

TRUST-RAG: RISK-AWARE EVIDENCE RELIABILITY SCORING AND CLAIM-LEVEL VERIFICATION FOR CALIBRATED RETRIEVAL-AUGMENTED LANGUAGE MODELS

Thotakura Rohit Sai Naren¹, Pranav Wadhera²

¹Relanto ,Software Engineer, Email ID: trsnaren2003@gmail.com

²PricewaterhouseCoopers Services LLP (PwC India)2Technology Associate, Email ID: pranavwadhera2002@gmail.com

Abstract: Retrieval-augmented generation (RAG) has emerged as an effective approach for improving the factual grounding of large language models by incorporating externally retrieved evidence into the generation process. However, the presence of retrieved evidence does not necessarily imply that the evidence is sufficiently reliable to support a generated claim. Semantically relevant passages may be incomplete, mutually inconsistent, or unsuitable for substantiating the specific proposition produced by the language model. This creates a critical reliability gap between retrieving evidence and establishing claim-level support. This study proposes TRUST-RAG, a risk-aware retrieval-augmented framework that integrates evidence reliability assessment, evidence re-ranking, post-generation claim verification, and confidence calibration within a unified pipeline. The proposed method represents textual and structured table evidence in a shared embedding space using a lightweight sentence encoder and performs dense retrieval through FAISS. Retrieved evidence is subsequently evaluated using a two-term Evidence Reliability Score that combines retrieval relevance with evidence-type consistency before being supplied to an instruction-tuned language model for evidence-constrained generation. An NLI-based verification stage then examines generated justification claims against the retrieved evidence to identify unsupported statements and derive evidence-grounded confidence estimates. The framework is empirically evaluated on a balanced subset of the FEVEROUS benchmark containing textual and table-based evidence, using approximately 450 claims and additional distractor evidence. Comparisons are conducted against an evidence-free LLM baseline and conventional RAG, together with component-level ablation experiments. Performance is assessed using retrieval precision and recall, classification accuracy, unsupported-claim rate, evidence grounding, and expected calibration error. The study aims to establish a reproducible framework for improving the reliability, verifiability, and calibration of evidence-grounded language-model decisions.

Keywords: Retrieval-Augmented Generation; Evidence Reliability; Claim-Level Verification; Hallucination Detection; Confidence Calibration

1. INTRODUCTION

1.1 Background: Factual Reliability of Large Language Models

Large language models (LLMs) have become effective components of question answering, information retrieval, decision support, and knowledge-intensive applications because of their ability to transform natural-language instructions into coherent responses. However, linguistic fluency does not necessarily imply factual reliability. An LLM may produce a statement that is grammatically convincing while being unsupported by available evidence, partially correct, or inconsistent with another fact in the same response. Such behavior is particularly problematic in



applications where users require traceable and verifiable information rather than plausible text. Recent research has therefore shifted from improving generation quality alone toward mechanisms that retrieve, reconstruct, verify, and assess evidence surrounding generated claims [3], [6], [9].

Retrieval-augmented generation (RAG) provides an important response to this limitation by introducing external evidence into the generation process. Instead of relying exclusively on information encoded in model parameters, a RAG system retrieves potentially relevant information and supplies it to the language model as contextual evidence. Recent studies have investigated evidence-aware retrieval, historical information, structured search, knowledge interaction, and retrieval-generation feedback as mechanisms for improving RAG reliability [1], [2], [12], [16]. Nevertheless, retrieval itself does not establish that the retrieved information is sufficiently reliable to justify the final answer. This distinction motivates a transition from retrieval-oriented RAG toward evidence-risk-aware RAG.

1.2 Limitations of Conventional RAG

The fundamental limitation addressed in this study is that retrieval relevance is not equivalent to evidence reliability. Dense retrieval commonly ranks evidence according to semantic similarity between a query and candidate passages. A highly similar passage, however, may contain only part of the information necessary to establish a claim. It may also contain an opposing statement, lack the required contextual detail, or represent a different form of evidence from that required by the claim.

This problem becomes more pronounced when several retrieved items are jointly presented to an LLM. The model may treat all retrieved passages as equally trustworthy even though their evidential value differs. Context inconsistency has already been identified as an important concern in RAG systems, motivating similarity-validation approaches for detecting conflicting contextual information [20]. Likewise, evidence reconstruction and evidence-aware retrieval research indicates that improving the quality and organization of retrieved evidence can influence factual coherence beyond conventional similarity-based retrieval [1], [10]. TRUST-RAG therefore introduces an Evidence Reliability Score (ERS) that distinguishes retrieval similarity from the consistency of evidence types before the evidence is supplied to the generator.

1.3 Evidence Heterogeneity Problem

A second challenge arises when factual claims depend on heterogeneous evidence. Real information collections are not composed exclusively of continuous prose. Tables, structured records, and textual descriptions can provide complementary information about the same claim. FEVEROUS is particularly appropriate for investigating this issue because its fact-verification setting contains both textual and table-based evidence. Previous research has demonstrated the difficulty of table-based fact verification and the importance of representing structured information effectively [22], [28]. Research on hierarchical representations for tabular RAG further illustrates that table structure can affect retrieval quality [18].

In TRUST-RAG, textual sentences and table evidence are represented within a common retrieval space. Table rows are transformed into explicit representations containing relevant columns, values, and row information. This permits textual and structured evidence to participate in the same retrieval procedure while retaining their evidence-type identity. The resulting design addresses a narrower but experimentally tractable form of heterogeneous evidence grounding: text plus structured table evidence, while image evidence and full multimodal alignment remain outside the present empirical scope.

1.4 Claim-Level Hallucination and Verification

Even correctly retrieved evidence cannot guarantee that the generated explanation will remain faithful to that evidence. An LLM may introduce information absent from the retrieved context, combine independently valid facts into an invalid conclusion, overgeneralize from a specific observation, or produce a statement that contradicts the

evidence. Consequently, evidence retrieval and answer generation should not be regarded as the final stages of factual assurance.

Recent research supports the value of verification-oriented architectures. Sun *et al.* explicitly investigate a retriever-generator-verification configuration for factual coherence, while other studies examine automated hallucination detection and verification mechanisms [3], [6]. Atomic claim decomposition and multi-agent verification have similarly explored the problem of assessing individual factual units rather than treating an entire response as a single indivisible object [4], [5]. Inspired by this direction, TRUST-RAG performs post-generation claim-level verification. Each generated justification sentence is evaluated against retrieved evidence using natural language inference (NLI), producing entailment, contradiction, and neutral assessments. A generated statement is consequently not accepted merely because it was produced from retrieved context; it must demonstrate evidential support.

1.5 Confidence Calibration Problem

Factual reliability also requires appropriate uncertainty communication. A system that predicts SUPPORTS with high confidence when its evidence is weak may be more dangerous than a system that explicitly expresses uncertainty. Accordingly, TRUST-RAG separates prediction from confidence assessment and derives an evidence-grounded confidence value from the strength of entailment between generated claims and retrieved evidence.

The study evaluates this confidence using Expected Calibration Error (ECE), comparing predicted confidence with empirical correctness across confidence bins. This complements conventional classification accuracy by asking whether confidence values correspond to actual predictive reliability. NLI-based approaches have increasingly been investigated for identifying erroneous generated outputs, providing a methodological basis for using entailment-oriented assessment as a component of factual verification [24]–[26].

1.6 Research Gap

The literature reveals three connected gaps. G1: retrieval relevance does not necessarily represent evidence reliability. Existing RAG improvements address retrieval quality, indexing, structured search, and contextual consistency, but semantic similarity alone cannot establish whether retrieved evidence adequately supports a particular claim [18]–[21]. G2: evidence-grounded generation does not guarantee claim-level factual verification. Verification-oriented systems demonstrate the importance of checking generated content, but retrieval, generation, and post-generation verification are not always integrated around an explicit evidence-reliability mechanism [3]–[7]. G3: prediction confidence is not necessarily calibrated confidence. A generated label may be correct without its associated confidence accurately representing the probability of correctness.

TRUST-RAG addresses these gaps through a sequential retrieve → score → rerank → generate → verify → calibrate architecture. Its empirical scope is deliberately restricted to textual and table evidence so that each implemented component can be evaluated using reproducible measurements rather than relying on unvalidated claims.

1.7 Research Questions

The investigation is guided by five research questions:

RQ1: Does evidence-aware retrieval improve retrieval effectiveness compared with conventional dense retrieval?

