific and medical settings may lead to misinformation, flawed interpretations, and potentially harmful decisions. Ensuring reliability in these high-stakes domains therefore requires systems that constrain outputs to verified evidence.

Conventional retrieval-based pipelines aim to improve grounding by consulting external documents during inference. In practice, however, models may ignore retrieved context, rely on internal parametric knowledge, or attach citations that do not genuinely support the claims being made. These challenges are particularly pronounced in biomedical domains, where technical terminology and multi-step reasoning are common.

To address these gaps, we introduce Citation-Driven Extractive Claim Verification (CD-ECV), a deterministic, evidence-first framework that strictly enforces grounding through extractive answer construction. CD-ECV is not a system for reducing hallucination in generative LLMs — it is a non-generative extractive baseline for citation-grounded scientific claim verification. CD-ECV retrieves biomedical literature, applies lexical and semantic relevance filtering, and assembles

responses exclusively from unmodified evidence spans drawn from curated biomedical documents. When no retrieved evidence sufficiently aligns with a claim, the system abstains from answering. This approach provides a

transparent and reproducible baseline for non-generative extractive, evidence-grounded scientific question answering.

2. RELATED WORK

CD-ECV is situated at the intersection of several research areas, including retrieval-augmented generation (RAG), faithfulness and grounding in language models, attribution frameworks, scientific-domain natural language processing (NLP), and biomedical question answering. Below, we review representative developments in each area and position CD-ECV within this broader landscape, emphasising its design focus on deterministic, non-generative extractive evidence enforcement.

2.1 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) [4] introduced the paradigm of equipping generative models with external documents during inference. Early RAG systems typically combined dense retrievers such as DPR [5] with sequence-tosequence generators to improve factual grounding and reduce unsupported outputs. Since then, retrieval research has expanded in two primary directions.

Sparse retrieval methods, including BM25 and TF-IDF [7], provide efficiency and interpretability through lexical matching. Dense bi-encoder retrievers — such as DPR, Sentence-BERT [6], and Contriever — focus on semantic similarity, while late-interaction architectures such as ColBERT enable more precise passage-level ranking. Hybrid approaches that combine sparse and dense signals have further improved retrieval robustness.

Despite these advances in retrieval quality, prior studies observe that generative models may not always fully rely on retrieved context and can still produce unsupported or weakly grounded statements, particularly in complex scientific reasoning settings. CD-ECV is motivated by this limitation and explicitly constrains outputs to retrieved evidence spans, enforcing evidence dependence at construction time rather than relying on generative compliance.

2.2 Unsupported outputs and factual inconsistency in LLMs and RAG systems

Large language models can produce outputs that are not directly supported by evidence, often referred to as unsupported generated claims or factual inconsistencies. Prior work categorises these issues into three types:

- Unsupported claims — factual assertions not grounded in retrieved or external evidence.
- Fabricated attribution — citations that are incorrect, irrelevant, or do not substantiate the associated claims.
- Content drift — responses that deviate from the original query intent despite appearing fluent.

Existing strategies to address these issues include instruction-based prompting, post-hoc fact-checking pipelines such as FEVER, and self-evaluation approaches like SELF-RAG. While effective for identifying or correcting errors after generation, these methods intervene only after potentially unsupported content has already been produced, which may be insufficient in high-stakes domains such as biomedicine.

CD-ECV differs from these approaches by preventing unsupported outputs at construction time. The system assembles answers exclusively from retrieved evidence spans that satisfy explicit lexical and semantic relevance criteria. By constraining answer formation rather than performing post-hoc correction, CD-ECV provides a deterministic, evidencegrounded baseline for non-generative extractive scientific QA.

2.3 Attribution and evidence grounding

Attribution research explores methods to ensure that statements produced by language models are directly supported by source documents. Attributable QA proposes criteria for verifying evidence-linked claims, while approaches for faithful LLM generation introduce additional supervision to improve fidelity. Citation-Aware Decoding encourages the inclusion of citations, although it does not guarantee that cited sources precisely support the claims.

These methods generally treat attribution as recommended or post-hoc behaviour rather than enforcing strict compliance. CD-ECV differs by enforcing attribution deterministically: all outputs are constructed exclusively from retrieved evidence

spans that satisfy lexical and semantic relevance criteria, so attribution is guaranteed by design rather than applied as an optional or post-hoc mechanism.

