PT Jurnal Cendekia Indonesia
Journal of Information Systems and Business Technology
Homepage: https://journal.jci.co.id/jisbt
Vol. 01 No. 02 (2026) Page: 119-129
E-ISSN: 3109-8886

Corrective RAG with Iterative Evidence Verification and Query Reformulation for Scientific Fact Checking

Hao Wang, Laura Robinson, Brandon King

Department of Computer Science, University of Maryland, College Park, MD, USA
Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA
School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA

hao.wang.pro@outlook.com

Abstract

Scientific claim verification requires a system to retrieve relevant literature, identify evidentiary sentences, and assign a SUPPORT, CONTRADICT, or not-enough-information (NEI) label. A single retrieval pass is inexpensive but leaves recoverable evidence outside the candidate set, whereas unconditional iterative retrieval increases cost and can introduce query drift. This study presents a corrective retrieval-augmented verification pipeline that couples stemmed BM25 retrieval, evidence sentence ranking, a calibrated verifier agent, training-fold evidence memory, query reformulation, and reciprocal-rank fusion. The corrective gate is calibrated inside each training fold. It initiates a second retrieval pass only for low-confidence or weak-evidence cases, accepts a changed label only when the second verifier provides a calibrated confidence gain, and preserves a previously evidence-backed label when the second pass returns NEI. Experiments use the SciFact corpus of 5,183 abstracts and 1,109 labeled claims under the provided five-fold partition. The second pass raises document Recall@5 from 90.3% to 92.8%. The gated system invokes that pass for 17.0% of claims, reaches 68.4% accuracy, 0.656 Macro-F1, and the highest evidence-conditioned joint score of 0.597. Compared with always applying two passes, it improves accuracy by 4.78 points and Macro-F1 by 5.51 points while reducing processed tokens by 41.8%. The results show that iterative retrieval is most effective when evidence accumulation, label replacement, and cost are governed by separate verification decisions.

Keywords: Scientific Fact Checking; Corrective RAG; Iterative Retrieval; Query Reformulation; Evidence Verification; Scifact; Calibrated Gating; Reciprocal Rank Fusion

A. INTRODUCTION

Scientific claims circulate through research articles, clinical summaries, news reports, and public discussion faster than experts can inspect the supporting literature. A useful system must retrieve the source, identify the sentences carrying the evidentiary burden, and bind its decision to them. SciFact formalized this setting with expert-written claims, scientific abstracts, document labels, and sentence rationales [1]. Evidence-grounded disclosure QA likewise treats provenance as part of answer correctness [2], while FEVER established the retrieval–evidence–classification decomposition at larger scale [3]. Long-document clause analysis extends the requirement to dispersed contractual evidence [4], and retrieval–summary code intelligence couples finding a source with explaining it [5]. A scientific fact-checking survey identifies provenance, terminology, and domain transfer as persistent difficulties [6].

Scientific verification systems have developed as pipelines or joint models. SCIVER evaluated abstract retrieval, rationale selection, and label prediction [7]; VerT5erini adapted a sequence-to-sequence model to all three subtasks [8], and MultiVerS used full-document context with weak supervision [9]. SciBERT [10], BERT [11], and Sentence-BERT [12] provide semantic representations, but representation quality does not remove the need to control evidence access. Adjacent high-stakes studies reinforce this distinction: DenseNet cancer classification depends on visual representation [13], and minimal-residual-disease prediction operates under clinical uncertainty [14]. Human-uncertainty distillation learns from label distributions [15], uncertainty-aware late fusion calibrates modality confidence [16], and spectral-estimator analysis formalizes uncertainty [17]. These studies motivate using confidence as a control signal, not as a substitute for evidence.

Retrieval-augmented generation introduced non-parametric memory into knowledge-intensive language processing [18], and REALM incorporated retrieval into pretraining [19]. Dense Passage Retrieval learned a dual encoder [20], while Fusion-in-Decoder aggregated multiple passages [21]. Retrieval-grounded diagnosis also appears outside question answering: HDFS analysis links anomalies to failure narratives and selective refusal [22]; OpenStack analysis retrieves normal prototypes [23]; CI diagnosis connects logs to repair recommendations [24]; and AIOps frames fault

Received: August 22, 2025 | Revised: August 30, 2025 | Published Online: August 31, 2025

analysis as evidence-driven operations [25]. Across these settings, a verifier cannot use absent evidence, yet indiscriminate retrieval can introduce distracting neighbors.

Classical information retrieval supplies controlled mechanisms for this trade-off. BM25 provides sparse probabilistic ranking [26]; reciprocal rank fusion combines rankings without score calibration [27]; the vector-space model places queries and documents in one term space [28]; and relevance feedback expands queries from retrieved material [29], [30]. Spectral rank-aggregation theory studies ranking stability [31]. Query2doc generates pseudo-documents for expansion [32], while approximately shared-feature analysis clarifies cross-domain transfer [33]. Rewrite-Retrieve-Read trains a rewriter [34]; FLARE triggers retrieval under uncertainty [35]; IRCoT interleaves retrieval and reasoning [36]; hypothetical-document embeddings construct retrieval targets [37]; and Self-RAG learns retrieval and critique decisions [38]. E-commerce recommendation further shows that neighbor selection depends on explicit representation, privacy, and cold-start assumptions [39].

The common insight is that a second acquisition step can recover a missed signal. The engineering question is when it should run and how it should alter an evidence-backed decision. Related systems allocate resources selectively: risk-aware auto-bidding balances utility, budget, and tail risk [40]; progressive PDF rendering prioritizes functional content [41]; and counterfactual learning-to-rank evaluates policies under biased exposure [42]. Hybrid-cloud LLM deployment [43], edge video-quality prediction [44], and digital-twin mobility control [45] pose analogous cost–quality choices. In scientific verification, always retrieving twice pays maximum cost and exposes stable claims to query drift.

This paper addresses that instability with a closed-loop corrective RAG design. The first pass retrieves abstracts, selects five claim-relevant sentences, and produces an evidence-aware prediction. A gate combines evidence probability, label confidence, maximum sentence similarity, and retrieval margin. Only routed claims enter a training-fold query rewriter; the expanded query drives a second stemmed-BM25 search, and reciprocal rank fusion preserves both rankings. A calibrated confidence-gain test controls replacement, while a monotonic evidence rule prevents second-pass NEI from erasing first-pass SUPPORT or CONTRADICT. Calibration [46] and selective classification [47] motivate prediction–routing separation. DriftGuard's monitor–rollback logic [48] and noisy-neighbor-aware VM risk modeling [49] motivate treating new evidence as a monitored proposal rather than an automatic replacement.