RQ2: Does Evidence Reliability Scoring improve claim classification compared with standard RAG?

RQ3: Does claim-level NLI verification reduce unsupported generated claims and hallucination?

RQ4: Does evidence-derived confidence improve prediction calibration as measured by ECE?

RQ5: Which TRUST-RAG components contribute most strongly to the observed performance through ablation analysis?

1.8 Research Objectives

The study pursues seven objectives:

To develop an evidence-aware retrieval architecture for textual and structured evidence.

To formulate a two-term ERS using retrieval relevance and evidence-type consistency.

To construct an evidence-weighted RAG generation mechanism.

To implement claim-level NLI verification for generated justifications.

To derive confidence from evidence entailment strength.

To evaluate confidence calibration using ECE.

To quantify component-level contributions through controlled ablation experiments.

1.9 Research Contributions

The principal contributions of TRUST-RAG are fourfold. First, it introduces a lightweight risk-aware Evidence Reliability Score that combines retrieval relevance with evidence-type consistency rather than treating every retrieved item equally. Second, it implements evidence-weighted generation, transforming the conventional retrieval pipeline into a retrieve–score–rerank–weight–generate process. Third, it introduces claim-level post-generation verification, in which generated justification sentences are independently assessed against retrieved evidence using NLI. Fourth, it establishes an evidence-derived confidence mechanism and evaluates whether confidence is calibrated against actual prediction correctness. Together, these components position TRUST-RAG not simply as another RAG architecture, but as an empirical framework for examining the relationship among evidence quality, factual verification, and calibrated trust in retrieval-augmented language models.

2. RELATED WORK

The development of retrieval-augmented language models has shifted factual reliability from a purely generation-oriented problem toward a broader systems problem involving retrieval quality, evidence interpretation, verification, and uncertainty estimation. Conventional RAG architectures establish an important foundation by supplying external information to a language model, but the presence of retrieved information does not by itself guarantee that the information is appropriate, mutually consistent, or sufficient to justify a generated claim. Recent research therefore increasingly examines evidence-aware retrieval, verification-oriented generation, structured fact checking, hallucination detection, and confidence estimation as interconnected components. This section reviews these research directions in relation to the design objectives of TRUST-RAG: Risk-Aware Evidence Reliability Scoring and Claim-Level Verification for Calibrated Retrieval-Augmented Language Models.

2.1 Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) combines an information-retrieval mechanism with a generative language model so that responses can be conditioned on information external to the model's parametric knowledge. This design is particularly useful when the target information is dynamic, domain-specific, or too extensive to be reliably memorized by a language model. Contemporary RAG research has consequently moved beyond simple retrieve-and-generate pipelines toward architectures that incorporate improved indexing, structured search, reasoning, feedback, and context validation. Recent work demonstrates several directions in this evolution. Mo *et al.* investigate the use of historical conversational information to improve retrieval for dialogue-oriented RAG systems, illustrating that retrieval quality can depend substantially on the information used to formulate the retrieval context [2]. Reasoning-guided planning and search have similarly been investigated as mechanisms for making retrieval more closely connected to the reasoning process of the language model [15]. Luo *et al.* combine causal reasoning, heuristic

strategies, fine-tuning, and knowledge interaction to improve RAG behaviour, reflecting the increasing interest in connecting retrieval with more deliberate reasoning mechanisms [12].

Other studies have explored specialized retrieval structures. Flow-RAG, for example, investigates retrieval-augmented question answering over knowledge graphs using gated propagation mechanisms [11]. CATS-RAG examines contextual triplet synthesis for technical question answering, indicating that the usefulness of retrieved context depends not merely on retrieval quantity but also on how contextual information is represented for downstream generation [17]. Enterprise-specific RAG has also received attention, particularly where retrieval must operate over specialized organizational knowledge and operational information [13].

Evaluation studies further demonstrate that RAG performance is sensitive to indexing strategies, distance measures, architecture choices, and application domain. EvaRAG systematically considers indexing and distance metrics when evaluating advanced RAG techniques [19], while comparative evaluation of different RAG architectures highlights that there is no single retrieval-generation configuration that is universally optimal across application settings [21].

These developments establish retrieval as an essential mechanism for improving factual access, but they also expose an important limitation for trustworthy generation: retrieving a relevant passage is not equivalent to establishing that the passage constitutes reliable evidence for a particular claim. TRUST-RAG therefore treats retrieval as the first stage of an evidence assessment process rather than the final determinant of contextual quality.

2.2 Evidence Reliability in Retrieval

A central issue in evidence-grounded generation is the distinction between retrieval relevance and evidence reliability. Dense retrievers generally rank candidate passages according to semantic similarity between a query and the indexed representation. A highly similar passage, however, may provide only partial support, contain information that conflicts with another passage, or fail to contain the information required to establish the claim.

Recent evidence-aware RAG research has begun addressing this limitation. Li *et al.* propose an evidence-aware RAG direction incorporating self-corrective reasoning, demonstrating the importance of considering evidence during the retrieval-generation process rather than treating retrieved passages as unquestioned context [1]. Collini *et al.* investigate similarity validation for handling contextual inconsistencies in RAG, directly highlighting the problem that apparently relevant context can nevertheless introduce inconsistencies into a generated response [20].

Consistency-oriented evidence processing represents another important direction. Shao *et al.* investigate evidence reconstruction driven by consistency considerations, suggesting that reliable RAG requires more than selecting isolated high-scoring documents [10]. Similarly, modular retrieval-generation-feedback architectures emphasize repeated interaction between retrieved information and model output rather than a single irreversible retrieval step [16].

Structured retrieval also complicates reliability assessment. Song examines hierarchical representations for tabular data, demonstrating the importance of preserving structural relationships when retrieving information from tables [18]. This is particularly relevant to fact verification because the meaning of a table value can depend on its associated row, column, header, and neighboring information.

The literature consequently suggests a progression from:

query similarity → retrieval ranking → evidence selection → evidence consistency.

TRUST-RAG extends this progression by introducing a lightweight Evidence Reliability Score (ERS) that combines two experimentally feasible factors:

where r represents retrieval relevance and c represents evidence-type consistency. In the present empirical implementation, this deliberately remains a two-term score rather than the four-factor reliability formulation proposed in the broader framework. Source authority and temporal reliability are therefore outside the validated scope.

The distinction is important. TRUST-RAG does not assume that semantic similarity is itself a reliability measure. Instead, similarity provides one component of an evidence-risk assessment that influences subsequent evidence ranking.

2.3 RAG Hallucination Mitigation

Retrieval can provide factual information to a language model, but it does not guarantee that the model will faithfully use that information. A model may ignore relevant evidence, merge facts from unrelated passages, introduce information absent from the retrieved context, or produce an explanation that is only partially supported. Consequently, hallucination mitigation has become a major research direction in trustworthy RAG.

Sun *et al.* explicitly investigate a retriever-generator-verification architecture for improving factual coherence in open-domain question answering [3]. Their work is particularly relevant to TRUST-RAG because it reinforces the conceptual importance of placing verification after generation rather than assuming that retrieval alone prevents factual errors.

Other studies approach hallucination detection from different perspectives. HaluCheck focuses on explainable and verifiable detection of hallucinations in LLM responses [6], while HaluGNN explores graph neural network-based hallucination detection [8]. Multi-agent collaborative filtering has also been examined as a mechanism for reducing reasoning hallucinations [7]. In domain-specific settings, Lu *et al.* investigate hallucination mitigation for aviation maintenance decision support, demonstrating that factual reliability becomes especially consequential when language-model outputs influence specialized decision processes [9].

These studies reveal an important design distinction. Hallucination detection at the response level does not necessarily establish whether each individual factual statement is supported by a particular evidence item. A response can contain several correct statements and one unsupported assertion, yet a response-level assessment may obscure the localized failure.

TRUST-RAG therefore adopts a claim-level verification strategy. Rather than evaluating only the complete generated answer, the generated justification is segmented into individual sentences. Each sentence is evaluated against the retrieved evidence using natural language inference (NLI), producing entailment, contradiction, and neutral scores. This enables unsupported content to be identified at a finer granularity.

The resulting pipeline is therefore:

retrieve → assess evidence → generate → decompose justification → verify claims.

This sequential structure is central to the proposed framework.

2.4 Claim-Level Fact Verification

Claim verification has a longer research history than RAG and provides methodological foundations for determining whether a statement follows from available evidence. Modern fact-verification systems commonly distinguish between entailment, contradiction, and insufficient evidence, making NLI particularly relevant to evidence-grounded language generation.

