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.

Hallucination Detection Using Evidence Retrieval and Natural Language Inference

Krish
Indian Institute of Technology Roorkee
Mayank Goswami
mayankfph@iitr.ac.in

Indian Institute of Technology Roorkee https://orcid.org/0000-0001-8829-9980

Research Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11701365/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.

Hallucination Detection Using Evidence Retrieval and Natural Language Inference

Krish[1] and Mayank Goswami [1, 2, #]
[1]Divyadrishti Imaging Laboratory, Department of Physics, IIT Roorkee
[2]*Mehta Family School of Data Science and Artificial Intelligence*, IIT Roorkee,
Roorkee, Uttarakhand-247667, India.
[#]mayank.goswami@ph.iitr.ac.in

Abstract

Large language models (LLMs) can generate fluent responses containing unsupported, incorrect, or unverifiable factual statements. TruthLens is an evidence-based claim-verification pipeline that examines factual statements individually rather than assigning one label to an entire answer. It extracts sentence-level claims, retrieves candidate evidence using BM25 lexical retrieval and MiniLM semantic retrieval, combines rankings with Reciprocal Rank Fusion (RRF), and applies a pretrained DeBERTa-v3 Natural Language Inference (NLI) model to the highest-ranked evidence sentence. Entailment, contradiction, and neutral predictions are mapped to SUPPORTED, CONTRADICTED, and UNKNOWN. On 19,998 FEVER development examples and a restricted 27,055-sentence evidence corpus, hybrid retrieval achieved Recall@1/5/10 of 56.51%/65.02%/65.67%, outperforming the individual retrieval methods in this experiment. A frozen hybrid top-1 NLI configuration achieved 62.25% accuracy and 0.6231 macro-F1 on a separate 2,000-example held-out subset. An exploratory HaluEval experiment showed poor transfer under the current question-answering evaluation formulation. The results demonstrate a working retrieval-and-verification prototype while exposing limitations in evidence coverage, claim extraction, coreference, uncertainty handling, and cross-dataset generalization.

1. Introduction

Large language models can produce text that is clear and well structured, but this does not necessarily mean that every statement is correct. An answer may contain several factual claims, some of which are supported by evidence while others are incorrect or difficult to verify. Giving the whole answer a single label can hide these differences. Claim-level verification addresses this by separating an answer into individual statements and checking each one against available evidence. Previous surveys describe hallucination as a broad problem across natural language generation and summarise several families of detection and mitigation methods [1] .Other work, such as SelfCheckGPT, explores consistency-based checking without relying on an external evidence collection [2].

This work presents TruthLens, a prototype for checking individual claims against a searchable evidence collection. It first extracts sentence-level claims, then searches for relevant passages using both lexical and semantic retrieval. Their rankings are combined before an NLI model estimates whether the selected evidence supports the claim, contradicts it, or does not provide enough information. It is important to distinguish this task from detecting truth in a general sense. TruthLens only assesses a claim in relation to the evidence it can retrieve from its configured corpus. In particular, an UNKNOWN result means that the available retrieved evidence is insufficient; it does not mean that the claim is false.

2. Problem Statement, Motivation and Objectives

2.1 Problem statement

The task considered in this paper is to take a generated answer and a searchable text corpus, identify the factual claims in the answer, find evidence relevant to each claim, and determine whether that evidence supports the claim, contradicts it, or is insufficient to reach either conclusion.

2.2 Motivation

Lexical retrieval can miss relevant evidence when wording differs. Semantic retrieval can capture paraphrases but may be less sensitive to exact names, dates, and terminology. Combining them may improve candidate retrieval. Retrieval relevance is not factual support: a relevant passage may agree with or contradict a claim. A separate NLI stage is therefore required.

2.3 Objectives

- Build a pipeline that decomposes an answer into individual claims.
- Compare lexical, semantic, and hybrid evidence retrieval.
- Combine BM25 and semantic rankings using RRF.
- Apply NLI to classify the evidence–claim relationship.
- Evaluate retrieval and verification using FEVER-derived data.
- Explore transfer to HaluEval and document limitations.