The experimental contribution is a five-fold evaluation over every labeled SciFact claim, with out-of-fold predictions, test outputs, retrieval and rationale metrics, token-cost accounting, depth sweeps, gate trade-offs, ablations, bootstrap intervals, and error diagnostics. The comparison includes word BM25, stemmed BM25, character TF-IDF, first-pass rank fusion, unconditional two-pass retrieval, and the corrective policy under one verifier. Following the lesson from IoT anomaly detection that aggregate accuracy does not expose failure modes [50], the study uses paired claim-level resampling [51] and reports recovery, drift, and cost separately. Predictive DDoS mitigation likewise illustrates why detection quality and action cost should be evaluated together [52]; risk-dashboard design further motivates presenting distinct diagnostic dimensions rather than one composite score [53].

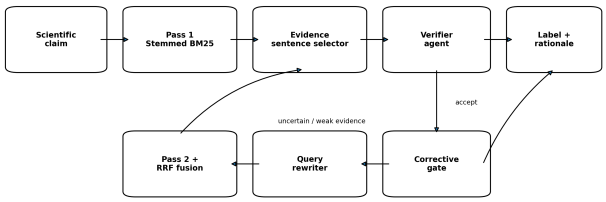
Method

A. Task formulation and data. Each example consists of a claim q, a corpus C of scientific abstracts, a gold label y in {SUPPORT, CONTRADICT, NEI}, and, for positive claims, one or more evidence rationales. The corpus contains 5,183 abstracts and 45,952 sentences. The labeled train and development files contain 1,109 distinct claims: 456 SUPPORT, 237 CONTRADICT, and 416 NEI. The release also contains 300 test claims without public labels. Quantitative evaluation therefore uses the dataset-provided five-fold partition over all 1,109 labeled claims; the final model trained on the full labeled set produces predictions for the 300 test claims. Every supervised component, nearest-claim memory, expansion source, threshold, and gate parameter is restricted to the training portion of its fold. Table 1 reports the split counts.

Table 1. SciFact claim splits used in the experiments.

Split	Claims	SUPPORT	CONTRADICT	NEI
Train	809	332	173	304
Development	300	124	64	112
Labeled total	1109	456	237	416
Unlabeled test	300	-	-	-

B. Closed-loop architecture. Figure 1 summarizes the pipeline. Pass 1 maps the claim to a stemmed sparse query and retrieves a ranked abstract list. A sentence selector compresses the top-d abstracts into five candidate sentences. The verifier estimates whether the claim has sufficient evidence and then assigns SUPPORT or CONTRADICT when evidence is present. The corrective gate accepts the first result or routes the claim to a query rewriter. Pass 2 retrieves with the expanded query, fuses the original and rewritten rankings, selects a new evidence set, and invokes the same verifier. Retrieval routing and label replacement are separate decisions: a claim can use second-pass sentences as its rationale while retaining the first label when the second prediction lacks a calibrated confidence advantage.



Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 1. Corrective retrieval loop with verifier-controlled routing, query reformulation, rank fusion, and conservative evidence accumulation.

C. First-pass document retrieval. The study evaluates four first-pass retrievers. Word BM25 tokenizes lowercase alphanumeric terms, removes English stop words while retaining negation, repeats the title twice, and uses k1 = 1.5 and b = 0.75. Stem-BM25 applies the Snowball English stemmer to the same representation. Character TF-IDF uses word-boundary character 3–5 grams, sublinear term frequency, L2 normalization, and a 120,000-feature limit. RRF-1P fuses the word, stem, and character rankings with reciprocal-rank fusion. For a document d and query q, BM25 sums term contributions based on term frequency, document length, average document length, and inverse document frequency [26]. RRF assigns document d the score Σr 1/(60 + rankr(d)) over the contributing rankings [27]. The fusion depth is 200 documents per ranking.

D. Evidence sentence selection. For each retrieved abstract, every sentence is represented by a word 1–2 gram TF-IDF vector. The selector evaluates the top d abstracts and scores sentence s in document rank r as 0.72 cos(q,s) + 0.18 cos(q,title) + 0.10/(r+1). It retains the five highest-scoring sentences across the candidate documents. This design makes sentence evidence sensitive to both local lexical agreement and document-level topic agreement. The selected context supplies the verifier text, the sentence-level evidence output, and the measured verifier token count. Evidence evaluation includes hit rates at one, three, and five sentences, union recall, precision at five, and complete-rationale recovery.

E. Verifier agent. The verifier decomposes the three-class decision into evidence availability and evidence stance. A balanced logistic classifier estimates p(E=1|q,S), where S is the selected context. Its sparse block concatenates claim word 1–3 grams, claim character 3–5 grams, context word 1–2 grams, prefixed claim–evidence relation tokens, and standardized dense features. Dense features include overlap, Jaccard similarity, negation counts, numeric mismatch, top-sentence similarity, retrieval score, and score margin. If evidence probability exceeds the calibrated threshold, a stance layer assigns SUPPORT or CONTRADICT; otherwise, the output is NEI. This separation parallels perception systems that distinguish fusion from interpretation [54]. Visual explanation cards similarly separate localization, classification, confidence, and retrieved precedents [55], while attention-enhanced driving detection evaluates feature extraction before decision use [56]. The textual verifier applies the same modular principle to retrieval, rationale selection, and stance.

F. Contrastive stance memory. Sparse claim–context features alone do not reliably distinguish a supported proposition from a minimally edited contradiction. The verifier therefore contains a training-fold contrastive memory. Positive training claims are encoded by concatenated word 1–3 gram and character 3–5 gram TF-IDF vectors. For a new claim, the nearest positive training claim supplies a provisional stance when cosine similarity is at least 0.20. A deterministic relation check reverses that stance when the two claims differ in explicit negation or in curated opposition sets such as increase/decrease, present/absent, before/after, and male/female. Below the similarity threshold, the evidence-based stance classifier supplies the label. The memory stores only training-fold claims and never accesses a held-out claim's annotation.

G. Evidence-linked query reformulation. The rewriter encodes training-fold claims with word 1–2 gram TF-IDF and retrieves the three nearest training claims. Neighbors below cosine 0.18 are ignored. For each retained neighbor, terms from its gold evidence titles receive weight 2·sim·idf, and terms from the first two evidence sentences receive weight 0.25·sim·idf. Terms already present in the input claim are removed. The twelve highest-weight terms are appended to the original claim, producing a transparent expansion rather than an unconstrained generated statement. A second stemmed-BM25 search runs on the expanded query, and RRF fuses the original and second rankings. This mechanism follows the relevance-feedback principle [29], [30] while enforcing fold-safe evidence provenance.