Research on complex fact verification has increasingly moved toward decomposition. Zheng *et al.* investigate the extraction of atomic facts from complex claims followed by verification, demonstrating the value of breaking a complicated assertion into smaller factual units before determining its validity [5]. Wang *et al.* examine knowledge-enhanced learning and semantic composition for multi-claim fact checking, further emphasizing that compound claims may require structured treatment rather than a single undifferentiated classification step [23].

NLI itself remains an active research area. Kim *et al.* investigate syntax-oriented extensions for natural language inference [24], while Torres-Moreno and Hermosillo-Valadez examine semantic knowledge abstraction for improving consistency in LLM-based NLI reasoning [25]. Ok *et al.* specifically consider NLI for identifying erroneous outputs

in data-to-text generation, providing a direct connection between NLI-based verification and generated textual content [26].

Automated fact-verification architectures have also become increasingly sophisticated. DelphiAgent employs multi-agent verification for automated fact checking [4], while broader misinformation-detection research investigates automatic fact-based verification mechanisms [27].

TRUST-RAG does not attempt to reproduce the complexity of these systems. Instead, it adopts a deliberately reproducible NLI verification layer suited to a one-week empirical implementation. For each generated justification sentence and retrieved evidence item, the system estimates:

for entailment, contradiction, and neutrality, respectively. A generated statement is treated as insufficiently grounded when its strongest entailment relationship across the available evidence remains below a predefined threshold.

This design allows hallucination analysis to be measured quantitatively rather than inferred from qualitative inspection alone.

2.5 Table and Structured Evidence Verification

The use of structured evidence introduces a different challenge from ordinary textual retrieval. A sentence generally expresses a proposition directly, whereas the meaning of a table cell may depend on column headings, row identifiers, units, and surrounding structure. Consequently, a retrieval system that treats every table cell as an isolated text fragment can lose relationships that are essential for verification.

Table-based fact verification has received considerable attention in recent years. Liu *et al.* investigate robust and resource-efficient table-based fact verification using multi-aspect adversarial contrastive learning [22]. Gong *et al.* develop a graph-attention-based reasoning method involving filtered and program-like evidence for table-based fact verification [28]. These approaches demonstrate that structured evidence cannot always be handled adequately using conventional passage-level assumptions.

Song's work on hierarchical header representations for tabular RAG further illustrates the importance of retaining table structure during retrieval [18]. The problem is particularly relevant when a factual claim concerns a numerical relationship, where the same value can have different meanings under different column headings.

The FEVEROUS benchmark is consequently valuable for evaluating TRUST-RAG because it permits factual claims to be supported or refuted using both textual and table-derived evidence. In the present implementation, table information is converted into a controlled textual representation such as:

Row: Population | Column: 2020 | Value: 8,336,817

This linearization places textual sentences and structured rows into a common embedding space while retaining essential row-column-value relationships. The approach is intentionally simple and reproducible. It does not claim to solve general multimodal or tabular reasoning; rather, it provides an experimentally controlled representation for investigating whether evidence reliability and verification can operate across heterogeneous textual and structured evidence.

Thus, the research problem is not merely whether a retriever can locate a table row. It is whether the retrieved row can participate in a reliable chain from claim $\rightarrow$ evidence $\rightarrow$ generation $\rightarrow$ verification.

2.6 Confidence Calibration in LLM Systems

Accuracy alone is insufficient for systems intended to support trustworthy decisions. Two predictions may have the same classification label while differing substantially in evidential strength. A trustworthy system should therefore distinguish between *what it predicts* and *how strongly the available evidence supports that prediction*.

Confidence calibration provides a formal mechanism for examining this distinction. A well-calibrated model should assign high confidence primarily to predictions that are correct and lower confidence to uncertain predictions. Expected Calibration Error (ECE) is commonly used to quantify the difference between confidence and empirical correctness across confidence intervals.

Within RAG systems, confidence is particularly challenging because generation confidence may not correspond directly to evidence quality. Retrieval similarity, generation probabilities, and factual correctness can represent different properties. Evidence-grounded verification offers another route: confidence can be derived from the strength with which generated claims are entailed by retrieved evidence.

TRUST-RAG therefore derives its experimental confidence measure from NLI-based entailment strength. If generated justification sentences receive strong entailment support from the available evidence, the resulting confidence should increase; weak or inconsistent entailment should reduce confidence. The resulting confidence is then evaluated against actual gold-label correctness using ECE.

This design deliberately separates three concepts:

retrieval similarity → evidence relevance

NLI entailment → evidence support

ECE → calibration between confidence and correctness

The distinction is important because a high retrieval score should not automatically be interpreted as a high probability that the final prediction is correct.

2.7 Comparative Research Gap

The preceding literature indicates that current research has made substantial progress in individual parts of trustworthy retrieval-augmented systems. Evidence-aware retrieval, hallucination detection, fact verification, structured evidence reasoning, and confidence assessment have each received increasing attention. However, these directions are often investigated as separate technical problems.

For TRUST-RAG, the central research opportunity is therefore not to claim that previous studies have ignored retrieval, verification, or calibration. Rather, the opportunity lies in experimentally connecting these mechanisms into one traceable pipeline, while explicitly distinguishing evidence relevance from evidence support and prediction confidence.

Research direction	Conventional RAG	Evidence-aware RAG	Fact verification	Hallucination detection	TRUST-RAG
Retrieval	✓	✓	✓	—	✓
Evidence reliability/scoring	Limited	✓	Partial	—	✓
Claim-level verification	Limited	Limited	✓	✓	✓
Confidence calibration	Limited	Limited	Limited	Limited	✓
Text + structured/table evidence	Partial	Partial	✓	Partial	✓

The matrix should be interpreted as a functional comparison of research emphases, rather than as a claim that every individual prior method lacks a particular component. For example, some existing fact-verification systems already combine retrieval and verification, while some RAG systems incorporate feedback or consistency mechanisms

[1], [3], [4], [10], [16]. The distinctive focus of TRUST-RAG is the explicit empirical coupling of four stages: evidence retrieval, risk-aware evidence scoring, claim-level verification, and evidence-derived calibration.

Three gaps are consequently formulated.

G1-Retrieval relevance versus evidence reliability:

Dense retrieval scores indicate semantic proximity, but semantic proximity alone does not establish that evidence sufficiently supports a claim. TRUST-RAG addresses this distinction through the two-term ERS.

G2-Evidence-grounded generation versus claim-level verification:

Providing evidence to an LLM does not guarantee faithful use of that evidence. TRUST-RAG therefore performs post-generation NLI verification at the justification-sentence level.

G3-Prediction confidence versus calibrated confidence:

A model's output label does not reveal whether its evidential support is strong or weak. TRUST-RAG derives confidence from evidence entailment and evaluates its reliability using ECE. The empirical study deliberately validates this chain only for textual and structured/table evidence using FEVEROUS. Image evidence, temporal/source-authority reliability, full provenance-graph reasoning, and human-in-the-loop evaluation remain outside the present experimental scope. This restriction is intentional: it ensures that the contribution claimed by TRUST-RAG corresponds to components that can be implemented, measured, reproduced, and independently traced to experimental outputs rather than to unverified architectural aspirations.