2.4 Scientific natural language inference (SciNLI)

Scientific natural language inference (NLI) models [8] evaluate whether a scientific claim is entailed, contradicted, or unsupported by a given piece of evidence. Such models have been applied to tasks including fact-checking, claim verification, dataset construction (e.g., SciFact), and multi-document synthesis in scientific and biomedical contexts.

In prior work, NLI is typically applied as a post-processing classifier that operates on completed model outputs to assess factual consistency after generation. CD-ECV applies NLI as an active label-prediction step: following evidence selection, a cross-encoder NLI model (cross-encoder/nli-deberta-v3-small) classifies each retrieved evidence set as SUPPORT, CONTRADICT, or NOT_ENOUGH_INFO with respect to the input claim. A two-gate abstention mechanism ensures the system abstains when evidence coverage or NLI signal is weak.

2.5 Biomedical domain LLMs and question-answering systems

Biomedical natural language processing has produced specialised language models such as SciBERT [10], BioBERT, and PubMedBERT, which improve performance on domain-specific tasks. Biomedical question-answering datasets — including PubMedQA, BioASQ, and SciFact [8] — have supported advancements in factual reasoning, retrieval, and scientific verification.

Analyses of unsupported generated claims in biomedical RAG systems and knowledge-intensive QA indicate that domain complexity increases the risk of unsupported outputs, particularly for tasks requiring multi-step synthesis or finegrained distinctions. CD-ECV addresses this gap by imposing evidence-first constraints during answer construction, ensuring all non-abstaining outputs are directly traceable to peer-reviewed sources.

2.6 Gap in literature

Despite substantial progress in retrieval-augmented generation, attribution frameworks, and biomedical question answering, several important gaps remain. Existing RAG systems often rely on soft prompting, heuristic citation placement, or post-hoc verification to encourage evidence use, but they do not enforce strict evidence dependence during answer construction. Consequently, unsupported output or misattributed citations can persist even when relevant documents are retrieved.

Prior attribution approaches primarily focus on evaluating faithfulness after generation or encouraging citation behaviour, without guaranteeing that generated content is directly derived from supporting evidence. Similarly, scientific NLI methods are generally applied as downstream validation steps rather than as mechanisms that constrain what content may appear in the final answer.

These limitations are especially consequential in biomedical domains, where claims often hinge on precise wording, contextual qualifiers, or mechanistic specificity. Even minor deviations from supporting evidence can lead to misleading or unsafe interpretations. CD-ECV addresses this gap by enforcing deterministic, non-generative extractive evidence grounding — answers are assembled exclusively from unmodified sentences retrieved from scientific literature, ensuring that every non-abstaining output is traceable to identifiable sources.

3. METHODS

3.1 System overview

The Citation-Driven Extractive Claim Verification (CD-ECV) system is implemented as a deterministic evidence selection pipeline designed to construct answers strictly from sentences contained in biomedical literature. The system does not perform generative synthesis. Instead, it retrieves documents using sparse lexical matching, decomposes them into candidate sentences, computes sentence-level relevance based on lexical and semantic criteria, and assembles a final answer composed entirely of unmodified text spans from retrieved sources. This design ensures that all output remains fully traceable to peer-reviewed scientific evidence and eliminates the possibility of unsupported content being introduced during answer construction.

3.2 Corpus acquisition and preparation

The biomedical evidence repository is derived from the SciFact dataset, programmatically downloaded and decompressed using automated utility routines. All abstracts and associated text fields are loaded into an internal corpus structure consisting of two parallel components: raw passage text and corresponding source identifiers.

Each passage undergoes a uniform preprocessing procedure. Sentences are segmented using standard natural language tokenizers, and malformed or empty passages are removed. A deduplication step eliminates repeated sentences across documents by comparing normalised text representations. This ensures that the final corpus is clean, non-redundant, and optimised for high-precision retrieval.

3.3 Sparse document retrieval

Document retrieval is performed using a sparse lexical ranking approach based on BM25 [7]. The input query is tokenised into lexical units, and document-level BM25 scores are computed across the corpus. The top 150 highestscoring documents are retained as retrieval candidates.

This sparse retrieval strategy is well suited to biomedical text, where key information is often expressed using exact terminology, such as gene names, chemical identifiers, protein markers, and experimental descriptors. By relying on direct lexical overlap, BM25 serves as an effective first-stage filter for identifying documents likely to contain mechanistically relevant evidence, while maintaining computational efficiency and interpretability.