H. Gate calibration and conservative correction. Each training fold is split into an 80% fit subset and a 20% stratified calibration subset. The binary and stance models are fitted on the fit subset, and the calibration subset selects the positive threshold and gate. The positive threshold is chosen from 0.20 to 0.65 in increments of 0.025 by Macro-F1. The trigger grid evaluates minimum positive probability, maximum accepted confidence, minimum sentence similarity, and document-score margin. Its objective is Macro-F1 − 0.015 × second-pass rate. After the second pass, a separate confidence-gain threshold decides whether its label replaces the first label. The acceptance grid ranges from −0.10 to 0.20. Finally, monotonic evidence preservation rejects any transition from first-pass SUPPORT or CONTRADICT to second-pass NEI. The rationale is evidential: failure to recover evidence under an alternative query does not invalidate evidence already retrieved. The full models are then refitted on the entire training fold and evaluated once on the held-out fold.

I. Evaluation and cost accounting. Document retrieval is measured by Recall@1, 3, 5, 10, and 20 and mean reciprocal rank over the 693 positive claims. Verification uses accuracy, Macro-F1, per-class precision/recall/F1, and ten-bin expected calibration error. The joint score counts NEI when its label is correct; a positive claim also requires at least one complete gold rationale among the five selected sentences. This objective is stricter than accuracy alone, as constrained budgeting must evaluate forecasts under feasible allocation rules [57], credit-risk modeling must connect classification to consequences [58], and uncertainty-aware market forecasting must preserve the evidence behind directional predictions [59]. Cost is the alphanumeric-token count in retrieved top-d abstracts plus claim–sentence tokens processed by the

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

verifier. Always-2P pays both pass costs for every claim; Corrective-RAG pays the second cost only when routed. Paired bootstrap intervals use 2,000 claim-level resamples, following NLP significance-testing guidance [51].

J. Implementation settings. All random operations use seed 42. Document, sentence, claim, and context indexes are fitted once within their permitted data scope and stored as sparse matrices. The experiment fixes five selected sentences, an RRF constant of 60, and the hyperparameters in Table 2. The median fold-calibrated positive threshold used for the final test model is 0.55. The recorded single-thread run completed in 109.3 seconds, including five-fold evaluation, depth sweeps, bootstrap analysis, figure generation, and final test prediction.

Table 2. Fixed model components and calibrated design settings.

Component	Setting
Word BM25	k1=1.5; b=0.75; title repeated twice
Stemmed BM25	Snowball English stemmer; k1=1.5; b=0.75
Character retrieval	char_wb 3-5 grams; max_features=120000
Sentence selector	0.72 sentence cosine + 0.18 title cosine + 0.10 reciprocal rank; top 5
Query rewriter	3 nearest training claims; min similarity 0.18; 12 expansion terms
Rank fusion	RRF k=60; top 200 from each pass
Verifier	balanced logistic regression; C=2; claim/context/relation features
Contrastive memory	nearest positive training claim; similarity threshold 0.20
Gate	fold-internal grid over trigger thresholds and confidence-gain acceptance; monotonic evidence preservation; objective Macro-F1 - 0.015 x second-pass rate

B. RESULTS AND DISCUSSION

A. Document retrieval. Table 3 shows that stemming is the strongest single sparse normalization for the early ranks. Word BM25 obtains Recall@5 = 0.885, while Stem-BM25 reaches 0.903 and raises MRR from 0.793 to 0.815. Character TF-IDF reaches the highest Recall@20 among one-pass methods (0.968) but has a lower MRR because character similarity moves some relevant abstracts down the earliest ranks. First-pass fusion improves Recall@3 and Recall@5 relative to either word BM25 or stemmed BM25 alone, yet its MRR remains below the stemmed ranking. The result indicates that heterogeneous recall gains do not guarantee a better first document.

Table 3. Out-of-fold document retrieval over positive SciFact claims.

Method	R@1	R@3	R@5	R@10	R@20	MRR	N
BM25-1P	0.726	0.837	0.885	0.929	0.944	0.793	693
Stem-BM25-1P	0.747	0.861	0.903	0.939	0.957	0.815	693
Char-TFIDF-1P	0.657	0.848	0.908	0.944	0.968	0.764	693
RRF-1P	0.739	0.877	0.909	0.944	0.957	0.813	693
Always-2P	0.797	0.893	0.928	0.960	0.977	0.855	693
Corrective-RAG	0.756	0.869	0.913	0.947	0.964	0.823	693

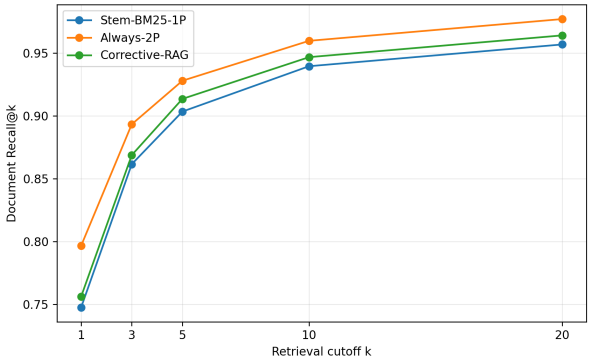


Figure 2. Document recall as the cutoff increases for the principal one-pass and iterative systems.

The evidence-linked rewrite produces a clear retrieval gain. Always-2P reaches Recall@1 = 0.797, Recall@5 = 0.928, and Recall@20 = 0.977. Relative to Stem-BM25, these are increases of 4.91, 2.45, and 2.02 percentage points. Corrective-RAG lies between one-pass and always-on retrieval because it uses the fused list only for gated claims. Its Recall@5 of 0.913 exceeds the one-pass value while preserving the original ranking for the majority of claims. This pattern confirms that the rewrite contains useful evidence vocabulary rather than random lexical expansion.

B. Sentence evidence recovery. Document recovery translates into sentence-level evidence improvement (Table 4). Stem-BM25 obtains an evidence hit at five of 0.714 and complete-rationale recovery of 0.693. Always-2P raises these values to 0.750 and 0.732. Corrective-RAG reaches 0.727 and 0.704, respectively. Its sentence union recall at five is 0.586, compared with 0.577 for the single-pass stemmed baseline. The difference is concentrated in claims whose first ranking lacks a gold abstract; when the document is already present, both methods often select the same decisive sentence.

Table 4. Evidence sentence selection and complete-rationale recovery.

Method	Hit@1	Hit@3	Hit@5	Sentence R@5	Sentence P@5	Complete rationale@5
BM25-1P	0.364	0.600	0.710	0.568	0.195	0.690
Stem-BM25-1P	0.372	0.609	0.714	0.577	0.198	0.693
Char-TFIDF-1P	0.361	0.596	0.704	0.571	0.196	0.681
RRF-1P	0.364	0.610	0.716	0.578	0.198	0.696
Always-2P	0.392	0.642	0.750	0.604	0.206	0.732
Corrective-RAG	0.374	0.619	0.727	0.586	0.201	0.704

C. End-to-end verification. Table 5 reports the principal verification results. Stem-BM25-1P is the strongest pure one-pass system with 68.3% accuracy and 0.654 Macro-F1.