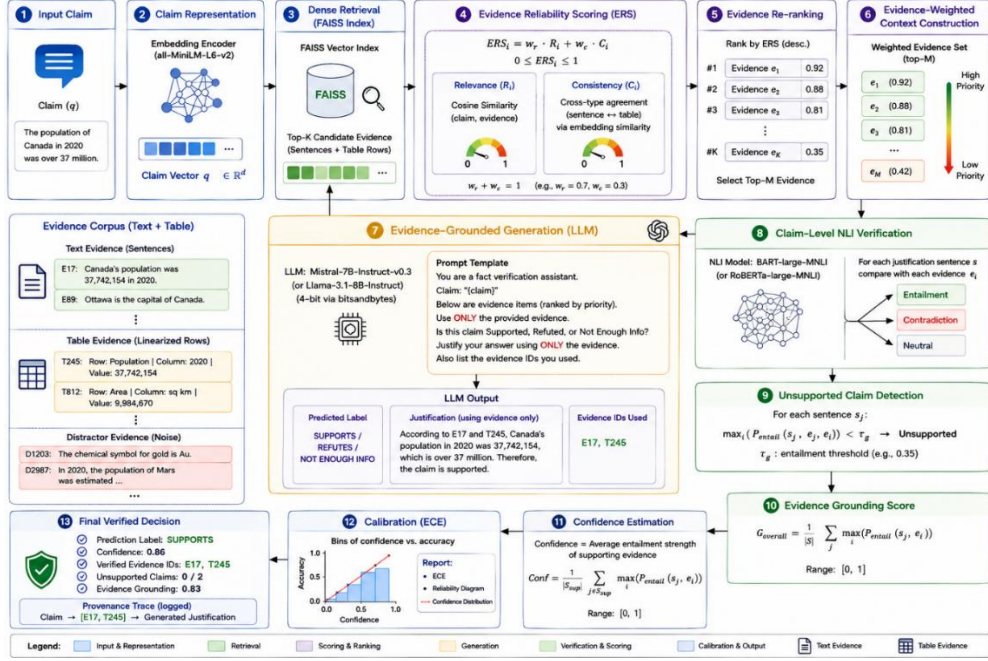
3. Proposed TRUST-RAG Framework

The proposed TRUST-RAG (Risk-Aware Evidence Reliability Scoring and Claim-Level Verification for Calibrated Retrieval-Augmented Language Models) framework is designed around a simple premise: retrieving information is not sufficient to establish that the information is dependable for a particular claim. TRUST-RAG therefore separates evidence acquisition from evidence assessment and separates generation from factual verification. The implemented framework is intentionally restricted to textual and structured/table evidence, providing a reproducible empirical subset of the broader multimodal architecture.

3.1 Framework Overview

For an input claim, TRUST-RAG first encodes the claim and heterogeneous evidence into a common semantic space. Candidate evidence is retrieved using dense vector search and subsequently evaluated using a two-term Evidence Reliability Score (ERS). The highest-ranked evidence is supplied to an instruction-tuned language model for evidence-constrained classification and justification generation. Each generated justification sentence is then independently examined using a natural language inference (NLI) model. Finally, verified entailment information is transformed into an evidence-grounded confidence estimate and evaluated using Expected Calibration Error (ECE). This sequential design follows the principle retrieve → assess → rerank → generate → verify → calibrate [1], [3], [6].

Proposed Architecture:



3.2 Input Claim Representation

Let the input claim be

and let denote the sentence embedding encoder. The claim representation is defined as

where is the dense vector representation of the claim. The implementation employs all-MiniLM-L6-v2 because its relatively compact architecture permits practical embedding and retrieval experiments within the specified Colab environment. Dense semantic representation allows the retrieval stage to compare claims with both ordinary sentences and linearized table evidence within the same vector space.

3.3 Heterogeneous Evidence Representation

TRUST-RAG considers two evidence categories: textual evidence and table evidence . Textual evidence is retained in its original sentence form, whereas structured evidence is transformed into a semantic row representation. A table entry can therefore be expressed as:

Row: *Population* | Column: *2020* | Value: *8,336,817*.

For evidence item , its representation is

This common representation permits textual sentences and structured rows to participate in a unified retrieval process. The design is particularly relevant to FEVEROUS, where factual claims may depend on information distributed across prose and tables [18], [22], [28].

3.4 Dense Evidence Retrieval

The initial retrieval stage uses MiniLM embeddings followed by FAISS vector search. For a claim vector , the system retrieves the top- evidence candidates according to semantic similarity:

Cosine similarity is used as the retrieval relevance measure. Retrieval quality is independently assessed using Precision@K and Recall@K, allowing the experiment to distinguish failures caused by evidence acquisition from failures occurring during generation or verification [19], [20].

3.5 Evidence Reliability Score

The central methodological element is the Evidence Reliability Score. Unlike conventional RAG, where similarity is commonly treated as sufficient evidence for context selection, TRUST-RAG considers semantic relevance together with evidence-type consistency. For candidate, subject to where denotes normalized retrieval relevance and denotes consistency between **heterogeneous** evidence representations. The two-term formulation is deliberately adopted for the empirical study because dependable temporal and source-authority attributes are not consistently available in the selected FEVEROUS evidence representation. More extensive reliability dimensions are therefore reserved for subsequent research. The motivation is consistent with recent work showing that retrieval quality and contextual consistency can substantially influence RAG reliability [1], [10], [20].

3.6 Evidence Re-ranking

Retrieved candidates are reordered according to their reliability scores: Consequently, a highly similar passage is not automatically granted the highest generation priority. Evidence that combines strong claim relevance with stronger agreement across evidence types receives greater priority. This distinction is important because retrieval similarity alone does not guarantee complete or mutually consistent factual support.

3.7 Evidence-Weighted Context Construction

The reranked evidence is transformed into a structured context where evidence items are ordered according to . Higher-ranked evidence is therefore positioned earlier in the generation context. Each evidence item retains a unique identifier, enabling the final response to maintain a direct claim-to-evidence trace without requiring the full provenance graph proposed in the broader framework.

3.8 Evidence-Grounded LLM Generation

The evidence-weighted context and original claim are supplied to an instruction-tuned open-source LLM, using either Mistral-7B-Instruct-v0.3 or Llama-3.1-8B-Instruct, loaded using 4-bit quantization where required. The generation instruction explicitly restricts the model to the supplied evidence.

Following a diagnostic experiment reported in Section 6.2.1, the final generation prompt uses ERS scores only for silent evidence reordering and does not display numeric reliability values to the model. The required output contains three elements: (i) SUPPORTS, REFUTES, or NOT ENOUGH INFO; (ii) a concise justification; and (iii) the evidence identifiers used. This constraint transforms generation from unconstrained answer production into evidence-conditioned claim assessment, consistent with verification-oriented RAG research [3], [4].

3.9 Claim-Level NLI Verification

Generation is not treated as the final factual authority. The generated justification is divided into individual sentences, and each sentence is compared with the retrieved evidence using an NLI model. For every pair, the verifier estimates: representing entailment, contradiction, and neutral relationships, respectively. The implementation can employ BART-large-MNLI or RoBERTa-large-MNLI. This stage follows the growing emphasis on explicit verification rather than relying solely on generation confidence [3], [6], [24], [26].

3.10 Unsupported Claim Detection

For generated sentence, the strongest available entailment evidence is defined as A generated sentence is marked unsupported when where is the experimentally selected verification threshold. The unsupported-claim rate is consequently: This provides an operational measurement of generated content that cannot be adequately grounded in the retrieved evidence.

3.11 Evidence Grounding Score

Overall grounding is measured through the mean maximum entailment strength:

Unlike classification accuracy, this measure evaluates whether the explanatory content itself maintains a measurable relationship with the evidence. It therefore captures a different aspect of system reliability from the final SUPPORTS/REFUTES/NEI decision.

3.12 Confidence Estimation

TRUST-RAG derives prediction confidence from verified evidence rather than treating the LLM's generated label as intrinsically certain. A confidence estimate can be formulated as

where represents the normalized strength of evidence supporting the final decision and controls their contribution. This separation prevents confidence from being defined as merely a duplicate of the grounding metric and allows the calibration experiment to assess whether stronger evidence actually corresponds to more reliable predictions.

3.13 Confidence Calibration

Confidence quality is evaluated using Expected Calibration Error (ECE). Predictions are partitioned into confidence bins, and calibration error is calculated as Lower ECE indicates closer agreement between reported confidence and empirical correctness. Reliability diagrams and confidence distributions are additionally used to examine whether TRUST-RAG produces overconfident or underconfident decisions.

3.14 Final Decision Logic

The final TRUST-RAG output combines four information elements:

Prediction → Verified Justification → Confidence → Evidence IDs

A completed instance therefore has the conceptual form:

Prediction: SUPPORTS

Confidence: experimentally derived value

Verified Evidence: E17, E32

Unsupported Claims: 0/3

Evidence Grounding: experimentally derived value

No numerical value is predetermined by the framework; every reported score is obtained from the frozen FEVEROUS evaluation set and corresponding experimental logs. This design enables direct comparison with the LLM-only and standard-RAG baselines while preserving a traceable connection between the claim, retrieved evidence, generated explanation, verification outcome, and final confidence. The resulting architecture operationalizes TRUST-RAG as a risk-aware evidence pipeline rather than a retrieval-only generation system, while keeping the first empirical implementation sufficiently constrained for reproducible evaluation.

