

Uncertainty-Gated Adaptive RAG for Scientific Claim Verification with Evidence-Calibrated Abstention

Lin Yang¹, Kevin Parker², David Huang³

¹Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA

²Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA

³School of Computer Science, Georgia Institute of Technology, Atlanta, GA, USA

lin.yang.cs@outlook.com

Abstract

Scientific claim verification must retrieve relevant evidence, assign SUPPORT, CONTRADICT, or NEI, and avoid confident decisions when evidence or the verifier is unreliable. We evaluate an uncertainty-gated adaptive RAG pipeline on 5,183 SciFact abstracts and 1,109 labeled claims through five official cross-validation folds. Four systems share one evidence selector and verifier: BM25 RAG, dense LSA RAG, always-on sparse-first hybrid RAG, and an adaptive system that begins with one BM25 document and two sentences, escalates uncertain claims to five hybrid documents and five sentences, arbitrates between fast and full posteriors, and applies calibrated abstention. The adaptive system answered 95.7% of claims with 54.4% selective accuracy and 47.9% macro-F1, versus 52.2% accuracy and 47.3% macro-F1 for BM25. Forced adaptive accuracy was 53.4%, while average context fell from 241.9 to 139.6 tokens, a 42.3% reduction. Abstention captured errors in 68.8% of rejected cases and reduced wrong answers by 8.7% relative to BM25. BM25 retained the strongest evidence recall, indicating that the adaptive benefit came mainly from calibrated routing and refusal rather than unconditional fusion. Contradiction reasoning remained the dominant error source.

Keywords : Scientific Claim Verification; Retrieval-Augmented Generation; Adaptive RAG; Uncertainty Calibration; Selective Prediction; Abstention; Scifact; Trustworthy AI

1. Introduction

Scientific claim verification combines document retrieval, sentence selection, and a three-way verdict. SciFact formalized this setting with expert-written claims, abstracts, evidence rationales, and SUPPORT, CONTRADICT, or NEI labels [1]. It extends FEVER's evidence-centered design [2], while PUBHEALTH [3], COVID-Fact [4], and HealthVer [5] show why biomedical verification must preserve the source behind a judgment. Specialized medical systems, including an AI-based multiple-myeloma MRD model [6], further illustrate that domain predictions remain meaningful only within their evidentiary and clinical context.

Retrieval is not a neutral preprocessing step: an irrelevant abstract can support a fluent but wrong verdict, and a missed negation can reverse the label. SciFact-Open makes this dependence explicit by expanding the search space [7]. Comparable evidence-linked pipelines appear in HDFS anomaly triage with failure narratives and selective refusal [8], the multi-stage VerT5erini system [9], and financial disclosure QA with citation constraints [10]. The scientific fact-checking survey in [11] similarly identifies evidence acquisition, rationale selection, veracity prediction, and explanation as coupled components. Accuracy, evidence recall, and retrieval cost must therefore be evaluated together.

RAG conditions a model on retrieved non-parametric evidence [12]. A fixed top-k context is simple, but it spends the same budget on easy and difficult claims and can

introduce distractors. Long-document contractual RAG [13], retrieval-summary integration for code intelligence [14], normal-prototype retrieval for OpenStack failures [15], and retrieval-guided CI diagnosis [16] all treat evidence selection as part of the decision process rather than a neutral preface. Scientific verification therefore needs a policy for when additional retrieval is warranted and when a low-confidence output should be withheld. Sparse and dense retrieval provide complementary inductive biases. BM25 emphasizes exact term overlap through probabilistic weighting [17], building on vector-space retrieval [18], whereas LSA projects term-document statistics into a lower-dimensional space [19]. RRF combines rankings without calibrated scores [20], and spectral rank-aggregation theory further shows that rank combination is itself an estimation problem [21]. DPR learns query-passage representations directly [22]. BERT [23] and SciBERT [24] supply general and scientific language encoders; approximately shared-feature methods address transfer across domains [25]; SPECTER [26] and Sentence-BERT [27] provide document and sentence representations. Because SciFact contains gene names, abbreviations, quantities, and direction-sensitive expressions, this study retains BM25 as the fast path and uses compact LSA as a deterministic dense comparator.