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Character retrieval and first-pass RRF improve some recall cutoffs but reduce verification performance, demonstrating that a verifier trained around concise high-ranked evidence is sensitive to context composition. Always-2P produces the strongest document retrieval but falls to 63.7% accuracy and 0.601 Macro-F1. The decline is not a contradiction in the metrics: additional relevant abstracts improve recall, while additional semantically adjacent sentences can weaken the discriminative boundary between evidence-bearing and NEI contexts.

Table 5. End-to-end label, joint evidence, and calibration results.

Method	Accuracy	Macro-F1	Joint score	ECE
BM25-1P	0.674	0.648	0.583	0.066
Stem-BM25-1P	0.683	0.654	0.591	0.055
Char-TFIDF-1P	0.629	0.587	0.557	0.075
RRF-1P	0.637	0.599	0.569	0.064
Always-2P	0.637	0.601	0.579	0.067
Corrective-RAG	0.684	0.656	0.597	0.058

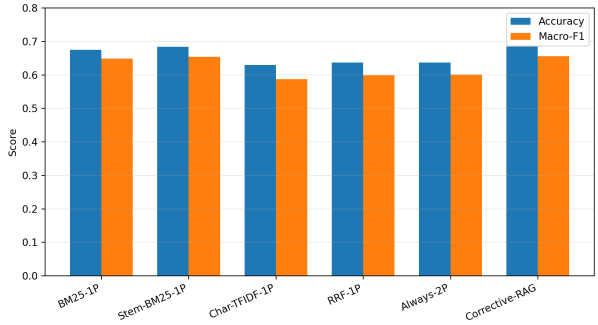


Figure 3. Accuracy and Macro-F1 for all evaluated retrieval policies.

Corrective-RAG attains 68.4% accuracy and 0.656 Macro-F1 while obtaining the highest joint score, 0.597. Relative to Always-2P, the improvements are 4.78 accuracy points and 5.51 Macro-F1 points. Relative to Stem-BM25-1P, the label difference is small (+0.09 percentage points in accuracy and +0.14 percentage points in Macro-F1), while the joint score rises by 0.63 points. The paired bootstrap interval for the Macro-F1 difference against Stem-BM25 spans zero, whereas the intervals against Always-2P and RRF-1P are strictly positive. The evidence-conditioned objective therefore captures the main benefit: the loop improves the chance that a correct label is accompanied by a complete rationale without paying the classification penalty of unconditional iteration.

D. Class behavior and conservative evidence accumulation. Table 6 separates the three labels. The one-pass stemmed model is especially strong on NEI, while SUPPORT and CONTRADICT remain more difficult because both require evidence presence plus stance discrimination. Corrective-RAG preserves the high NEI recall and changes only a small set of labels after the second verifier. The monotonic evidence rule is essential here. Without that rule, most harmful second-pass changes converted a correct positive label to NEI after query drift. Under the final policy, an alternate query is allowed to add or replace evidence, promote a weak NEI decision, or support a confident stance change, but it cannot erase previously observed evidence. This is an evidence aggregation principle rather than a class-specific heuristic: absence under one query is not proof of absence in the corpus.

Table 6. Per-class precision, recall, and F1.

Method	Label	Precision	Recall	F1	Support
BM25-1P	SUPPORT	0.683	0.634	0.658	456
BM25-1P	CONTRADICT	0.518	0.498	0.508	237
BM25-1P	NEI	0.745	0.820	0.780	416
Stem-BM25-1P	SUPPORT	0.704	0.616	0.657	456
Stem-BM25-1P	CONTRADICT	0.518	0.494	0.505	237
Stem-BM25-1P	NEI	0.744	0.865	0.800	416
Char-TFIDF-1P	SUPPORT	0.716	0.487	0.580	456
Char-TFIDF-1P	CONTRADICT	0.517	0.384	0.441	237
Char-TFIDF-1P	NEI	0.618	0.925	0.741	416
RRF-1P	SUPPORT	0.706	0.504	0.588	456
RRF-1P	CONTRADICT	0.508	0.418	0.458	237
RRF-1P	NEI	0.641	0.906	0.751	416
Always-2P	SUPPORT	0.709	0.487	0.577	456
Always-2P	CONTRADICT	0.528	0.430	0.474	237
Always-2P	NEI	0.633	0.918	0.750	416
Corrective-RAG	SUPPORT	0.704	0.616	0.657	456
Corrective-RAG	CONTRADICT	0.520	0.498	0.509	237
Corrective-RAG	NEI	0.745	0.865	0.801	416

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

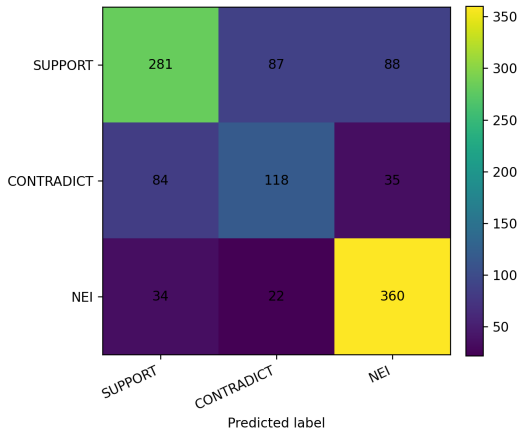


Figure 4. Out-of-fold confusion matrix for Corrective-RAG.

The confusion matrix shows that CONTRADICT remains the principal bottleneck. Contradictory claims frequently differ from supported claims by a polarity word, quantity, population qualifier, or direction of effect. Character and word overlap remain high in these pairs, so retrieval succeeds more often than stance classification. The contrastive memory improves this boundary by transferring a nearby training stance and reversing it under explicit negation or opposition, but it does not encode every biomedical antonym or numeric incompatibility. The residual errors therefore arise after successful retrieval as often as they arise from missing documents.

E. Retrieval cost. Table 7 quantifies the cost of each policy. Stem-BM25-1P processes 1,327.2 tokens per claim under the defined retrieval-plus-verifier accounting. Always-2P nearly doubles that amount to 2,676.2. Corrective-RAG triggers the second pass for 17.0% of claims and processes 1,558.5 tokens on average. This is 41.8% fewer tokens than Always-2P and 17.4% more than the one-pass stemmed baseline. The median remains close to the one-pass median because most claims terminate after pass 1; the p95 rises because routed claims pay both retrieval costs.

Table 7. Processed-token cost and second-pass utilization.

Method	Second-pass rate	Mean tokens	Median tokens	P95 tokens
BM25-1P	0.000	1,337.0	1,254.0	1,946.0
Stem-BM25-1P	0.000	1,327.2	1,237.0	1,927.8
Char-TFIDF-1P	0.000	1,233.3	1,172.0	1,787.0
RRF-1P	0.000	1,309.9	1,230.0	1,898.4
Always-2P	1.000	2,676.2	2,492.0	3,923.6
Corrective-RAG	0.170	1,558.5	1,306.0	3,096.6

The cost result distinguishes selective correction from simply reducing retrieval depth. Lower depth saves tokens for every claim but limits evidence recovery. Gating retains a stronger default depth and concentrates extra work on weak-similarity or low-confidence cases. This allocation follows selective classification [47]: computation rises with expected risk rather than uniformly. Comparable pressures arise when market forecasts fuse news uncertainty with macro variables [59] and financial-service models combine customer insight with risk analysis [60]. The present experiment isolates this principle at retrieval time by reporting evidence quality and workload together.