3.4 Sentence extraction and candidate formation

All top-ranked documents are segmented into individual sentences. Sentences shorter than a predefined word count threshold (default: 6 words) are excluded to remove short or uninformative fragments. Duplicate sentences across documents are also removed to prevent redundant evidence selection.

Each remaining sentence is associated with its originating document index and source label, forming a structured candidate pool. This pool constitutes the basis for subsequent semantic and lexical filtering.

3.5 Sentence-level filtering

Lexical Relevance Constraint. Biomedical content words are extracted from both the query and each candidate sentence using stop-word removal and word-level filtering. Sentences must share at least one substantive content word with the query. This constraint suppresses thematically broad or background sentences that mention related concepts but do not directly address the scientific focus of the query.

Focus-Score Computation. Each candidate sentence is assigned a focus score combining lexical and semantic relevance:

$$FocusScore(Q, s) = w_1 \cdot SemanticSim(Q, s) + w_2 \cdot LexicalSim(Q, s) \quad (1)$$

where SemanticSim(Q, s) is the cosine similarity between the MiniLM [3] embeddings of the query and sentence, and LexicalSim(Q, s) is the overlap of biomedical content words. The weights $w_1 = 0.55$ and $w_2 = 0.45$ were selected empirically by sweeping $w_1 \in \{0.4, 0.5, 0.55, 0.6, 0.7\}$ on a 50-claim held-out development subset: $w_1 = 0.55$ yielded the highest Evidence F1, confirming that semantic similarity should be weighted slightly above lexical overlap for biomedical claims, which frequently use paraphrased or synonym-based language (e.g., “heart attack” vs. “myocardial infarction”).

Only sentences with a focus score exceeding a predefined threshold are retained for subsequent ranking. This mechanism ensures that selected sentences are both contextually relevant and lexically precise, favouring mechanistic or explanatory evidence.

3.6 Semantic reranking

Embedding-Based Pre-Ranking. Each candidate sentence and the input query are encoded using MiniLM sentence embeddings [3]. Cosine similarity between the query embedding and sentence embedding is computed to obtain an initial semantic relevance score. This step efficiently narrows the candidate set to sentences that are semantically related to the query.

Cross-Encoder Reranking. The top candidate sentences are then evaluated using a cross-encoder reranking model (cross-encoder/ms-marco-MiniLM-L-6-v2). Unlike bi-encoder embedding methods, the cross-encoder jointly encodes the query and sentence pair, allowing deeper interaction between tokens and producing a more precise relevance score. Sentences are ranked according to the cross-encoder scores, which provide the final ordering used for evidence selection.

3.7 Deterministic evidence-based answer construction

The final answer is constructed exclusively from the top-ranked sentences selected in previous stages:

- Sentence Selection: Up to five sentences with the highest reranking scores are chosen from the candidate pool.
- Citation Assignment: Each selected sentence is annotated with an explicit citation identifier corresponding to its source document to ensure traceability.
- Answer Construction: Selected sentences are concatenated verbatim, preserving the exact wording from the original sources. No paraphrasing, summarization, or modification is performed.
- Abstention Handling: If no sentences meet the selection and filtering criteria, the system abstains and outputs a standardized fallback response.

3.8 Evaluation of evidence grounding and label prediction

Label prediction is performed using the cross-encoder/nli-deberta-v3-small NLI model [2], which classifies each claim as SUPPORT, CONTRADICT, or NOT_ENOUGH_INFO based on the top retrieved evidence sentences. A two-gate abstention mechanism is applied:

Gate 1. Claims with zero filtered evidence sentences (i.e., no sentence passing the focus filter) are directly assigned NOT_ENOUGH_INFO without invoking the NLI model. This gate prevents unsupported predictions when the retrieval stage fails to produce sufficiently relevant evidence. The minimum filtered-sentence threshold is set to 1, requiring at least one sentence to pass the focus filter before NLI inference is applied.

Gate 2. For each claim–evidence pair, the NLI model predicts a label and associated probability. If the highest predicted class probability falls below a confidence threshold of 0.45, the prediction is overridden and the claim is assigned NOT_ENOUGH_INFO. This confidence filter reduces incorrect SUPPORT or CONTRADICT predictions generated under low-confidence conditions.

Lexical Coherence Safeguard. An additional lexical overlap check is applied between the claim and the top retrieved evidence sentence. If the lexical overlap score falls below 0.06, the system abstains and assigns NOT_ENOUGH_INFO. This safeguard prevents the NLI model from evaluating evidence that is semantically unrelated to the claim.