Confidence creates a separate failure mode. Maximum class probability is often treated as probability of correctness even though modern classifiers can be miscalibrated [28]. Human-uncertainty distillation offers

one route to calibrated supervision [29], and transformer studies confirm the need for post-hoc validation [30]. Uncertainty quantification for spectral estimators [31] and broader evidence on architecture and distribution shift [32] show that confidence behavior depends on regime; uncertainty-aware late fusion likewise motivates confidence-conditioned combination [33]. Selective classification converts uncertainty into a reject decision [34], while news-macro fusion provides another mixed-signal setting in which uncertainty affects action [35]. SelectiveNet [36] and selective question answering [37] operationalize related reject options. Out-of-distribution scores [38], deep ensembles [39], and Bayesian dropout [40] are alternatives, but the present system uses temperature scaling, a retrieval gate, posterior arbitration, and one final abstention threshold.

This work contributes a five-fold comparison of BM25, dense LSA, always-hybrid, and uncertainty-gated RAG on all 1,109 labeled SciFact claims; a shared pipeline for retrieval, sentence selection, three-way verification, calibration, routing, and abstention; joint reporting of accuracy, macro-F1, evidence recall, context tokens, and wrong-answer risk; and fold-level predictions, calibration records, retrieval traces, figures, tables, and deterministic code. Holding the verifier constant makes the benefit and cost of each retrieval policy directly comparable.

2. METHOD

Figure 1 summarizes the adaptive verification workflow. Every claim first receives a compact sparse evidence context. Calibrated uncertainty controls whether broader hybrid retrieval is executed, posterior arbitration decides whether the expanded context improves the decision, and a final threshold determines whether the system answers or abstains.

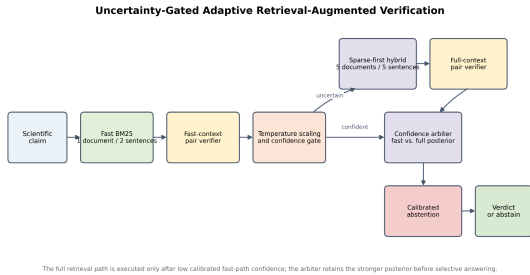


Fig. 1. Adaptive pipeline. Claims begin with a compact BM25 context; low calibrated confidence triggers broader hybrid retrieval and a second verification pass before selective answering.

2.1 Task Definition and Data

Let q denote a claim and $C=\{d_1, \dots, d_N\}$ the $N=5,183$ -abstract corpus. The label y is SUPPORT, CONTRADICT, or NEI. SUPPORT and CONTRADICT claims contain gold evidence documents and sentence rationales; an empty evidence field denotes NEI [1]. The 809 training and 300 development claims form 1,109 labeled examples: 456 SUPPORT, 237 CONTRADICT, and 416 NEI. The unlabeled test claims were excluded. Five official, disjoint evaluation folds cover every labeled claim once; each run trained on the other folds and evaluated on the held-out fold.

Within each fold, 20% of the fold-specific training set was selected by stratified sampling as a calibration subset; the remaining 80% trained the verifier. The calibration subset determined four temperatures, the fast-path uncertainty gate, the posterior-arbitration margin, and the final abstention threshold. No evaluation-fold label influenced these settings. Random operations used seed 2026 plus the fold index. Table I reports the data composition, and Fig. 2 shows the resulting class imbalance. CONTRADICT is the smallest class at 21.4%, making macro-F1 necessary in addition to accuracy.

TABLE I. SciFact corpus and labeled-claim composition.

Partition	Items	SUP	CON	NEI	Evidence claims	Gold docs	Gold sents
Corpus	5183	—	—	—	—	—	—
Official train	809	332	173	304	505	1.117	2.030
Official dev	300	124	64	112	188	1.112	1.947
All labeled	1109	456	237	416	693	1.115	2.007

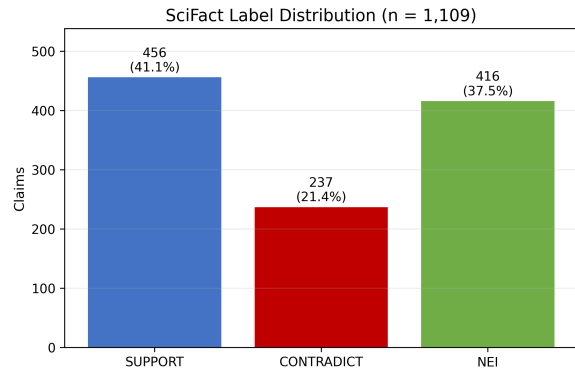


Fig. 2. Distribution of SUPPORT, CONTRADICT, and NEI labels across the 1,109 labeled claims.