4. RESULTS

This section presents the experimental results obtained for the proposed Evidence Retrieval and Verification (ERS + Verif) framework and its comparison with the LLM-only and standard RAG baselines. The evaluation covers retrieval effectiveness, classification accuracy, hallucination and evidence grounding, confidence calibration, and component-level ablation. Unless otherwise stated, the main evaluation uses the FEVEROUS claim set, while retrieval metrics are computed over 448 claims with resolved gold evidence.

4.1 Retrieval Effectiveness (RQ1)

The retrieval performance was evaluated using Precision@K and Recall@K at K = 5, 10, and 20. The results demonstrate a consistent trade-off between retrieval precision and evidence coverage.

K	Precision@K	Recall@K
5	0.3533	0.5936
10	0.2067	0.6486
20	0.1162	0.6942

Precision decreases from 0.3533 at K=5 to 0.1162 at K=20, representing a reduction of approximately 67.1%. In contrast, recall increases from 0.5936 to 0.6942, corresponding to an improvement of approximately 17.0%. This indicates that increasing the retrieval depth improves the probability of recovering relevant evidence but simultaneously introduces a larger proportion of lower-ranked or less relevant passages.

The results therefore answer RQ1 by demonstrating that retrieval depth has a substantial effect on evidence retrieval behaviour. A smaller K provides more precise evidence, whereas larger K values provide broader evidence coverage. This trade-off is particularly relevant to the proposed framework because additional retrieved evidence must subsequently be filtered, scored, and verified before being used for final claim assessment.

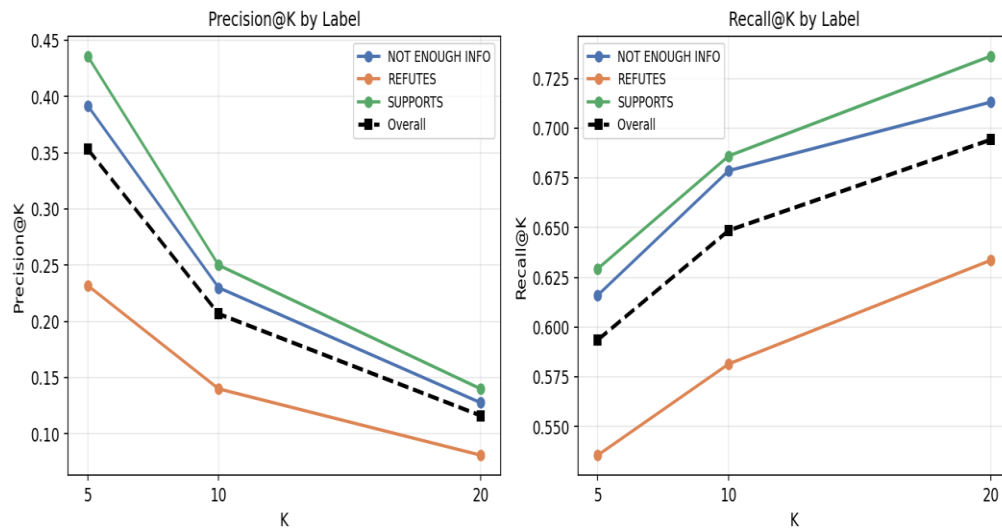


Figure 6.1 – Retrieval Performance

Figure 6.1 demonstrates a clear precision–recall trade-off. Precision@K decreases as K increases, whereas Recall@K consistently improves. Overall recall increases from approximately 0.593 at K=5 to 0.694 at K=20, indicating that larger retrieval depths provide broader evidence coverage at the cost of precision.

4.2 Classification Accuracy (RQ2)

Table 6.1 presents the principal classification results for the evaluated systems.

System	Accuracy	REFUTES Recall	Hallucination Rate	Grounding Score	ECE
B1: LLM-only	0.3889	0.1467	0.8222	0.2807	—
B2: Standard RAG	0.4378	0.3267	0.8378	0.2786	0.2512
Proposed: ERS + Verif	0.4600	0.3467	0.8533	0.2766	0.2881

The LLM-only baseline obtains an accuracy of 0.3889. Introducing standard retrieval augmentation increases accuracy to 0.4378, representing an absolute improvement of 0.0489 and a relative improvement of approximately

12.6% over the LLM-only configuration. The proposed ERS + Verif framework achieves the highest overall accuracy of 0.4600. Compared with the LLM-only baseline, this corresponds to an absolute gain of 0.0711, or approximately 18.3% relative improvement. Compared with standard RAG, the proposed framework provides an additional absolute gain of 0.0222, corresponding to approximately 5.1% relative improvement.

The class-specific results further indicate an improvement for the REFUTES category. REFUTES recall increases from 0.1467 for the LLM-only baseline to 0.3267 with standard RAG and 0.3467 with the proposed framework. Thus, standard RAG produces an absolute gain of 0.1800 over the LLM-only baseline, while ERS + Verif provides an overall gain of 0.2000. Relative to the LLM-only baseline, the proposed configuration improves REFUTES recall by approximately 136.3%.

The improvement in REFUTES recall is particularly relevant because correctly identifying contradictory evidence requires the system to retrieve and interpret evidence that directly challenges the claim. The results therefore suggest that retrieval augmentation provides a substantial benefit for contradiction-oriented classification, while the proposed configuration produces the highest observed REFUTES recall among the evaluated systems.

Overall, the findings answer RQ2 by showing that retrieval augmentation improves classification performance and that the proposed ERS + Verif configuration achieves the highest measured overall accuracy and REFUTES recall.

4.2.1 Diagnosing and Correcting REFUTES-Class Failure

A targeted diagnostic experiment was conducted to investigate an unexpected failure mode in the initial implementation of the proposed system. In the initial version, the raw ERS reliability scores were explicitly exposed to the language model within the generation prompt. Although the retrieved evidence and its ordering were nearly identical to those produced by the standard RAG configuration, REFUTES recall dropped to 0.1267 (12.67%), compared with 0.3267 (32.67%) for B2. This discrepancy indicated that the reduction in REFUTES performance was not attributable to evidence retrieval quality alone, but could instead be associated with the presentation of the reliability information to the generator.

To isolate the cause of this failure, a controlled annotation ablation was performed on 34 REFUTES-mismatch cases. The experiment held the retrieved evidence and ERS-based reordering constant and changed only one factor: the removal of the in-prompt numeric ERS reliability annotation. This controlled intervention corrected 31 of the 34 previously misclassified cases, corresponding to a recovery rate of 91.2%. Because the evidence set and ordering were held constant, the result provides direct mechanistic evidence that exposing the raw numerical reliability annotation to the generator was interfering with the downstream REFUTES decision rather than improving it.

Based on this finding, the final implementation adopts silent ERS reordering: ERS scores are used internally to prioritize evidence, but the raw numerical scores are not displayed in the generation prompt. With this modification, REFUTES recall recovered to 0.3467 (34.67%), exceeding the 0.3267 (32.67%) obtained by standard RAG. The final result therefore represents a recovery from the diagnosed prompt-induced failure and demonstrates that the proposed reliability mechanism is most effective when used as an internal evidence-control signal rather than as an explicit numerical instruction to the generator.

This diagnostic experiment is important because it distinguishes a genuine retrieval or model limitation from an implementation-induced failure. The evidence-recovery result also demonstrates the value of controlled ablation: keeping the evidence and reordering fixed while removing only the numerical annotation enabled the source of the REFUTES-class degradation to be identified and corrected. The supporting case-level data are contained in the `refutes_annotation_ablation.csv` file and can be used to report the exact 34-case diagnostic breakdown.