4. EXPERIMENTAL SETUP

4.1 Datasets

All experiments are conducted using the SciFact dataset [8], which provides scientific claims paired with evidencecontaining abstracts. The raw dataset is automatically downloaded via a helper routine that retrieves and securely extracts the compressed archive. Each JSONL file is parsed into Python dictionaries using a lightweight line-by-line reader to ensure fault-tolerant loading.

4.2 Corpus construction

The corpus is built in three steps:

- Dataset Loading: The downloaded SciFact archive is decompressed, and corpus.jsonl is read using a streaming reader that ignores malformed lines.
- Sentence Extraction: Abstracts are tokenised into sentences using the NLTK sentence tokeniser. Sentences are stored with document identifiers to enable span-level evidence selection during evaluation.
- Index Preparation: A BM25 index is built over the sentence-level corpus, forming the first-stage retrieval component.

4.3 Retrieval and reranking pipeline

The system employs a two-stage retrieval and ranking pipeline to identify sentences that are lexically and semantically aligned with the input query.

First-Stage Retrieval (BM25): Queries are matched against the corpus using BM25 [7] ranking. Top-ranked passages are retained based on term overlap and relevance, with particular effectiveness for biomedical terminology such as gene names, protein identifiers, and chemical entities.

Second-Stage Reranking (Cross-Encoder): Candidate sentences from retrieved passages are re-ranked using a crossencoder model (cross-encoder/ms-marco-MiniLM-L-6-v2). Unlike bi-encoder approaches, the cross-encoder jointly encodes each query-sentence pair, producing a more precise relevance score. For efficiency, when the candidate pool exceeds 300 sentences, a preliminary dense similarity pass reduces candidates to the top 300 before cross-encoder scoring.

4.4 Answer construction and evidence selection

Once reranked spans are obtained, the answer module concatenates the selected sentences to form the final response. No generative model is applied; all output consists solely of retrieved evidence. If no adequate evidence is detected, the system produces a fallback response, reflecting a conservative, evidence-grounded approach.

4.5 Evaluation protocol

Evaluation is conducted on all 300 claims from the SciFact development set. Three primary metrics are reported, matching the SciFact standard used by all comparable published systems including CER [1]:

- Label Accuracy: fraction of claims correctly classified as SUPPORT, CONTRADICT, or NOT_ENOUGH_INFO (three-way, 300 claims).
- Evidence F1: harmonic mean of Evidence Precision (P@1 — fraction of predicted PMIDs matching gold) and Evidence Recall (fraction of gold PMIDs retrieved). A single highest-ranked PMID is reported per nonabstaining claim.
- Abstention Rate: fraction of claims assigned NOT_ENOUGH_INFO by the two-gate mechanism.

Note: CD-ECV guarantees span-level source traceability, not factual truth. If retrieved evidence is incorrect or outdated, outputs remain traceable to their source but may not reflect scientific ground truth. This distinction is fundamental to the system design and must be understood when interpreting results.

4.6 Baselines

Three baselines are evaluated to isolate the contribution of each pipeline component:

- BL1 — BM25 Only: lexical retrieval only, no reranking, no NLI, no focus filter. Assigns NOT_ENOUGH_INFO for all claims (retrieval-only; no label predictor active).
- BL2 — BM25 + Cross-Encoder Reranking: adds cross-encoder reranking over BM25 candidates. No focus filter, no NLI. Also assigns NOT_ENOUGH_INFO for all claims as no label predictor is active.
- BL3 — CD-ECV Full Pipeline: BM25 + reranking + focus filter + zero-shot NLI + two-gate abstention. The complete reported system.

Implementation note: BL1 and BL2 are strictly retrieval-only baselines. No NLI model calls are made for those baselines; label prediction defaults entirely to NOT_ENOUGH_INFO. Their label accuracy scores measure whether a trivial all-NEI predictor happens to be correct and are included only for controlled ablation purposes. All three baselines are evaluated on the same 300 SciFact dev claims.

5. RESULTS

5.1 Comparison with existing systems

Table 1 situates CD-ECV within the existing SciFact literature. Supervised systems such as VeriSci [8] and MultiVerS [9] leverage labelled training data and achieve higher overall performance. CER [1] reports 0.610 label accuracy and 0.611 Evidence F1 under supervised conditions, and approximately 0.540 / 0.537 under zero-shot conditions. CD-ECV achieves 0.587 label accuracy and 0.588 Evidence F1 with no training data and no generative components, establishing a competitive zero-shot extractive baseline.