2.2 Retrieval and Evidence Selection

The sparse retriever used BM25 with $k_1=1.5$ and $b=0.75$. Document text concatenated the title twice and the abstract once, which gives the title modest additional weight without a separate fielded index. English stop words were removed, and the vocabulary contained 35,617 terms. For query q and document d , the score was computed as

$$BM25(q,d) = \sum_{t \in q} IDF(t) \cdot f(t,d) \cdot (k_1 + 1) / \{f(t,d) + k_1 [1 - \frac{b + b|d|/avgdl}{b + b|d|/avgdl}]\}$$

where $f(t,d)$ is the term frequency and $avgdl$ is mean document length. A binary query vector ensured that repeated words in a short claim did not receive additional query-side weight. The BM25 baseline retrieved five documents. Candidate sentences from those documents were then ranked by claim-sentence TF-IDF cosine similarity, augmented by 0.08 times the normalized document score and a small reciprocal document-rank prior. The top five sentences formed the evidence context. The dense branch implemented latent-semantic retrieval rather than a pretrained neural encoder. Document titles and abstracts were represented with unigram-bigram TF-IDF, limited to 50,000 features with minimum document frequency two and sublinear term frequency. Truncated

singular-value decomposition produced 256-dimensional vectors, which were L2-normalized. The dense score was

$$s_{dense}(q,d)=\cos(z_q,z_d)(2)$$

where z_q and z_d are the normalized LSA projections. The dense baseline also retrieved five documents and used the identical sentence selector, thereby isolating document retrieval as the main difference from BM25. LSA is less expressive than DPR [22] or scientific transformer encoders [24], [26], and [27], but it supplies a deterministic dense semantic baseline with no remote weights, GPU requirement, or task-specific pretraining.

The always-on hybrid system fused the top 100 BM25 and dense rankings through weighted reciprocal rank fusion. Preliminary retrieval measurements on the labeled corpus showed that BM25 was markedly stronger than LSA, so the fusion placed 0.85 weight on BM25 and 0.15 on dense retrieval. With rank $r_m(d)$ under method m , the fusion score was

$$s_{hyb}(d)=0.85/[60+r_{BM25}(d)] + 0.15/[60+r_{dense}(d)] \quad (3)$$

The constant 60 follows the stabilizing form of RRF [20]. The always-on system retrieved five hybrid-ranked documents and selected five evidence sentences. The adaptive fast path used BM25 (by itself only) the highest-ranked document and its two highest-scoring sentences. Table II lists all retrieval and verification settings.

TABLE II. System configurations and fixed hyperparameters.

Component	Retriever / representation	Docs	Sents	Policy	Role
BM25 RAG	BM25	5	5	fixed	5-document sparse context
Dense RAG	TF-IDF + 256-d LSA	5	5	fixed	5-document dense context
Always-RAG	0.85/0.15 weighted RRF	5	5	fixed	5-document hybrid context
Adaptive fast	BM25	1	2	gated	default context
Adaptive full	weighted RRF	5	5	conditional	executed after low fast confidence
Verifier	word/character TF-IDF + 13 numeric features	—	—	shared	SGD logistic classifier

2.3 Evidence-Conditioned Pair Verifier

All four systems shared one verifier architecture within each fold. Training examples paired each claim with five context variants: the gold rationale context, BM25 context, dense context, always-hybrid context, and fast context. This mixed-context design exposed the classifier to both precise and noisy evidence while holding the downstream model constant across retrieval comparisons. For NEI training claims, the gold-context variant selected the three claim-most-similar sentences from the claim's cited

abstracts because no supporting or contradicting rationale exists.

The verifier represented 'claim ... evidence ...' with word unigram-bigram TF-IDF (40,000 features), character-within-word three- to five-gram TF-IDF (30,000 features), and thirteen standardized features summarizing lexical overlap, negation and numeric mismatch, context size, and retrieval and sentence scores. An L2-regularized SGD logistic classifier used balanced class weights, alpha 2×10^{-5} , 250 iterations, tolerance 10^{-4} , and averaged weights in scikit-learn [41]. High-capacity visual detectors such as RFACnv with triplet attention represent a different modeling regime [42]; the sparse verifier keeps evidence and confidence behavior auditable. Pretrained passage reranking could strengthen semantic matching [43].