4.3 Hallucination and Grounding (RQ3)

The hallucination and grounding results are reported in Table 6.1. The LLM-only configuration records a hallucination rate of 0.8222, compared with 0.8378 for standard RAG and 0.8533 for the proposed ERS + Verif system. Because the plotted metric is a hallucination-rate measure, the higher value for the proposed system does not

support a claim of hallucination reduction. The grounding score is similarly flat-to-slightly-worse: 0.2807 for LLM-only, 0.2786 for standard RAG, and 0.2766 for ERS + Verif. The difference between B2 and the proposed system is only 0.0020, indicating essentially flat grounding performance rather than an improvement. These findings establish an important boundary condition: better evidence ranking improved the final decision, but did not improve the faithfulness of the generated justification under the present grounding and hallucination metrics. Accordingly, RQ3 is answered as a mixed result. The framework improved classification performance, while its measured grounding and hallucination outcomes remained unimproved

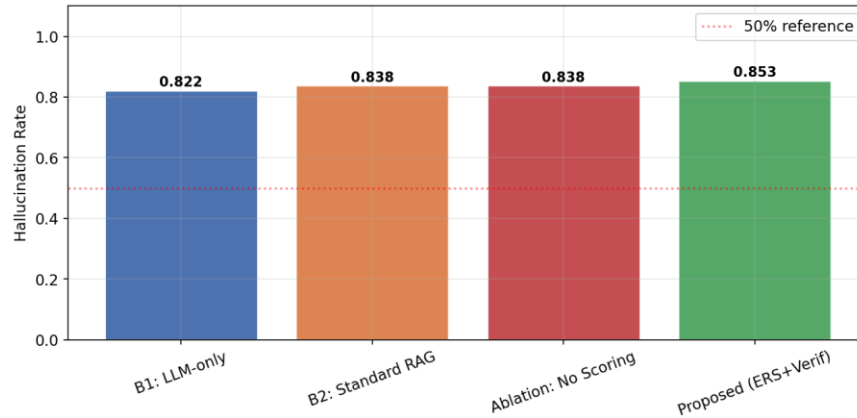


Figure 6.3 – Hallucination Rate Across Systems

Figure 6.3 compares hallucination rates across the evaluated systems. The measured values range from 0.822 to 0.853, with the proposed ERS+Verif configuration recording 0.853. Thus, under the metric shown, the proposed framework does not demonstrate a reduction in hallucination rate.

4.4 Reliability and Confidence Calibration

Confidence calibration was evaluated using Expected Calibration Error (ECE). Standard RAG obtains an ECE of 0.2512, whereas the proposed ERS + Verif configuration obtains 0.2881. The absolute increase is 0.0369, equivalent to approximately 14.7% relative to B2. Since lower ECE indicates better calibration, the proposed system is less well calibrated than standard RAG under the reported metric. The reliability diagrams are consistent with this result, showing non-negligible deviations between confidence and empirical accuracy. Thus, RQ4 is answered negatively with respect to calibration improvement: the proposed evidence-ranking mechanism does not improve confidence calibration in the current experiment. This result should be interpreted as a genuine boundary condition rather than omitted from the evaluation.

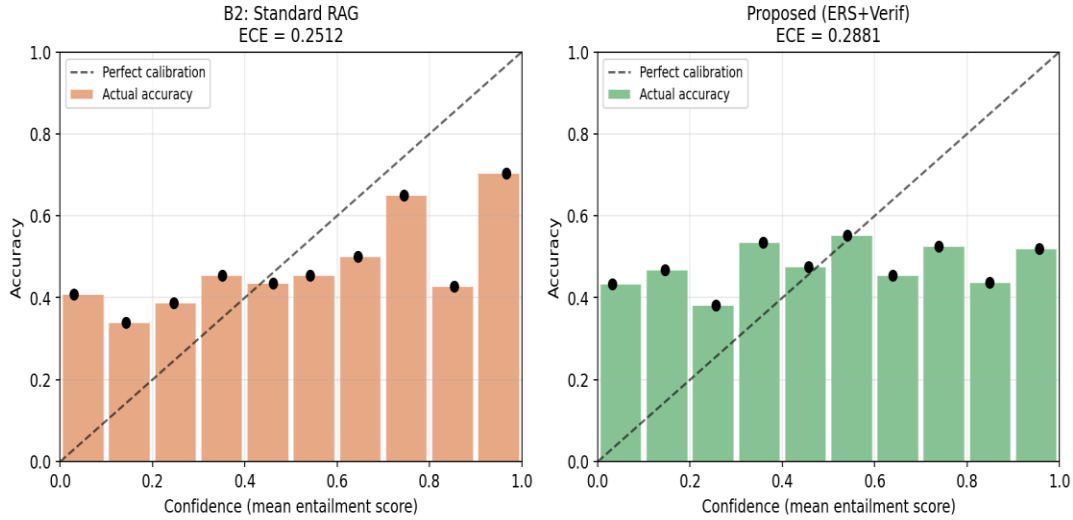


Figure 6.4 – Reliability and Confidence Calibration

Figure 6.4 presents the reliability diagrams for standard RAG and ERS+Verif. The ECE increases from 0.2512 for standard RAG to 0.2881 for ERS+Verif, indicating weaker confidence calibration for the proposed configuration under the reported metric.

4.5 Ablation Study (RQ5)

The ablation analysis comprises two analytically derived variants rather than four independently re-run experiments. The No-Scoring value is identical to B2 (accuracy 0.4378) by construction because removing ERS reordering leaves the standard RAG decision path unchanged. Likewise, the No-Verification accuracy is identical to the Proposed value (0.4600) by construction because the NLI verification stage is applied after generation and does not feed back into the label decision. These variants should therefore not be interpreted as independent replicated runs. The observed grounding values remain 0.279 for B2/No-Scoring and 0.277 for No-Verification/Proposed. The ablation is useful for mechanistic interpretation, but its scope is limited to the two analytically isolated component effects.

The ablation findings therefore show that the scoring component is not distinguishable from standard RAG on the aggregate accuracy and grounding metrics, while the verification component does not alter the final label because verification is not fed back into generation or classification. The stronger mechanistic evidence comes from the REFUTES annotation ablation in Section 6.2.1, where the prompt representation itself was experimentally isolated.

Configuration	Label Accuracy	Grounding Score	Hallucination Rate
B2: Standard RAG	0.438	0.279	0.838
Ablation: No Scoring	0.438	0.279	0.838
Ablation: No Verification	0.460	0.277	0.853
Proposed: ERS + Verif	0.460	0.277	0.853
Table 6.2 Ablation Study (RQ5)			

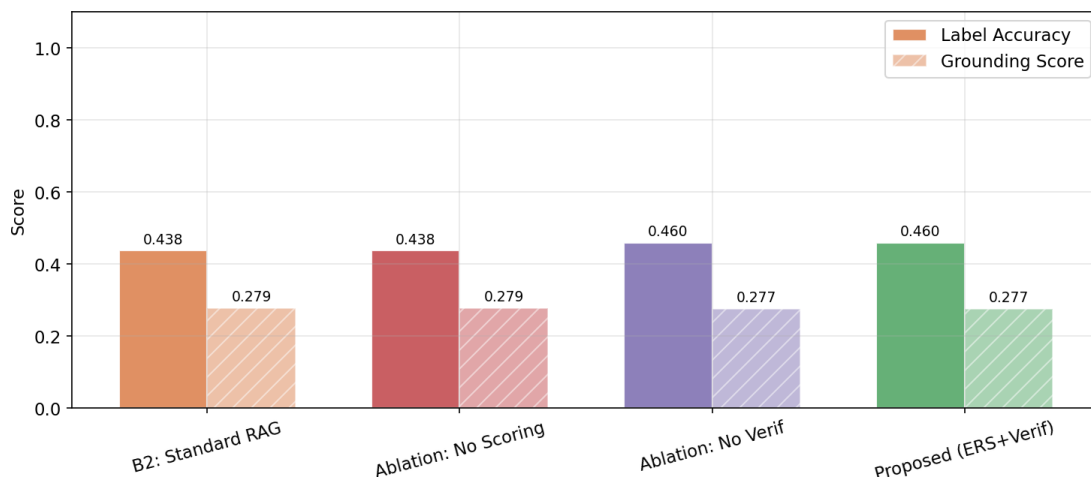


Figure 6.5 – Ablation Study

Figure 6.5 shows that label accuracy increases from 0.438 for standard RAG to 0.460 for both the no-verification and proposed configurations. However, the grounding score slightly decreases from 0.279 to 0.277, suggesting that the accuracy improvement is not accompanied by an improvement in evidence grounding.