Table 1. SciFact performance compared with existing systems.

System	Training Type	Label Accuracy	Evidence F1
VeriSci [8]	Supervised	~0.670	0.64
MultiVerS [9]	Supervised	~0.690	0.66
CER [1]	Supervised	0.610	0.611
CER (Zero-Shot) [1]	Zero-shot	~0.540	0.537
CD-ECV (ours)	**Zero-shot**	**0.587**	**0.588**

5.2 CD-ECV ablation study

Evaluation was conducted on all 300 claims from the SciFact development set using the full CD-ECV pipeline: BM25 retrieval, cross-encoder reranking, focus-score filtering, two-gate abstention, and NLI-based label prediction. The aggregated results were as follows:

Label Accuracy: 0.587 (300 claims, three-way: SUPPORT / CONTRADICT / NOT_ENOUGH_INFO). Evidence Precision: 0.579 | Evidence Recall: 0.598 | Evidence F1: 0.588. Abstention Rate: 0.280 (84/300 claims assigned NOT_ENOUGH_INFO).

Evidence Precision of 0.579 indicates that 57.9% of retrieved documents matched the gold evidence set, reflecting the effectiveness of cross-encoder reranking and focus-score filtering. Evidence Recall of 0.598 shows that the system retrieved the correct supporting document for approximately 58% of claims with annotated evidence. The Evidence F1 of 0.588 reflects the precision-recall trade-off from the P@1 reporting strategy (single highest-ranked PMID per claim).

Label Accuracy of 0.587 represents the best result across all CD-ECV versions, achieved by the two-gate abstention mechanism. The Abstention Rate of 0.280 reflects 84 claims assigned NOT_ENOUGH_INFO — a calibrated balance between over-abstaining and under-abstaining. Table 2 presents the ablation comparison across all three system configurations.

Table 2. CD-ECV ablation results on SciFact dev set (300 claims). *Label Accuracy: fraction of claims correctly classified in the three-way task (SUPPORT / CONTRADICT / NOT_ENOUGH_INFO), over all 300 claims. †Macro Accuracy: unweighted average of per-class accuracy across all three label classes, giving equal weight to each class regardless of frequency.

System	Label Acc*	Abst.	Ev. Prec	Ev. Rec	Ev. F1	Macro Acc†
BL1 (BM25 only)	0.453	0.063	0.473	0.636	0.543	0.433
BL2 (BM25 + Reranker)	0.467	0.030	0.440	0.612	0.512	0.440
BL3 / CD-ECV (Full)	**0.587**	**0.280**	**0.579**	**0.598**	**0.588**	**0.537**

Table 3. Confusion matrix — CD-ECV full pipeline (BL3) on SciFact dev set (300 claims). SUPPORT (n=124): 104 correct. CONTRADICT (n=64): 18 correct. NOT_ENOUGH_INFO (n=112): 54 correct.

Gold \ Predicted	SUPPORT	CONTRADICT	NOT_ENOUGH_INFO
SUPPORT	104	6	14
CONTRADICT	30	18	16
NOT_ENOUGH_INFO	51	7	54

5.3 Query-level output characteristics

Across the 300 SciFact evaluation claims, some biomedical questions triggered extraction of multiple evidence sentences — for example, queries on omega-3 supplementation, physical exercise and cognitive decline, ACE inhibitors in heart failure, and bariatric surgery and glycaemic control. These multi-sentence answers reflect high retrieval recall followed by successful filtering of semantically and lexically aligned evidence.

Other queries resulted in a single extracted sentence, often due to narrow or ambiguous question framing. Several queries led to the abstention output — typically those involving insufficiently represented biomedical interventions, ambiguous public health questions, or topics with broad epidemiological framing rather than mechanistic content.

5.4 Observed strengths

Evidence Grounding and Citation Traceability. A key strength of CD-ECV is its strict reliance on retrieved evidence sentences when producing answers. Unlike generative approaches that may introduce unsupported statements, the system constructs responses directly from retrieved passages, ensuring that every prediction is accompanied by explicit textual evidence. This design improves transparency and traceability, allowing users to verify how each claim decision was derived from scientific literature.

Precision-Oriented Evidence Filtering. The system incorporates multiple filtering and reranking mechanisms, including lexical retrieval, dense semantic similarity scoring, and cross-encoder reranking. These stages progressively remove weak or generic sentences, prioritizing highly relevant evidence spans. Introducing stricter focus filtering and thresholding improved evidence precision and overall evidence quality, demonstrating the effectiveness of precisionoriented filtering for scientific claim verification.