F. Ablation analysis. Table 8 compares complete operating points rather than removing individual feature coordinates. Word BM25 loses the stemming benefit and the closed loop. Stem-BM25 isolates the strongest single pass. RRF-1P retains multi-retriever fusion but removes reformulation and gating. Always-2P keeps reformulation but removes selectivity. The large gap between Always-2P and Corrective-RAG shows that the gate is not merely a cost optimization; it protects the verifier from context shift. The small difference between Stem-BM25 and Corrective-RAG on label metrics, combined with the positive joint-score difference, shows that the loop acts primarily as evidence repair rather than as a wholesale replacement for the initial classifier.

Table 8. Pipeline ablations and operating-point comparisons.

Configuration	Removed or changed	Accuracy	Macro-F1	Joint score	Δ Macro-F1 vs. CRAG
BM25-1P	stemming, iterative retrieval, and corrective gate	0.674	0.648	0.583	-0.007
Stem-BM25-1P	query rewriting and corrective gate	0.683	0.654	0.591	-0.001
RRF-1P	query rewriting and corrective gate; retain first-pass fusion	0.637	0.599	0.569	-0.056
Always-2P	corrective gate; force all claims through pass 2	0.637	0.601	0.579	-0.055
Corrective-RAG	none	0.684	0.656	0.597	0.000

G. Retrieval depth–accuracy–cost trade-off. Table 9 and Figures 5–6 vary the abstract depth from one to twenty while keeping five selected sentences. Stem-BM25 reaches its best label accuracy near depth five and changes little after depth three. Additional documents increase the scanned-token cost but do not reliably improve the five-sentence context, because lower-ranked sentences compete with already strong

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

evidence. Always-2P shows the same saturation at twice the retrieval cost. Corrective-RAG follows the one-pass accuracy curve closely while remaining well below the unconditional cost curve.

Table 9. Retrieval depth, verification quality, and processed-token cost.

Depth	Method	Accuracy	Macro-F1	Joint score	Second-pass rate	Mean tokens
1	Always-2P	0.628	0.590	0.576	1.000	848.7
1	Corrective-RAG	0.670	0.641	0.590	0.213	515.1
1	Stem-BM25-1P	0.668	0.639	0.583	0.000	423.3
3	Always-2P	0.642	0.608	0.591	1.000	1,763.7
3	Corrective-RAG	0.680	0.652	0.602	0.183	1,039.3
3	Stem-BM25-1P	0.679	0.651	0.594	0.000	876.5
5	Always-2P	0.637	0.601	0.579	1.000	2,676.2
5	Corrective-RAG	0.684	0.656	0.597	0.170	1,558.6
5	Stem-BM25-1P	0.683	0.654	0.591	0.000	1,327.2
10	Always-2P	0.638	0.602	0.582	1.000	4,984.0
10	Corrective-RAG	0.683	0.653	0.598	0.172	2,920.4
10	Stem-BM25-1P	0.682	0.652	0.593	0.000	2,477.9
20	Always-2P	0.634	0.597	0.576	1.000	9,552.0
20	Corrective-RAG	0.683	0.655	0.598	0.170	5,582.1
20	Stem-BM25-1P	0.682	0.653	0.593	0.000	4,751.5

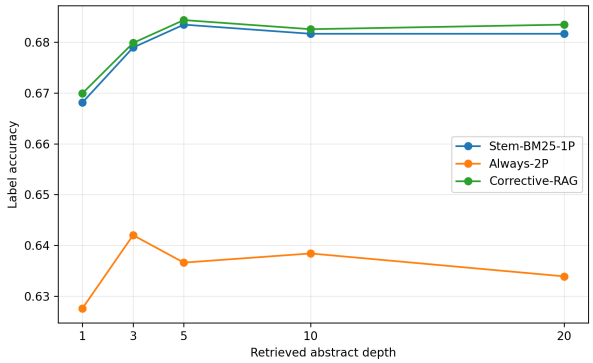


Figure 5. Retrieval depth versus label accuracy.

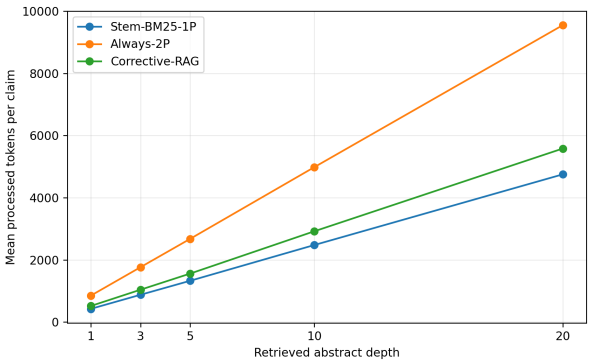


Figure 6. Retrieval depth versus processed-token cost.

Depth three produces the strongest joint score for the corrective policy in the sweep, while depth five produces the strongest one-pass accuracy and remains the primary operating point. This difference is expected: the joint metric rewards complete evidence, and a rewritten top-three list can concentrate more relevant abstracts than a broader first-pass list. The depth-five setting was retained for the main comparison because it gives the verifier a consistent candidate budget and matches the retrieval analysis at Recall@5.

H. Query-rewrite recovery and drift. Table 10 analyzes every positive claim at document cutoff five. The first stemmed pass retrieves a gold document for 626 of 693 positive claims. The fused second pass retrieves one for 643, recovering 23 initial misses and drifting on 6 initial hits. The net gain is 17 claims. Expansions contain 7.64 terms on average, below the maximum of twelve because some claims have no neighbor above the similarity threshold. The asymmetry between recovery and drift validates the use of training-fold evidence titles as expansion anchors, while the nonzero drift count justifies retaining the original ranking in RRF and preserving first-pass evidence.

Table 10. Query-reformulation recovery, drift, and routing diagnostics.

Positive claims	Initial top-5 hits	Second top-5 hits	Recovered	Drifted	Gate trigger rate	Mean expansion terms
693	626	643	23	6	0.170	7.639

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

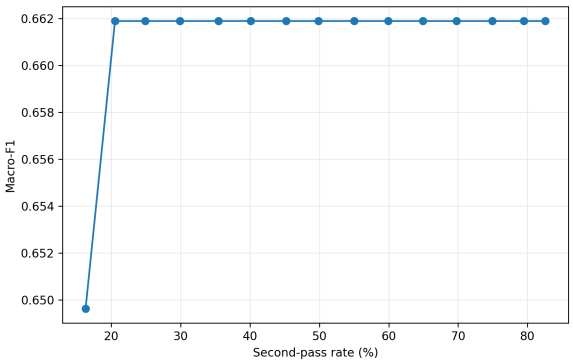


Figure 7. Calibration-set Macro-F1 across second-pass utilization levels.