2.4 Temperature Scaling, Adaptive Routing, and Abstention

Raw logits from sparse linear classifiers were highly overconfident. Each context type therefore received a scalar temperature fitted on the fold's calibration subset by minimizing multiclass negative log-likelihood, following temperature scaling [28]. For logits z and temperature $T > 0$,

The optimization searched log T between -5 and 8. Separate temperatures were fitted for BM25, dense, always-hybrid, and fast contexts because their score distributions differed. Evaluation probabilities were transformed only with temperatures learned from the corresponding training-fold calibration subset.

Adaptive routing operated in two stages. First, a claim escalated from the fast path when its calibrated maximum probability c_{fast} fell below gate threshold τ_g . The calibration procedure searched gate quantiles from 0.20 to 0.60 and constrained escalation to 15–65% of claims. Second, an arbiter accepted the full-context posterior only when c_{full} was at least c_{fast} plus margin δ . Otherwise, the fast posterior remained the final posterior even though broader retrieval had been executed. The calibration objective combined accuracy, macro-F1, context tokens, and a mild preference for an escalation rate near 35%. The routing rule was

$$escalate \Leftrightarrow c_{fast} < \tau_g; \text{ use full } \Leftrightarrow escalate \wedge c_{full} \geq c_{fast} + \delta \quad (5)$$

The arbiter prevents a larger retrieved context from automatically replacing a more confident compact-context decision. Confidence-aware fusion is also explicit in autonomous-driving perception systems [44], but here the fused objects are retrieval-conditioned posteriors rather than sensor streams. Retrieval cost for an escalated claim still includes the full context because both retrieval and verification passes occurred.

The final selective decision abstained when the maximum probability of the arbitrated posterior fell below τ_a :

$$answer(q) = \underset{y}{argmax} p_{final}(y|q) \text{ if } \underset{y}{max} p_{final} \geq \tau_a; \text{ otherwise abstain} \quad (6)$$

The threshold search required at least 90% coverage on the calibration subset and maximized a score combining selective accuracy, macro-F1, and coverage. Abstention accuracy is defined as the proportion of abstained evaluation claims whose forced prediction was incorrect. A value above 50% indicates that rejected cases are enriched for errors. Wrong-answer rate is the number of emitted

incorrect predictions divided by all claims, so abstentions lower this rate only when they preferentially remove mistakes.

2.5 Evaluation Protocol

Document retrieval used mean gold-document recall at $k \in \{1, 3, 5, 10, 20\}$; sentence selection used gold-sentence recall and evidence MRR at $m \in \{1, 3, 5, 10\}$. End-to-end metrics were accuracy and macro-F1. For the adaptive system, selective metrics were computed on answered claims and forced accuracy on all claims. Coverage, abstention accuracy, wrong-answer rate, ECE, Brier score, context size, evidence recall, and evidence MRR completed the evaluation.

Context tokens were counted with a deterministic word-and-punctuation tokenizer over the exact verifier evidence string, so they measure retrieval-conditioned input rather than API cost or generation latency. Fold statistics describe split variation, and pooled metrics concatenate the five disjoint predictions. Paired comparisons used 5,000 bootstrap resamples [45] on forced predictions. Reliability plots used ten equal-width confidence bins, and risk-coverage curves retained claims in descending confidence order.

3. Results and Discussion

3.1 Retrieval Quality

Table III and Fig. 3 show that BM25 retrieved 72.2% of gold documents at rank 1 and 87.9% by rank 5. Dense LSA reached 39.0% and 64.6%, confirming that broad semantic similarity often lost exact biomedical entities and directional language. Sparse-first hybrid retrieval reached 63.5% at rank 1, 86.3% at rank 5, and 93.0% at rank 10. It therefore recovered complementary dense documents deeper in the list but weakened the top position.

TABLE III. Gold-document recall (%) over evidence-bearing claims.

Retriever	R@1	R@3	R@5	R@10	R@20
BM25	72.2	83.6	87.9	92.5	94.5
Dense LSA	39.0	56.1	64.6	74.8	85.6
Sparse-first hybrid	63.5	79.7	86.3	93.0	94.9

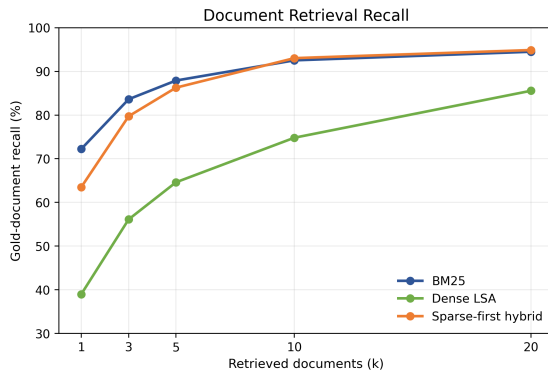


Fig. 3. Gold-document recall as the retrieval cutoff increases. BM25 dominates the dense LSA branch at shallow cutoffs, while sparse-first fusion converges at larger k .