3. Theoretical Background

3.1 Fact verification and FEVER

FEVER (Fact Extraction and VERification) is a benchmark containing claims labelled as Supported, Refuted, or Not Enough Information, with evidence annotations for supported and refuted claims [3]. TruthLens uses FEVER as a dataset for development and evaluation; FEVER is not itself the TruthLens system.

3.2 BM25 lexical retrieval

BM25 is a probabilistic lexical ranking function. It scores candidate texts using query-term frequency, inverse document frequency, and document-length normalization [4]. It is useful when informative words overlap between a claim and evidence, but may miss paraphrases with limited word overlap.

3.3 MiniLM embeddings and FAISS

A sentence-embedding model maps text to a dense numerical vector. TruthLens uses a pretrained MiniLM sentence-transformer to encode claims and evidence. MiniLM is a compact Transformer model developed using deep self-attention distillation [5]. FAISS indexes and searches dense vectors efficiently [6]. MiniLM creates the embeddings; FAISS searches them.

3.4 Reciprocal Rank Fusion

BM25 and semantic retrieval produce rankings with different score scales. RRF combines rankings using positions rather than raw scores. For evidence item d: $RRF(d) = \Sigma r\ 1 / (k + rankr(d))$, where rankr(d) is the rank in retrieval list r and k is a smoothing constant. TruthLens uses $k = 60$. Up to 50 candidates are retrieved from each method, fused, and reduced to the top 10.

3.5 Natural Language Inference

NLI predicts the relationship between a premise and a hypothesis. In TruthLens, retrieved evidence is the premise and the claim is the hypothesis. The classes are entailment, contradiction, and neutral, mapped respectively to SUPPORTED, CONTRADICTED, and UNKNOWN. The system uses a pretrained DeBERTa-v3 NLI model; DeBERTa introduces disentangled attention and an enhanced mask decoder [7].

4. Related Work

Thorne et al. introduced FEVER as a large-scale benchmark for evidence-based fact verification [3]. Robertson and Zaragoza describe the probabilistic relevance framework underlying BM25 [4]. Wang et al. introduced MiniLM, a compact pretrained Transformer approach based on self-attention distillation [5], while Johnson et al. presented efficient large-scale vector similarity search methods associated with FAISS [6]. Cormack et al. proposed Reciprocal Rank Fusion as a method for combining ranked lists from different retrieval systems [8]. He et al. introduced DeBERTa, which uses disentangled attention [7]. Li et al. presented HaluEval as a benchmark for studying hallucination recognition in generated text [9].

Broader work on hallucination detection provides context for the problem addressed here. Cormack et al. proposed Reciprocal Rank Fusion as a method for combining ranked lists from different retrieval systems [8]. He et al. introduced DeBERTa, which uses disentangled attention [7], and Li et al. presented HaluEval as a benchmark for studying hallucination recognition in generated text [9]. Ji et al. survey hallucination in natural language generation and discuss evaluation and mitigation approaches across several generation tasks [1]. Manakul et al. propose SelfCheckGPT, which uses consistency across sampled model outputs to identify potentially unreliable statements without an external knowledge base [2]. TruthLens takes a different, evidence-retrieval-based approach: it searches a configured corpus and uses NLI to assess the relationship between a claim and retrieved evidence. Its contribution is the implementation and experimental comparison of retrieval and verification choices in this pipeline, rather than a new neural architecture.

5. Methodology

5.1 System overview

An input answer is divided into sentence-level claims. Each claim is independently processed by BM25 and MiniLM-based semantic retrieval. BM25 returns up to 50 candidates; the semantic retriever encodes the claim and evidence using MiniLM and searches a FAISS IndexFlatIP index of normalized vectors for up to 50 candidates. RRF with k = 60 combines the rankings and returns the top 10 fused candidates. The final architecture selected for verification passes the highest-ranked (top-1) evidence sentence to NLI.

5.2 Data preparation and evidence corpus

The FEVER training split used in this project contains 145,449 records: 80,035 SUPPORTS (55.03%), 29,775 REFUTES (20.47%), and 35,639 NOT ENOUGH INFO (24.50%). A full Wikipedia-derived evidence corpus was constructed with 5,416,536 pages and 25,247,890 sentences (approximately 4.55 GB). A smaller development corpus was built from FEVER development references: 2,860 pages and 27,055 sentences; 33 target pages were unavailable.