The gate curve shows that additional second-pass coverage does not monotonically improve Macro-F1. Low-coverage policies miss some recoverable cases, but high-coverage policies expose more stable claims to query drift. The selected fold-specific policies occupy the low-to-moderate coverage region. Their median trigger settings require positive probability of at least 0.50 and label confidence below 0.55 or weak sentence evidence. Label replacement is stricter than retrieval routing, so most routed claims receive improved evidence without changing their class.

I. Calibration and statistical interpretation. Corrective-RAG has an expected calibration error of 0.058, compared with 0.055 for Stem-BM25 and 0.067 for Always-2P. Figure 8 shows the empirical reliability bins. The confidence score combines the binary evidence probability with stance confidence; it is therefore a routing score rather than a calibrated posterior over three mutually exclusive labels. Even so, its ordering is useful for selective computation. Bootstrap analysis gives a Corrective-RAG minus Always-2P Macro-F1 difference of 0.055, with a 95% interval from 0.037 to 0.074. Against Stem-BM25, the interval crosses zero. The data support a strong conclusion against unconditional iteration and a narrower conclusion against the best single pass: corrective routing improves evidence-conditioned verification without a statistically resolved label-metric difference.

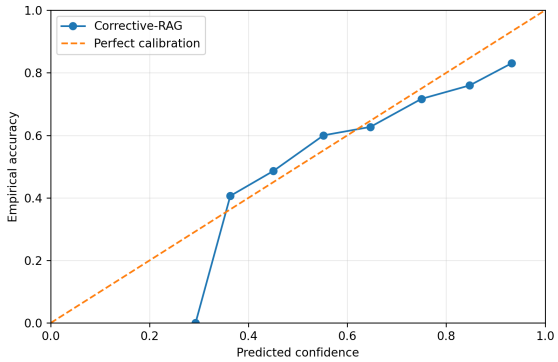


Figure 8. Reliability diagram for the Corrective-RAG routing confidence.

J. Overall interpretation. The experiments separate three effects that are often collapsed into one RAG score. Retrieval quality asks whether the corpus candidate set contains a gold abstract. Evidence quality asks whether the compressed context contains a complete rationale. Label quality asks whether the verifier interprets that evidence correctly. Query reformulation improves the first two effects. It does not automatically improve the third because the verifier receives a different context distribution. The corrective design succeeds by treating the second pass as an evidence proposal that must be verified, not as an authoritative replacement. This distinction explains why the method can outperform Always-2P while using fewer tokens, and why its main gain appears in the joint evidence-conditioned score rather than in a large standalone accuracy increase.

Limitations

The experiments are confined to SciFact, whose claims and evidence are predominantly biomedical and whose corpus contains abstracts rather than full articles. The results establish behavior for compact scientific evidence retrieval, not the same operating point for guidelines, long reports, tables, figures, or cross-document synthesis. Multimodal extension would require separate localization, fusion, and display validation. Driving perception fusion [54], retrieved traffic-sign explanation cards [55], and attention-enhanced image detection [56] illustrate why these stages should not collapse into one retrieval score. The rewriter also depends on labeled training-fold evidence; deployment in a new domain therefore requires an annotated memory, a high-precision evidence store, or a separately validated expansion model.

The verifier is deliberately lightweight and CPU-reproducible. Sparse lexical features, logistic classification, and a curated contrastive relation check expose retrieval-control effects without requiring external model weights, but they leave substantial room for stronger scientific entailment modeling. CONTRADICT performance remains lower than SUPPORT and NEI because many contradictions depend on quantities, study populations, temporal qualifiers, and domain-specific antonyms not represented by the opposition lists. A transformer verifier can replace the current classifier while retaining the same gate, rewrite, fusion, and evidence-preservation protocol.

Quantitative evaluation uses the provided five-fold partition because public labels are absent for the 300 test claims. Cross-validation gives an out-of-fold prediction for every labeled claim and prevents a claim's evidence from entering its own rewrite memory, but it does not reproduce a hidden leaderboard evaluation. The final test predictions are included as system outputs rather than scored results. The reported bootstrap intervals quantify claim-sampling variation within SciFact and do not capture variation across scientific domains or annotation policies.

The cost metric counts retrieved abstract tokens and verifier claim–sentence tokens. It is a workload proxy under identical

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

local code, not a hosted-model dollar estimate. Index construction and training are excluded, and runtime depends on the environment. The study evaluates automated labels and rationales rather than scientists' decisions; it does not measure review speed, trust calibration, or downstream judgment. Interface design matters: risk-dashboard research shows that visual hierarchy and explainability shape interpretation [53], bilingual computer-science education shows that language choice can affect engagement and learning [61], and enterprise marketing research illustrates how predictions ultimately become human decisions [62]. Multilingual and expert-facing presentation therefore require dedicated human-subject evaluation.

C. CONCLUSION

This study evaluated a complete corrective RAG loop for scientific fact checking on SciFact. Evidence-linked query reformulation recovered documents and rationales missed by a strong stemmed-BM25 pass, but applying the second pass to every claim reduced verification quality and doubled the token workload. The verifier-controlled policy resolved that conflict by routing only low-confidence or weak-evidence claims, separating retrieval from label replacement, and preserving evidence already found by the first pass. Under five-fold evaluation, the resulting system used the second pass for a minority of claims, retained the label quality of the best single-pass baseline, improved the evidence-conditioned joint score, and substantially outperformed always-on iteration at lower cost. The central design lesson is that iterative retrieval requires an explicit evidence acceptance policy. Query reformulation supplies candidates; the verifier determines whether those candidates add evidence, justify a label change, or should be ignored.

D. REFERENCES

[1] Cohan, and H. Hajishirzi, "Fact or Fiction: Verifying Scientific Claims," in Proc. 2020 Conf. Empirical Methods in Natural Language Processing, 2020, pp. 7534–7550.

[2] S. Zhou, Z. Li, and E. Wang, "Evidence-Grounded RAG for Tokenized Trade Receivable Disclosure QA under U.S. Capital Market Standards," J. Adv. Comput. Syst., vol. 3, no. 7, pp. 41–57, Jul. 2023, doi: 10.69987/JACS.2023.30704.

[3] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A Large-Scale Dataset for Fact Extraction and VERification," in Proc. 2018 Conf. North American Chapter of the Association for Computational Linguistics, 2018, pp. 809–819.