Sentence results follow the same pattern. At five sentences, BM25, hybrid, and dense LSA recalled 54.7%, 52.0%, and 39.7% of gold rationales; at ten, recall rose to 69.8%, 66.7%, and 49.9%. BM25 also led evidence MRR at five sentences (0.470 versus 0.434 and 0.348). Because BM25 document recall@5 was 87.9% but sentence recall@5 only 54.7%, sentence ranking remained a separate bottleneck: a missed qualifier could push the exact rationale below the cutoff.

TABLE IV. Gold evidence-sentence recall (%) and mean reciprocal rank.

Retriever	R@1	R@3	R@5	R@10	MRR@5
BM25	23.3	43.7	54.7	69.8	0.470
Dense LSA	18.0	31.4	39.7	49.9	0.348
Sparse-first hybrid	20.5	41.6	52.0	66.7	0.434

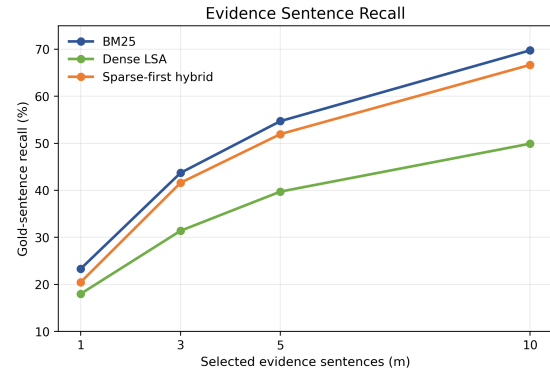


Fig. 4. Evidence-sentence recall under a shared sentence selector. Exact lexical matching remains strongest throughout the tested cutoffs.

3.2 End-to-End Verification and Context Cost

Table V and Fig. 5 report pooled verification. BM25 was the strongest fixed baseline at 52.2% accuracy and 47.3% macro-F1; dense RAG reached 49.5% and 43.5%, and always-hybrid RAG 49.6% and 44.9%. The adaptive system answered 95.7% of claims with 54.4% selective accuracy and 47.9% macro-F1. Its forced accuracy was 53.4%, 1.2 points above BM25, so the gain did not arise solely from withholding rejected cases.

TABLE V. Pooled five-fold verification, evidence, cost, and selective-safety results (% except tokens).

System	Cov.	Acc.	Macro-F1	Forced acc.	Ev. recall	Token s	Wrong ans.	Abs t. acc.
BM25 RAG	100.0	52.2	47.3	52.2	54.7	241.9	47.8	—
Dense RAG	100.0	49.5	43.5	49.5	39.7	237.5	50.5	—
Always-RAG	100.0	49.6	44.9	49.6	52.0	241.5	50.4	—
Uncertain y-gated RAG	95.7	54.4	47.9	53.4	41.8	139.6	43.6	68.8

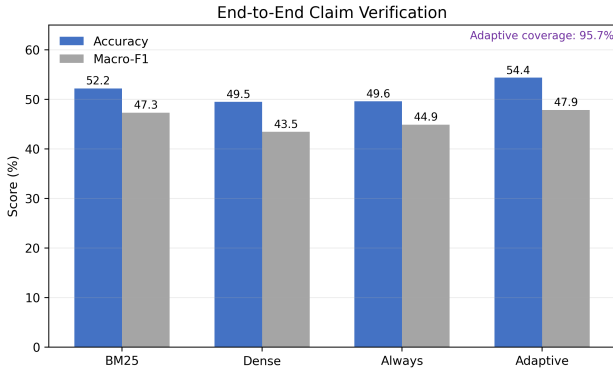


Fig. 5. Accuracy and macro-F1 for the four evaluated systems. Adaptive scores are calculated on answered claims at 95.7% coverage.