Both the reported retrieval experiment and the held-out evaluation use the 27,055-sentence development corpus. The full 25.25-million-sentence corpus is a separate corpus artifact; the reported metrics must not be interpreted as results from searching all 25 million sentences.

5.3 Verification and decision policy

The top-1 evidence sentence and its claim are passed to DeBERTa-v3. The model outputs probabilities for contradiction, entailment, and neutral. The pipeline applies a confidence threshold of 0.50 and a margin of 0.10 as part of its decision policy, avoiding a definite label when the prediction is insufficiently decisive. The outputs are SUPPORTED, CONTRADICTED, and UNKNOWN.

5.4 Software implementation

The project is modular. config.py stores paths, model identifiers, device settings, and parameters. dataset_loader.py and preprocess.py load and prepare FEVER records and map labels. build_dev_corpus.py and build_evidence_corpus.py construct the development and full corpora. build_bm25_index.py and build_semantic_index.py build retrieval indexes. hybrid_retriever.py performs retrieval and RRF. nli_verifier.py runs NLI inference, while pipeline.py coordinates claim extraction, retrieval, verification, and answer-level summaries. app/app.py provides a Gradio interface.

6. Experimental Setup

Retrieval evaluation used 19,998 FEVER development examples and the 27,055-sentence development corpus. Retrieval performance is reported as Recall@k. Architecture selection used a 1,000-example development subset. After selecting the architecture, a separate 2,000-example subset was used for held-out evaluation. A sentence-level retrieval analysis used 74 usable examples from a 100-example subset; 26 skipped examples were NOT ENOUGH INFO records with placeholder evidence.

The implementation used Python 3.11, PyTorch, Transformers, Sentence-Transformers, FAISS, rank-bm25, and Gradio. The recorded environment included an NVIDIA GeForce RTX 4060 Laptop GPU with CUDA support.

7. Results

7.1 Retrieval comparison

Retriever	Recall@1	Recall@5	Recall@10
BM25	50.72%	62.69%	64.37%
MiniLM semantic	55.69%	64.22%	65.11%
Hybrid (RRF)	56.51%	65.02%	65.67%

Hybrid retrieval achieved the highest recall at all three cutoffs in this experiment. At Recall@1, it improved by 5.79 percentage points over BM25 and 0.82 points over MiniLM semantic retrieval. The gain over semantic retrieval is modest and should not be generalized beyond this evaluation setting.

7.2 Sentence-level retrieval subset

On the 74 usable examples, BM25 achieved Recall@1/5/10 of 59.46%/78.38%/83.78%; semantic retrieval achieved 58.11%/74.32%/81.08%; and hybrid retrieval achieved 63.51%/82.43%/87.84%. This small subset is exploratory.

7.3 NLI architecture ablation

Configuration	Accuracy	Macro-F1	Weighted F1
BM25 + NLI (V2)	0.4740	0.4482	0.4495
Semantic + NLI (V2)	0.4930	0.4827	0.4838
Hybrid + NLI (V2)	0.5070	0.4960	0.4971
Hybrid + NLI (V1)	0.4740	0.3736	0.3757

BM25 + top-1 NLI	0.5880	0.5887	—
Semantic + top-1 NLI	0.5920	0.5943	—
Hybrid + top-1 NLI	0.6080	0.6100	0.6106
Hybrid + Top-5 NLI + V1	0.4740	0.3736	0.3757
Hybrid + Top-5 NLI + V2	0.5070	0.4960	0.4971

On the 1,000-example architecture-selection subset, hybrid retrieval with top-1 NLI was the strongest tested configuration, with accuracy 0.6080 and macro-F1 0.6100. This configuration was selected for held-out evaluation.

7.4 Held-out evaluation

On the separate 2,000-example held-out subset, the frozen system achieved 62.25% accuracy, 64.40% macro-precision, 62.18% macro-recall, 62.31% macro-F1, and 62.46% weighted F1.

Class	Precision	Recall	F1	Support
SUPPORTED	0.82	0.58	0.68	680
CONTRADICTED	0.63	0.77	0.69	669
UNKNOWN	0.48	0.52	0.50	651

The confusion matrix, with rows as true labels and columns as predicted labels in the order SUPPORTED, CONTRADICTED, UNKNOWN, was [[393, 58, 229], [15, 514, 140], [69, 244, 338]]. UNKNOWN had the weakest F1 (0.50), indicating difficulty separating insufficient evidence from support or contradiction.