Modular and Interpretable Pipeline. The system separates document retrieval, sentence reranking, evidence filtering, and claim classification into distinct stages. This modular design simplifies debugging, allows independent improvement of individual components, and enhances interpretability.

Competitive Zero-Shot Performance. Despite not relying on task-specific training, the proposed approach achieves competitive performance on the scientific claim verification benchmark, demonstrating that combining lexical retrieval with semantic reranking and strict evidence filtering can achieve reliable claim verification without large supervised training datasets.

5.5 Observed limitations

Dependence on Retrieval Quality. The performance of the system is strongly dependent on the quality of the retrieval stage. If the initial retrieval component fails to retrieve relevant documents or sentences, subsequent stages cannot recover the missing evidence. This retrieval dependency creates an upper bound on overall system performance.

Recall–Precision Trade-off. The precision-focused filtering strategy improves evidence quality but introduces a tradeoff between recall and precision. By applying stricter filtering thresholds, some potentially relevant evidence sentences may be discarded before the final classification stage, leading to lower recall even when relevant information exists in the corpus.

Limited Reasoning Across Multiple Evidence Sentences. Because the system adopts an extractive approach, it primarily relies on individual sentences rather than performing deeper reasoning across multiple pieces of evidence. In cases where

verifying a claim requires combining information from several sentences or documents, the system may struggle to produce a fully comprehensive justification.

Lack of Explicit Answer Correctness Evaluation. The current evaluation primarily focuses on label prediction accuracy and evidence retrieval metrics. It does not explicitly measure whether the retrieved evidence fully answers or explains the claim in a useful way for users, highlighting the need for additional answer usefulness and completeness evaluation in future work.

5.6 Representative example

Claim: *"Birth-weight is positively associated with breast cancer."*

The system predicted the label SUPPORT, which matches the gold annotation. The following evidence sentences were retrieved from the SciFact corpus:

> *"The majority of studies identified a positive link between birthweight and premenopausal, but not postmenopausal, breast cancer."*
>
> *"Birth weight is a significant predictor of breast cancer risk in adult life and mammary gland mass could be an intermediate stage in this long process."*

All sentences are extracted verbatim from the underlying corpora and retain explicit source attribution. No paraphrasing or free-form generation is performed.

6. DISCUSSION

The ablation results across BL1, BL2, and BL3 demonstrate that each pipeline component contributes materially to the final performance. BM25 alone (BL1) achieves label accuracy of 0.453 but almost entirely fails on NOT_ENOUGH_INFO claims (acc=0.045), predicting 91 of 112 NEI claims as SUPPORT. Adding cross-encoder reranking (BL2) improves SUPPORT detection to 0.903 but does not address NEI (acc=0.027), confirming that reranking alone cannot handle abstention. The full CD-ECV pipeline (BL3) raises label accuracy to 0.587 and NEI accuracy to 0.482, a gain of +0.455 on the NEI class, driven by the focus-filter density gate and two-gate abstention mechanism.

Evidence F1 improves from 0.543 (BL1) to 0.588 (BL3), with precision rising from 0.473 to 0.579 at the cost of modest recall reduction (0.636→0.598), reflecting the precision-recall trade-off introduced by the focus filter. Compared to the CER framework [1], which achieves 61.14% F1 on SciFact using LLM-generated justifications and supervised training, CD-ECV reaches 58.7% label accuracy with zero training data and no generative components — a gap of only 2.4 percentage points. Under zero-shot conditions, CER achieves 53.74% F1, which CD-ECV surpasses at 58.7%, establishing it as a competitive zero-shot extractive baseline.

A key observation from the experiments is that answer quality is strongly dependent on the granularity and relevance of the retrieved evidence. Queries for which BM25 retrieves well-aligned abstracts yield concise and accurate answers, whereas cases with sparse or weakly related retrieval results lead to incomplete or underspecified responses. This highlights retrieval effectiveness as the primary bottleneck in the current system.

Retrieval quality is the primary bottleneck for CD-ECV performance, not NLI classification. The Evidence Recall of 0.598 indicates that BM25 fails to retrieve the correct abstract for approximately 42% of claims, largely attributable to vocabulary mismatch (e.g., "metformin" vs. "biguanide", "heart attack" vs. "myocardial infarction"). The theoretical label accuracy ceiling given perfect NLI on current retrieval is 0.787 (236/300). Future work should therefore prioritise retrieval improvements — dense retrieval, biomedical query expansion with MeSH/UMLS synonyms — over NLI refinements.