Efficiency separates the systems more clearly. BM25, dense, and always-hybrid contexts averaged 241.9, 237.5, and 241.5 tokens; adaptive routing used 139.6 tokens, 2.14 documents, and 2.85 sentences, a 42.3% reduction from BM25. Most claims used one document and two sentences, while only uncertain cases paid for broader retrieval. Resource-aware allocation also appears in hybrid-cloud LLM deployment [46], edge VMAF distillation [47], digital-twin dispatching [48], and progressive PDF rendering [49]. Although the tasks differ, all make computation allocation part of system quality. Index fitting took 11.30 seconds, and the five verifier fits averaged 8.59 seconds.

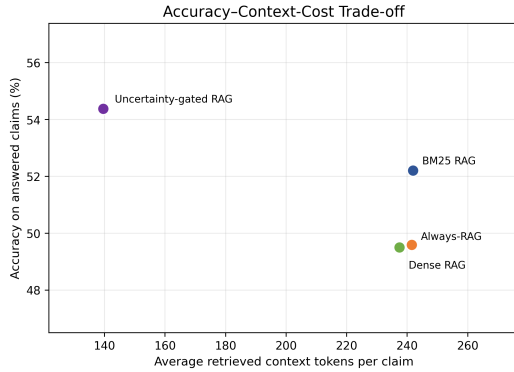


Fig. 6. Selective accuracy versus average retrieved context tokens. The adaptive system uses substantially less evidence context while attaining the highest answered-claim accuracy.

3.3 Class-Wise Behavior

Table VI shows the class imbalance behind aggregate accuracy. BM25 F1 was 0.527 for SUPPORT, 0.203 for CONTRADICT, and 0.689 for NEI. Dense retrieval increased NEI recall to 0.767 but reduced SUPPORT recall to 0.421. Adaptive selective prediction reached NEI precision 0.680, recall 0.814, and F1 0.741; SUPPORT F1 remained 0.515 and CONTRADICT F1 0.180. The gain therefore came mainly from uncertain boundary cases, not solved contradiction reasoning.

TABLE VI. Pooled per-class results on answered claims.

System	Class	Precision	Recall	F1	N answered
BM25	SUPPORT	0.489	0.572	0.527	456

System	Class	Precision	Recall	F1	N answered
BM25	CONTRADICT	0.224	0.186	0.203	237
BM25	NEI	0.723	0.659	0.689	416
Dense	SUPPORT	0.500	0.421	0.457	456
Dense	CONTRADICT	0.209	0.160	0.181	237
Dense	NEI	0.587	0.767	0.665	416
Always	SUPPORT	0.475	0.496	0.485	456
Always	CONTRADICT	0.204	0.181	0.192	237
Always	NEI	0.666	0.675	0.671	416
Adaptive	SUPPORT	0.529	0.502	0.515	442
Adaptive	CONTRADICT	0.209	0.159	0.180	227
Adaptive	NEI	0.680	0.814	0.741	392

3.4 Calibration and Abstention

Temperature scaling preserved argmax labels but sharply improved confidence quality. Fold-mean ECE fell from 0.469 to 0.049 for BM25, 0.498 to 0.046 for dense, and 0.491 to 0.063 for hybrid; Brier scores fell from 0.940-0.999 to 0.588-0.611, and negative log-likelihood approached 1.0. Large fitted temperatures reflected sharp averaged-SGD logits. The adaptive answered subset achieved pooled ECE 0.040 and Brier score 0.571, although Fig. 7 still shows residual bin-level deviations.

TABLE VII. Fold-mean calibration before and after temperature scaling on evaluation folds.

Context	ECE raw	ECE scaled	Brier raw	Brier scaled	NLL raw	NLL scaled
BM25	0.469	0.049	0.940	0.588	12.838	0.966
Dense	0.498	0.046	0.999	0.611	14.157	1.015
Always	0.491	0.063	0.985	0.603	13.217	0.998
Fast path	0.446	0.053	0.895	0.571	12.990	0.951

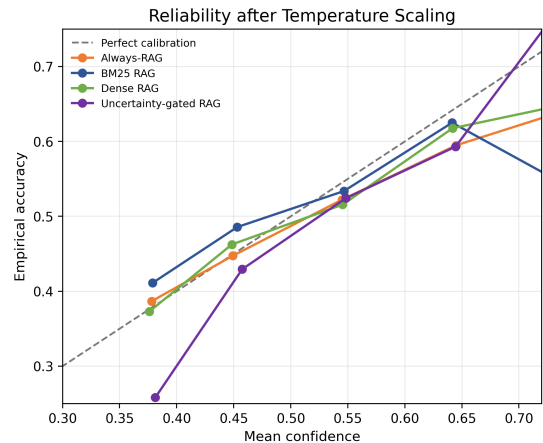


Fig. 7. Reliability curves after temperature scaling. Points aggregate confidence bins across the five disjoint evaluation folds.