7.5 Error analysis

In a manual analysis of 100 examples, 57 were correct, 6 were retrieval misses, 28 were NLI/verification errors, and 9 had unclear error types. Among the 91 examples assigned a meaningful category, retrieval misses represented 6.6% and NLI/verification errors 30.8%. In this inspected sample, verification and decision-making were larger sources of error than retrieval alone.

An initial formulation that did not include the question context achieved accuracy 0.4655, precision 0.4258, recall 0.1980, F1 0.2703, and ROC-AUC 0.5119 on 1,000 examples (2,000 answer comparisons). The revised question-aware formulation achieved accuracy 0.4525, precision 0.3662, recall 0.1300, F1 0.1919, and ROC-AUC 0.5065. Both results are close to chance-level discrimination and indicate that this direct transfer of the FEVER-oriented NLI formulation was unsuccessful in the tested QA setup.

7.6 HaluEval exploratory experiment

An exploratory experiment used 1,000 HaluEval QA examples (2,000 answer comparisons). The question-aware formulation achieved accuracy 0.4525, precision 0.3662, recall 0.1300, F1 0.1919, and ROC-AUC 0.5065. This indicates that the current FEVER-oriented NLI formulation did not transfer effectively to this HaluEval QA setup. It is a limitation of the present formulation and adaptation, not a definitive assessment of every possible HaluEval-based method.

The top-5 NLI aggregation experiments produced accuracy/macro-F1 of 0.4740/0.3736 for V1 and 0.5070/0.4960 for V2. The top-1 configuration achieved 0.6080 accuracy and 0.6100 macro-F1 in the same architecture-selection setting. These results motivated the final top-1 design for this implementation; they do not establish that top-1 evidence is universally preferable for all verification tasks.

A separate pipeline smoke test used three illustrative claims: “Paris is the capital of France” was classified as SUPPORTED (confidence 0.9921); “Paris is the capital of Germany” was classified as CONTRADICTED (confidence 0.9997); and “Paris is exactly 10 metres” was classified as UNKNOWN (confidence 0.9702). A multi-claim answer was also exercised to confirm that the pipeline could return one result per extracted claim. These are functional smoke tests, not substitutes for benchmark evaluation.

8. Discussion

In the FEVER development experiment, combining the lexical and semantic rankings gave the best evidence recall of the three retrieval approaches tested. However, retrieving a relevant passage did not automatically lead to a correct verification decision. The held-out results and error analysis show that deciding how evidence relates to a claim remains challenging, particularly for UNKNOWN cases. This is particularly challenging for claims that need more than one evidence sentence, depend on when an event occurred, or refer to entities indirectly through pronouns or other expressions. A passage can be topically relevant without supporting a claim, which is why TruthLens separates retrieval from NLI.

9. Conclusion

This paper presented TruthLens, a modular claim-level verification prototype combining sentence-level claim extraction, BM25, MiniLM embeddings, FAISS vector search, RRF, and DeBERTa-v3 NLI. On the restricted 27,055-sentence evidence corpus, hybrid retrieval achieved the highest recall among the tested retrieval methods. The selected hybrid top-1 NLI configuration achieved 62.25% accuracy and 0.6231 macro-F1 on a separate 2,000-example held-out FEVER development subset. The results demonstrate the feasibility of integrating hybrid retrieval and NLI, while showing that verification errors and insufficient evidence remain substantial challenges. TruthLens should be regarded as an evidence-grounded assistance system, not an authoritative or universal detector of factual truth.

Additional directions documented for the project include cross-encoder evidence reranking, fine-tuning NLI on factual-verification data, FEVEROUS evaluation, and live web evidence retrieval with source provenance. These are proposed extensions and are not implemented results of the current prototype.