[4] S. Zhou, Z. Li, and E. Wang, "Long-Document RAG for Contractual and Insurance Clause Analysis in Receivables RWA Structures," J. Adv. Comput. Syst., vol. 4, no. 8, pp. 88–104, Aug. 2024, doi: 10.69987/JACS.2024.40810.

[5] Y. Li, "Findable then Explainable: Retrieval–Summary Integration for Code Intelligence on a Lightweight CodeSearchNet Subset," J. Adv. Comput. Syst., vol. 4, no. 7, pp. 65–82, Jul. 2024, doi: 10.69987/JACS.2024.40706.

[6] J. Vladika and F. Matthes, "Scientific Fact-Checking: A Survey of Resources and Approaches," in Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 6215–6230.

[7] Wadden and K. Lo, "Overview and Insights from the SCIVER Shared Task on Scientific Claim Verification," in Proc. Second Workshop on Scholarly Document Processing, 2021, pp. 124–129.

[8] R. Pradeep, X. Ma, R. Nogueira, and J. Lin, "Scientific Claim Verification with VerT5erini," in Proc. 12th International Workshop on Health Text Mining and Information Analysis, 2021, pp. 94–103.

[9] D. Wadden, K. Lo, L. L. Wang, A. Cohan, I. Beltagy, and H. Hajishirzi, "MultiVerS: Improving Scientific Claim Verification with Weak Supervision and Full-Document Context," in Findings of the Association for Computational Linguistics: NAACL 2022, 2022, pp. 61–76.

[10] Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," in Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th International Joint Conf. Natural Language Processing, 2019, pp. 3615–3620.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics, 2019, pp. 4171–4186.

[12] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," in Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th International Joint Conf. Natural Language Processing, 2019, pp. 3982–3992.

[13] Z. Zhong, M. Zheng, H. Mai, J. Zhao, and X. Liu, "Cancer Image Classification Based on DenseNet Model," J. Phys.: Conf. Ser., vol. 1651, no. 1, Art. no. 012143, Nov. 2020, doi: 10.1088/1742-6596/1651/1/012143.

[14] J. Chen, J. Xiong, Y. Wang, Q. Xin, and H. Zhou, "Implementation of an AI-Based MRD Evaluation and Prediction Model for Multiple Myeloma," Front. Comput. Intell. Syst., vol. 6, no. 3, pp. 127–131, Jan. 2024, doi: 10.54097/zJ4MnbWW.

[15] Z. S. Zhong, R. Ma, and H. Zhao, "Human-Uncertainty Distillation for Calibrated Vision Models on CIFAR-10H," J. Adv. Comput. Syst., vol. 3, no. 2, pp. 77–89, Feb. 2023, doi: 10.69987/JACS.2023.30206.

[16] Q. Xin, "Uncertainty-Aware Late Fusion for 3D Perception (Confidence Calibration + Fusion Rule

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Learning)," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 215–238, Apr. 2025, doi: 10.51903/jtie.v4i1.485.

[17] Z. S. Zhong and S. Ling, "Uncertainty Quantification of Spectral Estimator and MLE for Orthogonal Group Synchronization," arXiv:2408.05944, Aug. 2024.

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

[19] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," in Proc. 37th International Conf. Machine Learning, 2020, pp. 3929–3938.

[20] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense Passage Retrieval for Open-Domain Question Answering," in Proc. 2020 Conf. Empirical Methods in Natural Language Processing, 2020, pp. 6769–6781.

[21] Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering," in Proc. 16th Conf. European Chapter of the Association for Computational Linguistics, 2021, pp. 874–880.

[22] J. Nie and D. Zheng, "Ambiguity-Aware HDFS Log Anomaly Detection with Retrieval-Augmented Failure Narratives and Selective Refusal," J. Adv. Comput. Syst., vol. 3, no. 1, pp. 66–80, Jan. 2023, doi: 10.69987/JACS.2023.30105.

[23] Q. Xin, "Explaining OpenStack Failure-Injection Log Anomalies with Retrieved Normal Prototypes," Emerg. Inf. Sci. Technol., vol. 6, no. 2, pp. 125–146, Nov. 2025, doi: 10.18196/eist.v6i2.31232.

[24] Zhang, "LLM-Driven CI Failure Diagnosis and Automated Repair: From GitHub Actions Logs to Patch Recommendation," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 190–214, Apr. 2025, doi: 10.51903/jtie.v4i1.484.

[25] Y. He, Y. Pan, Y. Wang, S. Du, and Q. Xin, "Intelligent Fault Analysis with AIOps Technology," J. Theory Pract. Eng. Sci., vol. 4, no. 1, pp. 94–100, Feb. 2024, doi: 10.53469/jtpes.2024.04(01).13.

[26] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009.

[27] V. Cormack, C. L. A. Clarke, and S. Büttcher, "Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods," in Proc. 32nd International ACM SIGIR Conf. Research and Development in Information Retrieval, 2009, pp. 758–759.

[28] G. Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," Communications of the ACM, vol. 18, no. 11, pp. 613–620, 1975.

[29] J. Rocchio, "Relevance Feedback in Information Retrieval," in The SMART Retrieval System—Experiments in Automatic Document Processing, G. Salton, Ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 1971, pp. 313–323.

[30] V. Lavrenko and W. B. Croft, "Relevance-Based Language Models," in Proc. 24th Annual International ACM SIGIR Conf. Research and Development in Information Retrieval, 2001, pp. 120–127.

[31] Z. S. Zhong and S. Ling, "Improved Theoretical Guarantee for Rank Aggregation via Spectral Method," Inf. Inference: J. IMA, vol. 13, no. 3, Art. no. iaae020, Sep. 2024, doi: 10.1093/imaiai/iaae020.

[32] L. Wang, N. Yang, and F. Wei, "Query2doc: Query Expansion with Large Language Models," in Proc. 2023 Conf. Empirical Methods in Natural Language Processing, 2023, pp. 9414–9423.

[33] Z. S. Zhong, X. Pan, and Q. Lei, "Bridging Domains with Approximately Shared Features," in Proc. 28th Int. Conf. Artificial Intelligence and Statistics, PMLR, vol. 258, pp. 559–567, 2025.

[34] X. Ma, Y. Gong, P. He, H. Zhao, and N. Duan, "Query Rewriting in Retrieval-Augmented Large Language Models," in Proc. 2023 Conf. Empirical Methods in Natural Language Processing, 2023, pp. 5303–5315.

[35] Z. Jiang, F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, and G. Neubig, "Active Retrieval Augmented Generation," in Proc. 2023 Conf. Empirical Methods in Natural Language Processing, 2023, pp. 7969–7992.

[36] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions," in Proc. 61st Annual Meeting of the Association for Computational Linguistics, 2023, pp. 10014–10037.

[37] L. Gao, X. Ma, J. Lin, and J. Callan, "Precise Zero-Shot Dense Retrieval without Relevance Labels," in Proc. 61st Annual Meeting of the Association for Computational Linguistics, 2023, pp. 1762–1777.