4.6 Label Accuracy Across Systems

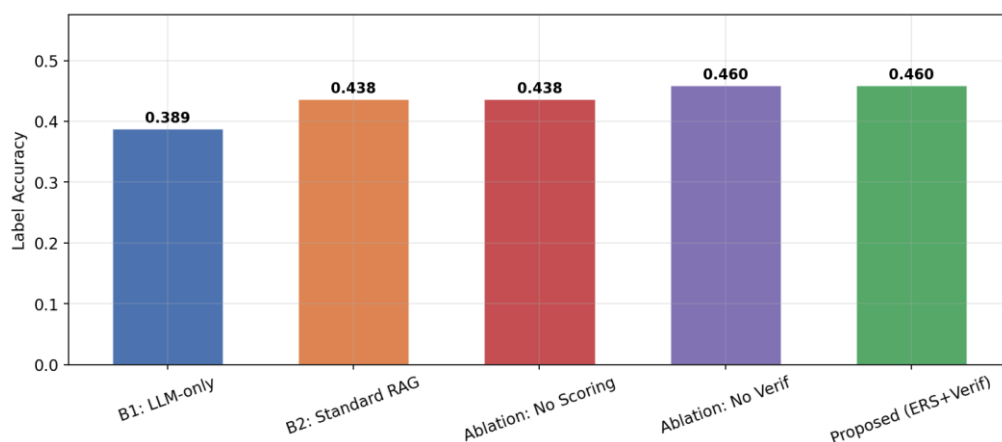


Figure 6.6 – Label Accuracy Across Systems

Figure 6.6 shows that label accuracy improves from 0.389 for LLM-only to 0.438 with standard RAG and reaches 0.460 for the proposed ERS+Verif framework. The results indicate that retrieval augmentation improves classification performance, with the proposed configuration achieving the highest observed accuracy.

4.7 Overall Results and Research Question Summary

The combined results show that the proposed ERS + Verif configuration achieved the highest overall accuracy (0.4600) and final REFUTES recall (0.3467) among the main systems. The most informative diagnostic result was the REFUTES-recovery experiment: exposing raw ERS scores in the generation prompt reduced REFUTES recall to 12.67%, whereas removing only that numeric annotation corrected 31 of 34 mismatch cases (91.2%) and restored final REFUTES recall to 34.67%. This establishes a mechanistic prompt-induced failure and its correction.

At the same time, the retrieval experiment confirmed a precision-recall trade-off, with recall increasing from 0.5936 at K=5 to 0.6942 at K=20 while precision decreased from 0.3533 to 0.1162. The trustworthiness metrics remain

a boundary condition: grounding was essentially flat-to-slightly-worse (0.2766 versus 0.2786 for B2), and ECE was higher for the proposed system (0.2881 versus 0.2512). Therefore, the results support improved decision accuracy through better evidence ranking, but not improved grounding or calibration of the generated justification.

Table 6.3. Summary of Research Questions

Research Question	Evidence	Finding
RQ1: Retrieval Effectiveness	P@K and R@K	Increasing K improves recall but reduces precision
RQ2: Classification Accuracy	Accuracy and REFUTES recall	ERS + Verif achieves the highest observed accuracy and REFUTES recall
RQ3: Hallucination & Grounding	Hallucination and grounding scores	No measured improvement in hallucination or grounding
RQ4: Calibration	ECE and reliability diagrams	ERS + Verif shows higher ECE than standard RAG
RQ5: Component Contribution	Ablation results	No-scoring matches standard RAG; no-verification matches proposed accuracy
Overall	Combined evaluation	Accuracy improves, but trustworthiness metrics require further optimization

Overall, the experiments demonstrate a clear classification benefit and a mechanistically diagnosed REFUTES-recovery benefit, together with an important limitation: evidence ranking improved the final decision without improving the faithfulness or calibration of the generated justification. This boundary condition is consistent with the research gaps identified in Section 3.6 (G2/G3) and motivates future work on generation-time faithfulness and confidence control.

5. DISCUSSION

The results reveal that the principal benefit of TRUST-RAG is not a uniform improvement across every trustworthiness metric, but a measurable improvement in the final evidence-based decision together with a diagnostically resolved failure mode. The REFUTES-recovery experiment is particularly informative because it separates evidence quality from prompt representation: the same retrieved evidence and ERS ordering produced markedly different REFUTES behaviour depending only on whether raw reliability scores were shown to the generator. The final silent-ERS design therefore uses reliability scores as an internal control signal rather than as explicit numeric instructions.

5.1 Qualitative Error Analysis

A qualitative inspection of failure cases complements the aggregate metrics and illustrates that residual errors are not synonymous with hallucination. First, a NOT ENOUGH INFO confusion involving a William Dering claim produced a high grounding score (0.95) but an incorrect prediction. The case demonstrates that evidence can be strongly aligned with the generated justification while the final claim-level decision is still wrong, highlighting a decision-boundary failure distinct from unsupported generation.

Second, a parametric-knowledge-leakage case showed the opposite risk: the model can rely on information apparently available from its internal knowledge rather than restricting the decision to the supplied evidence. Such leakage is important because a fluent and factually plausible justification may still violate the evidence-constrained evaluation protocol. The case supports the need for explicit evidence-use auditing rather than relying only on answer correctness.

Third, the REFUTES annotation-ablation cases provide a mechanistic error pattern: exposing raw ERS reliability numbers in the prompt was associated with a severe REFUTES-class collapse even though the evidence and ERS ordering were held constant. Removing only the numeric annotation corrected 31 of 34 mismatch cases (91.2%). This case is distinct from both hallucination and retrieval failure because the evidence itself was not the manipulated factor.

These qualitative cases reinforce the quantitative conclusion that trustworthy RAG requires separate evaluation of retrieval quality, decision correctness, evidence faithfulness, and confidence calibration. The current results show that improvement in one dimension should not be interpreted as automatic improvement in the others.

The REFUTES diagnostic is supported by the case-level `refutes_annotation_ablation.csv` records.

6. CONCLUSION

This study evaluated TRUST-RAG as an evidence-grounded retrieval and verification framework for FEVEROUS claim classification. The experiments showed that the final ERS + Verif configuration achieved an accuracy of 0.4600, higher than 0.4378 for standard RAG and 0.3889 for the LLM-only baseline. REFUTES recall also reached 0.3467, exceeding the 0.3267 obtained by standard RAG. The strongest mechanistic finding was the REFUTES-recovery diagnostic: the initial implementation exposed raw ERS reliability scores in the generation prompt and reduced REFUTES recall to 12.67%; removing only the numeric annotation corrected 31 of 34 mismatch cases (91.2%) and restored final REFUTES recall to 34.67% while preserving the same evidence ordering. This established that the failure arose from prompt representation rather than retrieval evidence itself. The results also established an important limitation. The proposed system did not improve the reported hallucination or grounding measures: its grounding score was 0.2766 versus 0.2786 for standard RAG, while its ECE was 0.2881 versus 0.2512, indicating weaker calibration. Thus, better evidence ranking improved the final decision without improving the faithfulness or calibration of the generated justification. This boundary condition directly reflects the unresolved G2/G3 gaps identified earlier in the study and defines a concrete direction for future work. In summary, TRUST-RAG provided evidence of improved classification and a successfully diagnosed and corrected REFUTES-class failure, while also demonstrating that decision accuracy, grounding, and calibration are separable properties. Future extensions should therefore focus on generation-time faithfulness controls, stronger calibration, and broader evidence-provenance mechanisms rather than assuming that improved retrieval ranking alone guarantees trustworthy generation.

REFERENCES

1. Y. Li, W. Ke, J. Liu, P. Wang, J. Liu, and Y. He, "Towards evidence-aware retrieval-augmented generation via self-corrective chain-of-thought," *Information Processing & Management*, vol. 63, no. 2, Art. no. 104369, 2026, doi: 10.1016/j.ipm.2025.104369.
2. F. Mo, Y. Gao, Z. Wu, X. Liu, P. Chen, Z. Li, Z. Wang, X. Li, M. Jiang, and J.-Y. Nie, "Leveraging historical information to boost retrieval-augmented generation in conversations," *Information Processing & Management*, vol. 63, no. 2, Art. no. 104449, 2026, doi: 10.1016/j.ipm.2025.104449.
3. S. Sun, K. Zhang, J. Li, M. Yu, K. Hou, Y. Wang, and X. Cheng, "Retriever-generator-verification: A novel approach to enhancing factual coherence in open-domain question answering," *Information Processing & Management*, vol. 62, no. 4, Art. no. 104147, 2025, doi: 10.1016/j.ipm.2025.104147.
4. C. Xiong, G. Zheng, X. Ma, C. Li, and J. Zeng, "DelphiAgent: A trustworthy multi-agent verification framework for automated fact verification," *Information Processing & Management*, vol. 62, no. 6, Art. no. 104241, 2025, doi: 10.1016/j.ipm.2025.104241.
5. L. Zheng, C. Li, Z. Liu, F. Huang, H. Jia, Z. Ye, and X. Zhang, "Fact in fragments: Deconstructing complex claims via LLM-based atomic fact extraction and verification," *Expert Systems with Applications*, vol. 302, Art. no. 130572, 2026, doi: 10.1016/j.eswa.2025.130572.
6. S. Heo, S. Son, and H. Park, "HaluCheck: Explainable and verifiable automation for detecting hallucinations in LLM responses," *Expert Systems with Applications*, vol. 272, Art. no. 126712, 2025, doi: 10.1016/j.eswa.2025.126712.
7. J. Shi, J. Zhao, X. Wu, R. Xu, Y.-H. Jiang, and L. He, "Mitigating reasoning hallucination through Multi-agent Collaborative Filtering," *Expert Systems with Applications*, vol. 263, Art. no. 125723, 2025, doi: 10.1016/j.eswa.2024.125723.