10. Limitations and Future Work

- The main retrieval and held-out results use a restricted 27,055-sentence corpus, not the full 25,247,890-sentence corpus.
- Evaluation is primarily FEVER-style; real LLM answers can contain longer, more complex and context-dependent claims.
- Sentence-level claim extraction may not split compound claims or resolve pronouns and references; a coreference-related failure was observed.
- The final architecture uses only top-1 evidence, although some claims require multiple passages or multi-hop reasoning.
- UNKNOWN is the weakest held-out class (F1 0.50).
- The HaluEval experiment showed poor transfer under the current QA formulation.
- Thresholds and margins are configuration choices and require calibration on representative validation data.
- Future work includes full-corpus evaluation, improved claim decomposition and coreference, multi-evidence NLI, uncertainty calibration, and broader cross-domain evaluation.

11. Additional Implementation Details

11.1 Exact pretrained model identifiers

The semantic encoder is sentence-transformers/all-MiniLM-L6-v2. It generates dense sentence embeddings used by the semantic retrieval component. The NLI verifier uses cross-encoder/nli-deberta-v3-base. These are pretrained models loaded for inference; the project does not claim to train these base models from scratch.

11.2 Retrieval and verification responsibilities

The implementation separates candidate retrieval from semantic verification. BM25 and MiniLM/FAISS produce candidate rankings; hybrid_retriever.py combines those rankings using RRF. The fused retriever exposes a top-10 candidate list, while the selected final verification configuration supplies only the top-ranked evidence sentence to the NLI model. This top-1 choice was selected through development ablation. It reduces the number of claim–evidence pairs evaluated by NLI and avoids diluting a strong evidence signal with irrelevant or neutral candidates under the tested aggregation strategies.

11.3 Claim extraction and answer-level summaries

For claim extraction, the current version uses a simple sentence-level segmentation approach. Each extracted sentence is verified separately. The dashboard also gives a summary of the answer, showing how many claims—and what proportion of the total—fall into each category. The contradicted fraction is specifically the fraction of processed claims labelled CONTRADICTED; it is not a universally validated hallucination rate.

11.4 Dashboard and user-facing outputs

The Gradio dashboard accepts an AI-generated answer and displays extracted claims, final classifications, model confidence, decision margin, NLI class probabilities, retrieved evidence, Wikipedia page information, sentence identifiers, and retrieval scores. The application can be launched from the repository root with `python app/app.py` and is served locally by Gradio.

11.5 Testing and software quality

The repository includes unit and integration tests in tests/test_pipeline.py and tests/test_integration.py. The latest recorded test run completed with 13 passing tests. The integration tests exercise the retrieval-to-NLI-to-decision workflow. The project also includes requirements.txt for dependency installation and requirements-lock.txt for a more reproducible pinned environment.

12. Ethical Considerations

TruthLens may produce incorrect classifications, so its output should not be treated as a final judgement about whether a statement is true. A correct claim could be marked as contradicted, while a false or unsupported claim could be missed. The system may also reflect gaps or biases in the evidence corpus. Its confidence scores need particular care: they are probabilities produced by the NLI model, not objective measures of the chance that a claim is true. Showing the retrieved evidence and retaining an UNKNOWN category are intended to make the system easier to review, but human judgement is still necessary.

13. Digital Resource:

Video Explanation: https://youtu.be/vCD4NtDIpm0

References

[1] Z. Ji *et al.*, “Survey of Hallucination in Natural Language Generation,” *ACM Comput. Surv.*, vol. 55, no. 12, 2023, doi: 10.1145/3571730.

[2] P. Manakul, A. Liusie, and M. Gales, “SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2023, pp. 9004–9017. doi: 10.18653/v1/2023.emnlp-main.557.

[3] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "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*, 2018, pp. 809–819. doi: 10.18653/v1/N18-1074.

[4] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Foundations and Trends in Information Retrieval*, vol. 4, no. 1–2, pp. 1–174, 2009, doi: 10.1561/1500000019.

[5] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers," in *Advances in Neural Information Processing Systems*, 2020.

[6] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," 2017. doi: 10.48550/arXiv.1702.08734.

[7] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," in *International Conference on Learning Representations*, 2021.

[8] G. V Cormack, C. L. A. Clarke, and S. Büttcher, "Reciprocal rank fusion outperforms Condorcet and individual rank learning methods," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009, pp. 758–759. doi: 10.1145/1571941.1572114.

[9] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, "HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 6449–6464. doi: 10.18653/v1/2023.emnlp-main.397.