The adaptive system abstained on 48 of 1,109 claims, producing 95.7% coverage. Among these 48, 33 forced predictions were wrong, so abstention accuracy was 68.8%. The wrong-answer rate was 43.6%, compared with 47.8% for BM25. This is an 8.7% relative reduction in emitted errors. Removing abstention but retaining routing increased coverage to 100% and lowered accuracy from 54.4% selective to 53.4% forced, while wrong-answer rate increased from 43.6% to 46.6%. Thus the final threshold rejected a small, error-enriched subset rather than indiscriminately reducing coverage.

3.5 Routing and Risk-Coverage Trade-offs

Table VIII reports the learned thresholds, routing rates, and construction times. Across folds, the fast-path rate was 71.5% and escalation was 28.5%; the full posterior replaced the fast posterior for 20.1% of all claims, while 8.4% were escalated but retained the fast decision. The gate threshold averaged 0.473, the arbiter margin -0.008, and the abstention threshold 0.407. Fold coverage ranged from 90.1% to 98.2%. The staged safeguard resembles drift early warning with safe rollback [50], noisy-neighbor VM risk modeling [51], and AIOps fault analysis [52], although the trigger here is calibrated claim uncertainty.

TABLE VIII. Adaptive thresholds, routing rates, and measured construction times.

Routing / runtime quantity	Mean $\pm$ SD or measured value
Gate threshold	0.473 $\pm$ 0.037
Arbiter margin	-0.008 $\pm$ 0.034
Abstention threshold	0.407 $\pm$ 0.019
Fast-path share	71.5 $\pm$ 6.3%
Escalation share	28.5 $\pm$ 6.3%
Full posterior used	20.1 $\pm$ 8.6%
Abstention share	4.3 $\pm$ 3.3%
Coverage	95.7 $\pm$ 3.3%
Fast temperature	68.103 $\pm$ 8.438
Full temperature	68.845 $\pm$ 14.188
Retrieval index fit	11.30 s
Verifier fit per fold	8.59 s mean

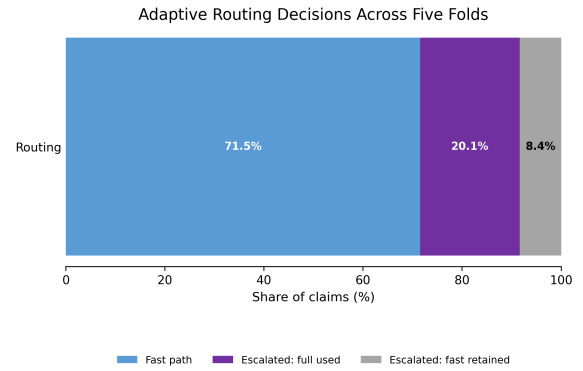


Fig. 8. Mean adaptive routing decisions. The gray segment represents escalated claims for which the arbiter retained the fast posterior.

The coverage sweep in Table IX and the risk-coverage curves in Fig. 9 show a smooth selective trade-off. At 79.8% actual coverage, adaptive accuracy was 57.3%; at 89.7% it was 56.0%; at 94.7% it was 54.8%; and forced 100% coverage produced 53.4%. The highest-confidence subset was consistently safer, although absolute risk remained substantial because the sparse verifier faced three-way scientific inference. This budget-risk framing is analogous to constrained auto-bidding [53] and data-driven budgeting [54], while the controlled resource here is retrieved evidence context.

TABLE IX. Adaptive risk and performance at alternative target coverages (fold means, % except threshold).

Target cov.	Actual cov.	Threshold	Accuracy	Macro-F1	Risk
80.0	79.8	0.469	57.3	49.5	42.7
85.0	84.8	0.456	56.6	48.9	43.4
90.0	89.7	0.440	56.0	48.9	44.0
95.0	94.7	0.414	54.8	48.2	45.2
100.0	100.0	0.000	53.4	47.2	46.6

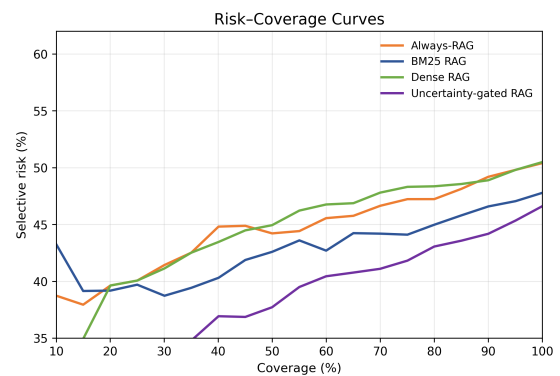


Fig. 9. Selective risk as coverage increases. Curves rank claims by calibrated confidence; lower risk is better.