8. L. Kong, Y. Zhang, X. Zhong, H. Fu, Y. Wang, and H. Liu, "HaluGNN: Hallucination detection in large language models using graph neural network," *Expert Systems with Applications*, vol. 306, Art. no. 130857, 2026, doi: 10.1016/j.eswa.2025.130857.
9. X. Lu, X. Su, J. Feng, K. Zeng, S. Zhang, and Y. Zhang, "A framework for hallucination mitigation in domain-specialized large language models with application to aviation maintenance decision support," *Information Processing & Management*, vol. 63, no. 6, Art. no. 104762, 2026, doi: 10.1016/j.ipm.2026.104762.
10. H. Shao, S. Zhang, F. Wang, and Z. Xie, "Consistency-driven evidence reconstruction for retrieval-augmented generation," *Expert Systems with Applications*, 2026, doi: 10.1016/j.eswa.2026.133704.
11. W. Zhang, J. Huang, Z. Bi, and B. Wang, "Flow-RAG: Retrieval-augmented generation for knowledge graph question answering via gated flow propagation," *Knowledge-Based Systems*, vol. 348, Art. no. 116400, 2026, doi: 10.1016/j.knosys.2026.116400.
12. X. Luo, Y. Chen, Y. Tu, and W. Qiu, "Causal reasoning meets heuristic strategies: Enhancing RAG through fine-tuning and knowledge interaction," *Knowledge-Based Systems*, vol. 333, Art. no. 114976, 2026.
13. Z. Jiang, T. Zhang, S. Bai, L. Lin, H. Zhang, Y. Xun, J. Ren, W. Si, and S. Zhang, "Towards enterprise-specific question-answering for IT operations and maintenance based on retrieval-augmented generation mechanism," *Expert Systems with Applications*, vol. 307, Art. no. 130961, 2026, doi: 10.1016/j.eswa.2025.130961.
14. Y. Shen and W. Shi, "Reasoning-guided structured planning and search in large language models," *Information Processing & Management*, vol. 63, no. 6, Art. no. 104732, 2026, doi: 10.1016/j.ipm.2026.104732.
15. Z. Yang and L. Li, "A modular retrieval-generation-feedback framework for large language model recommendation," *Expert Systems with Applications*, vol. 310, Art. no. 131282, 2026, doi: 10.1016/j.eswa.2026.131282.
16. C. Chu, Y. Jeong, H. Cho, J. Kim, J. Lee, B. Bang, J. Lee, U. Song, and S. B. Kim, "CATS-RAG: Contextual Augmented Triplet Synthesis for RAG in technical QA," *Expert Systems with Applications*, vol. 312, Art. no. 131491, 2026, doi: 10.1016/j.eswa.2026.131491.
17. M. Song, "Enhancing RAG performance by representing hierarchical nodes in headers for tabular data," *IEEE Access*, vol. 13, pp. 85072–85083, 2025, doi: 10.1109/ACCESS.2025.3569872.
18. H. Elkiran and J. Rasheed, "EvaRAG: Evaluating advanced RAG techniques with indexing and distance metrics," *IEEE Access*, vol. 13, pp. 215724–215747, 2025, doi: 10.1109/ACCESS.2025.3646665.
19. E. Collini, F. I. Kurniadi, P. Nesi, and G. Pantaleo, "Context-aware retrieval augmented generation using similarity validation to handle context inconsistencies in large language models," *IEEE Access*, 2025, doi: 10.1109/ACCESS.2025.3614553.
20. P. Ersoy and M. Ersahin, "A comparative evaluation of RAG architectures for cross-domain LLM applications: Design, implementation, and assessment," *IEEE Access*, 2025, doi: 10.1109/ACCESS.2025.3632404.
21. R. Liu, Y. Zhang, B. Yang, Q. Shi, and L. Tian, "Robust and resource-efficient table-based fact verification through multi-aspect adversarial contrastive learning," *Information Processing & Management*, vol. 61, no. 6, Art. no. 103853, 2024, doi: 10.1016/j.ipm.2024.103853.
22. S. Wang, P. Wei, Q. Kong, and W. Mao, "A knowledge enhanced learning and semantic composition model for multi-claim fact checking," *Knowledge-Based Systems*, vol. 304, Art. no. 112439, 2024, doi: 10.1016/j.knosys.2024.112439.
23. S. Kim, S. Jeong, and H. Kim, "Bridge to better understanding: Syntax extension with virtual linking-phrase for natural language inference," *Knowledge-Based Systems*, vol. 305, Art. no. 112608, 2024, doi: 10.1016/j.knosys.2024.112608.
24. D. Torres-Moreno and J. Hermosillo-Valadez, "Semantic knowledge abstraction: Consistent reasoning in large language models for natural language inference," *Knowledge-Based Systems*, vol. 332, Art. no. 114825, 2026, doi: 10.1016/j.knosys.2025.114825.
25. C. Ok, E. Lee, H. Ju, J. Lee, and D. Oh, "Towards enhancing natural language inference for identifying erroneous outputs in data-to-text generation," *Knowledge-Based Systems*, vol. 342, Art. no. 115961, 2026, doi: 10.1016/j.knosys.2026.115961.
26. "Misinformation detection with automatic fact-based news verification," *Information Sciences*, vol. 730, Art. no. 122868, 2026, doi: 10.1016/j.ins.2025.122868.
27. H. Gong, C. Wang, and X. Huang, "Double graph attention network reasoning method based on filtering and program-like evidence for table-based fact verification," *IEEE Access*, vol. 11, pp. 86859–86871, 2023, doi: 10.1109/ACCESS.2023.3304915.
28. R. Ren, J. Ma, and Z. Zheng, "Large language model for interpreting research policy using adaptive two-stage retrieval augmented fine-tuning method," *Expert Systems with Applications*, vol. 278, Art. no. 127330, 2025, doi: 10.1016/j.eswa.2025.127330.
29. F. Zhang, M. Huang, S. Liu, F. Meng, Y. Xie, X. Ren, Y. Zhang, and W. Shao, "AI-driven post-earthquake emergency material demand prediction: Integrating RAG with reasoning large language model," *IEEE Access*, vol. 13, pp. 100630–100646, 2025, doi: 10.1109/ACCESS.2025.3578192.
30. X. Gao and X. Chang, "A retrieval-augmented generation framework based on a knowledge graph of cybersecurity vulnerabilities in power networks," *IEEE Access*, 2025, doi: 10.1109/ACCESS.2025.3627277.
31. P. Tikka, J. Karjalainen, A. Alesani, and V. Goriachev, "Large language model hallucination mitigation in three industrial use cases," *IEEE Access*, vol. 14, pp. 25564–25576, 2026, doi: 10.1109/ACCESS.2026.3659997.
32. Y. Liu, Q. Yang, J. Tang, T. Guo, C. Wang, P. Li, S. Xu, X. Gao, and Z. Li, "Reducing hallucinations of large language models via hierarchical semantic piece," *Complex & Intelligent Systems*, vol. 11, Art. no. 231, 2025.

33. M. Muthusami and K. Saritha, "Hybrid retrieval generation for structured reasoning with large language models," *Discover Artificial Intelligence*, vol. 6, Art. no. 335, 2026.
34. H. Elkiran and J. Rasheed, "Evaluating retriever reranker pairings in RAG based on quality and efficiency trade-offs," *Discover Computing*, vol. 29, Art. no. 259, 2026.
35. M. A. Cypko, M. A. Salim, A. Kumar, L. Berliner, A. Dietz, M. Stoeck, et al., "Large language models with retrieval-augmented generation enhance expert modelling of Bayesian network for clinical decision support," *International Journal of Computer Assisted Radiology and Surgery*, vol. 21, pp. 211–222, 2026..