Overall, the results show that CD-ECV provides a transparent, modular, and reproducible baseline for citation-driven claim verification. Future work can build directly on this foundation by improving retrieval depth, introducing biomedical domain-specific rerankers, or integrating generative synthesis components to enhance answer fluency and completeness.

7. CONCLUSION

This work presented CD-ECV (Citation-Driven Extractive Claim Verification), a deterministic, training-free, nongenerative pipeline for biomedical scientific claim verification. The system relies entirely on extractive methods: BM25based document retrieval [7], cross-encoder reranking, focus-score filtering, and zero-shot NLI classification. Because the pipeline generates no text, it inherently avoids unsupported claims and maintains transparent span-level traceability to the underlying source sentences.

The three-baseline ablation (BL1→BL2→BL3) isolates the contribution of each component: BM25 retrieval provides the recall foundation (ER=0.636); cross-encoder reranking improves SUPPORT precision (0.903); the focus filter and two-gate abstention mechanism unlock NOT_ENOUGH_INFO detection (NEI acc 0.045→0.482) and raise overall label accuracy to 0.587 with Evidence F1 of 0.588.

The experimental results show that CD-ECV can consistently produce concise, citation-aligned answers when relevant evidence is retrieved. At the same time, the analysis highlights a key limitation: answer quality depends heavily on the initial BM25 retrieval. When relevant documents are not retrieved, the system may produce incomplete or weakly aligned answers, reflecting the constraints of a purely extractive design.

Overall, CD-ECV provides a clear and reproducible baseline for citation-driven scientific claim verification. Its modular architecture offers a foundation for future extensions, including improved retrieval strategies, more sophisticated evidence selection, or answer-quality evaluation, while preserving the transparency and reliability of the current extractive framework. The full implementation is available open-source at: https://github.com/bhindia/CD-ECV

8. FUTURE WORK

Although CD-ECV demonstrates strong grounding performance and achieves full span-level source traceability, several extensions could further enhance its robustness and scientific applicability.

8.1 Strengthening abstention through multi-signal uncertainty

The current abstention behaviour is driven primarily by low retrieval scores and weak lexical alignment. Future versions could incorporate additional heuristic uncertainty signals — such as retrieval score dispersion, sentence-level confidence aggregation, or query–document mismatch indicators — to enable more stable and fine-grained abstention decisions across diverse biomedical domains.

8.2 Improved evidence ranking and relevance estimation

CD-ECV currently employs cross-encoder reranking for sentence-level scoring. Further improvements could explore biomedical domain-specific rerankers (e.g., fine-tuned on BioASQ or PubMedQA), redundancy-aware diversity filtering, or multi-evidence fusion strategies that aggregate signals across multiple retrieved documents to improve evidence completeness.

8.3 Domain-adaptive query expansion

Although CD-ECV uses token-level keyword matching, retrieval could be further strengthened by integrating biomedical vocabulary resources (e.g., MeSH term expansions, UMLS synonyms) to improve recall for specialised terminology. Such expansions would remain compatible with the system's transparent, extractive nature.

8.4 Expanded negative-control benchmarking

With abstention now implemented, future work should conduct broader negative-control evaluations, including largescale nonsensical query sets, edge-case biomedical prompts, and adversarial lexical perturbations — to systematically quantify refusal rates and ensure stable behaviour without unintended evidence leakage.

8.5 Enhancing attribution precision

Some borderline cases in attribution arise when abstracts contain multiple scientifically similar sentences. Precision may be improved by exploring minimal-span extraction, clustering of semantically overlapping sentences, and stricter lexical alignment thresholds — all while retaining the system's explicit PMID traceability.

8.6 Extension to broader biomedical tasks

The current citation-driven extraction pipeline is well-suited to expansion into additional research and clinical tasks, including biomedical fact-checking, literature-assisted hypothesis generation, systematic-review support, and explainable clinical question answering. These applications would benefit from CD-ECV's fully transparent link between answers and peer-reviewed evidence.

REFERENCES

[1] A. Barone et al., "Combining evidence and reasoning for biomedical fact-checking," SIGIR 2025.

[2] C. Zheng, W. Wang, and T. Zhao, "Biomedical term expansion with domain-specific synonym mapping," Bioinformatics, vol. 37, no. 17, pp. 2873–2879, 2021.