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

[38] Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, "Intelligent Classification and Personalized Recommendation of E-Commerce Products Based on Machine Learning," Appl. Comput. Eng., vol. 64, no. 1, pp. 143–149, May 2024, doi: 10.54254/2755-2721/64/20241365.

[39] H. Zhang, "Risk-Aware Budget-Constrained Auto-Bidding under First-Price RTB: A Distributional Constrained Deep Reinforcement Learning Framework," J. Adv. Comput. Syst., vol. 4, no. 6, pp. 30–47, Jun. 2024, doi: 10.69987/JACS.2024.40603.

[40] H. Wang, Y. Ren, and X. Chang, "Layout-Aware Progressive PDF Rendering: AI Prioritization of PDF Slices to Reduce Time-to-Functional-First-Frame on FUNSD," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 425–446, Aug. 2025, doi: 10.51903/jtie.v4i2.523.

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

[41] H. Zhang, "Counterfactual Learning-to-Rank for Ads: Off-Policy Evaluation on the Open Bandit Dataset," J. Adv. Comput. Syst., vol. 5, no. 12, pp. 1–11, Dec. 2025, doi: 10.69987/JACS.2025.51201.

[42] Q. Xin, "Hybrid Cloud Architecture for Efficient and Cost-Effective Large Language Model Deployment," J. Inf. Syst. Informatics, vol. 7, no. 3, pp. 2182–2195, Sep. 2025, doi: 10.51519/journalisi.v7i3.1170.

[43] B. Zhou, H. Wang, and X. Chang, "Distilling VMAF into an Edge-Deployable Quality Predictor: A Pilot Shot-Level Proxy with LLM-Ready Quality Tokens," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 447–463, Aug. 2025, doi: 10.51903/jtie.v4i2.522.

[44] Zhang, R. Ma, and P. Greg, "Digital-Twin Dispatching for Urban Mobility via Spatio-Temporal Transformers and Offline Reinforcement Learning," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 337–363, Aug. 2025, doi: 10.51903/jtie.v4i2.501.

[45] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in Proc. 34th International Conf. Machine Learning, 2017, pp. 1321–1330.

[46] Y. Geifman and R. El-Yaniv, "Selective Classification for Deep Neural Networks," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 4878–4887.

[47] H. Zhang, "DriftGuard: Multi-Signal Drift Early Warning and Safe Re-Training/Rollback for CTR/CVR Models," J. Adv. Comput. Syst., vol. 3, no. 7, pp. 24–40, Jul. 2023, doi: 10.69987/JACS.2023.30703.

[48] J. Nie and D. Zheng, "Noisy-Neighbor-Aware VM Degradation Risk Modeling with Unsupervised Residual Fusion," J. Adv. Comput. Syst., vol. 4, no. 4, pp. 112–123, Apr. 2024, doi: 10.69987/JACS.2024.40409.

[49] Q. Xin, Z. Xu, L. Guo, F. Zhao, and B. Wu, "IoT Traffic Classification and Anomaly Detection Method Based on Deep Autoencoders," Appl. Comput. Eng., vol. 69, no. 1, pp. 64–70, Jul. 2024, doi: 10.54254/2755-2721/69/20241511.

[50] R. Dror, G. Baumer, S. Shlomov, and R. Reichart, "The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing," in Proc. 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 1383–1392.

[51] B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, "Predictive Optimization of DDoS Attack Mitigation in Distributed Systems Using Machine Learning," Appl. Comput. Eng., vol. 64, no. 1, pp. 89–94, May 2024, doi: 10.54254/2755-2721/64/20241350.

[52] Z. Li, S. Zhou, and Z. Zhou, "Financial Risk Dashboard Design for Institutional RWA Investors: Visual Hierarchy, Chart Comprehension, and Explainability in FinChart-Bench," Int. J. Graph. Des., vol. 3, no. 1, pp. 196–210, May 2025, doi: 10.51903/ijgd.v3i1.3715.

[53] Y. Wang, S. Du, Q. Xin, Y. He, and W. Qian, "Autonomous Driving System Driven by Artificial Intelligence Perception Fusion," Acad. J. Sci. Technol., vol. 9, no. 2, pp. 193–198, Feb. 2024, doi: 10.54097/e0b9ak47.

[54] R. Ma, L. Zhang, and T. Song, "Computer-Vision-Informed Visual Explanation Cards for Autonomous-Driving Traffic-Sign Alerts: Localization, Classification, and Retrieved Evidence on GTSDB," Int. J. Graph. Des., vol. 3, no. 2, pp. 455–473, Oct. 2025, doi: 10.51903/ijgd.v3i2.3991.

[55] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of Autonomous Driving Image Detection Based on RFAConv and Triplet Attention," Appl. Comput. Eng., vol. 77, no. 1, pp. 210–217, Jul. 2024, doi: 10.54254/2755-2721/77/2024MA0067.

[56] Y. Lu, H. Zhou, and Y. Zhang, "A Constrained, Data-Driven Budgeting Framework Integrating Macro Demand Forecasting and Marketing Response Modeling," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 493–520, Dec. 2025, doi: 10.51903/jtie.v4i3.466.

[57] Q. Xin, R. Song, Z. Wang, Z. Xu, and F. Zhao, "Enhancing Bank Credit Risk Management Using the C5.0 Decision Tree Algorithm," J. Comput. Technol. Appl. Math., vol. 1, no. 4, pp. 100–107, Nov. 2024, doi: 10.5281/zenodo.14032041.

[58] H. Zhou and K. Zhang, "News-Based Uncertainty and Macro-Market Fusion for VIX Direction Forecasting: Evidence from 2015–2024 FRED Panel," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 487–501, Aug. 2025, doi: 10.51903/jtie.v4i2.540.

[59] T. Yang, Q. Xin, X. Zhan, S. Zhuang, and H. Li, "Enhancing Financial Services through Big Data and AI-Driven Customer Insights and Risk Analysis," J. Knowl. Learn. Sci. Technol., vol. 3, no. 3, pp. 53–62, Jul. 2024, doi: 10.60087/jklst.vol3.n3.p53-62.

[60] G. S. Raj, H. Zhang, V. Abhyankar, S. Mukherjee, E. Zhang, J. Williams, R. Halverson, and J. M. Patel, "Impact of Bilingual CS Education on Student Learning and Engagement in a Data Structures Course," in Proc. 19th Koli Calling Int. Conf. Comput. Educ. Res., 2019, pp. 1–10, doi: 10.1145/3364510.3364518.

[61] J. Wang, Q. Xin, Y. Liu, J. Wang, and T. Yang, "Predicting Enterprise Marketing Decision Making with Intelligent Data-Driven Approaches," J. Ind. Eng. Appl. Sci., vol. 2, no. 3, pp. 12–19, Jun. 2024, doi: 10.5281/zenodo.11357252.

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)