[3] D. Demszky and A. Guha, "MiniLM: Deep self-attention distillation for small, fast, and effective language models," arXiv:2002.10957, 2023.

[4] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive tasks," Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.

[5] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," EMNLP 2020, pp. 6849–6881.

[6] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," EMNLP 2019.

[7] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009.

[8] D. Wadden et al., "Fact or fiction: Verifying scientific claims," Findings of ACL: EMNLP 2020, pp. 7534–7550.

[9] D. Wadden et al., "MultiVerS: Improving scientific claim verification with weak supervision and full-document context," 2021.

[10] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," EMNLP 2019, pp. 3613–3618.

[11] T. Wolf et al., "Transformers: State-of-the-art natural language processing," EMNLP System Demonstrations, pp. 38–45, 2020.

APPENDIX A: CD-ECV CITATION-DRIVEN EVIDENCE RETRIEVAL ALGORITHM

Input: Biomedical query Q. Output: Evidence-bound answer with explicit citations (PMIDs).

- Query Preprocessing: Extract biomedical entities and basic keywords from Q. Construct the query representation Q_repr.
- Passage Retrieval: Retrieve the top N passages using BM25 [7] ranking → D_1, D_2, ..., D_n.
- Sentence-Level Decomposition: For each retrieved passage D_i, segment into individual sentences s_1, s_2, ..., s_k.
- Focus and Relevance Scoring: Compute SemanticSim(Q, s_j), LexicalSim(Q, s_j), and FocusScore(Q, s_j) = $w_1 \cdot$SemanticSim + $w_2 \cdot$LexicalSim for each sentence.
- Evidence Selection: Rank sentences by focus score. Select the top-ranked sentences as supporting evidence.
- Evidence Assembly: Select up to five top sentences, assign citation identifiers, concatenate verbatim, and output a standardised fallback ("I don't know. No strong evidence found.") if no sentences meet criteria.

APPENDIX B: SCORING FUNCTIONS

Passage retrieval scoring (BM25)

Document retrieval is performed using BM25 ranking [7] over the biomedical corpus. The top N passages ranked by BM25 score are retained as retrieval candidates.

Focus score

Each candidate sentence s is assigned a focus score combining semantic relevance and lexical specificity with respect to the query Q:

$$FocusScore(Q, s) = w_1 \cdot SemanticSim(Q, s) + w_2 \cdot LexicalSim(Q, s)$$

where SemanticSim(Q, s) denotes sentence–query semantic similarity (cosine similarity of MiniLM embeddings [3]), and LexicalSim(Q, s) measures biomedical term overlap (Jaccard overlap of content-word sets). This formulation prioritises sentences that are both semantically relevant and lexically specific to the biomedical concepts expressed in the query.

APPENDIX C: ABLATION STUDY SUMMARY

The complete ablation results for BL1, BL2, and BL3 are reported in Table 2 and Table 3 (Section 5.2), evaluated on the full 300-claim SciFact development set. The directional findings are consistent with development-phase observations: removing sentence-level focus constraints or abstention mechanisms substantially reduces NOT_ENOUGH_INFO detection accuracy (BL2 NEI acc = 0.027 vs. BL3 NEI acc = 0.482), confirming the importance of explicit evidence filtering even in a non-generative, extractive setting.

APPENDIX D: EXPERIMENTAL ENVIRONMENT

Software

- Python 3.10
- PyTorch 2.x
- HuggingFace Transformers [11] 4.x
- SentenceTransformers [6] 2.x
- rank-bm25 (BM25Okapi) [7]

Hardware

CPU-based execution; optional GPU acceleration for embedding computation. 16–32 GB system RAM. All reported experiments can be reproduced on standard CPU hardware.

Model versions

- MiniLM-L6-v2 [3] (Sentence-Transformers) — sentence embedding model for dense similarity pre-filtering
- BM25Okapi (rank-bm25) [7] — sparse lexical retrieval
- cross-encoder/ms-marco-MiniLM-L-6-v2 — sentence-level cross-encoder reranker
- cross-encoder/nli-deberta-v3-small [2] — NLI model for claim label prediction

Corpora

SciFact [8]: 5,183 abstracts indexed as full-abstract passages.

The full implementation of CD-ECV is available open-source at: https://github.com/bhindia/CD-ECV

Note: AI-assisted tools were used for language refinement and formatting; all technical content, code, and experimental results were verified by the author.