3.6 Ablation and Statistical Comparison

Table X isolates the components. Without abstention, adaptive context remained 139.6 tokens but accuracy fell to 53.4% and wrong-answer rate rose to 46.6%. Always retrieving hybrid context increased cost to 241.5 tokens and reduced accuracy to 49.6%. BM25 outperformed always-hybrid, while the dense-only branch was weakest at 49.5%. The full result therefore requires a strong sparse fast path,

selective broader retrieval, and calibrated rejection; unconditional fusion is insufficient.

TABLE X. Component ablation using evaluated five-fold predictions.

Configuration	Acc.	Macro-F1	Cov.	Tokens	Wrong ans.
Full uncertainty gate + calibrated abstention	54.4	47.9	95.7	139.6	43.6
- abstention (forced answers)	53.4	47.3	100.0	139.6	46.6
- adaptive gate (always retrieve)	49.6	44.9	100.0	241.5	50.4
- dense branch (sparse retrieval only)	52.2	47.3	100.0	241.9	47.8
- sparse branch (dense retrieval only)	49.5	43.5	100.0	237.5	50.5

Table XI reports paired bootstrap comparisons. Adaptive forced predictions exceeded dense RAG by 3.86 points (95% CI 1.62-6.22; $p=0.0008$). The 1.19-point gain over BM25 was inconclusive (95% CI -0.90-3.16; $p=0.2692$). Always-RAG was 2.61 points below BM25 ($p=0.0100$) and did not differ from dense RAG. Thus the strongest adaptive evidence is lower context, error-enriched abstention, and superiority to dense RAG, not a statistically established forced-accuracy advantage over BM25.

TABLE XI. Paired 5,000-resample bootstrap on forced predictions.

Paired comparison	Δ accuracy (pp)	95% CI (pp)	p	Result
Uncertainty-gated RAG – BM25 RAG	1.19	[-0.90, 3.16]	0.2692	not significant
Uncertainty-gated RAG – Dense RAG	3.86	[1.62, 6.22]	0.0008	significant
Always-RAG – BM25 RAG	-2.61	[-4.51, -0.72]	0.0100	significant
Always-RAG – Dense RAG	0.06	[-2.16, 2.34]	0.9992	not significant

3.7 Error Analysis and Practical Implications

The rule-based attribution in Table XII and Fig. 10 assigned every adaptive forced-prediction error to the first applicable category. Negation or contradiction mismatch was largest with 127 cases, followed by 104 document retrieval misses, 95 sentence selection misses, 92 entailment reasoning errors, 85 NEI boundary errors, and 14 numeric mismatches. Retrieval and sentence selection together accounted for 199 errors, but 318 errors remained after relevant evidence was available. Better retrieval alone therefore cannot close the performance gap; the verifier must model contradiction, scope, and evidence sufficiency more precisely.

TABLE XII. First-applicable error attribution for adaptive forced predictions.

Error category	Count	Share
negation/contradiction error	127	24.6%
document retrieval miss	104	20.1%
sentence selection miss	95	18.4%
entailment reasoning error	92	17.8%
NEI boundary ambiguity	85	16.4%
numeric mismatch	14	2.7%

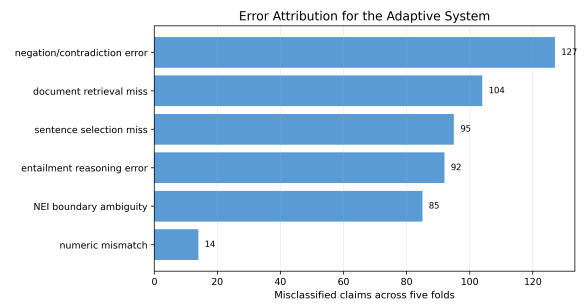


Fig. 10. Error attribution for the adaptive system. Negation and contradiction remain the dominant verifier failure mode.

The class results follow the error profile. CONTRADICT claims often differ from evidence through a directional verb, negation, population restriction, or numeric comparator; surface n-grams do not reliably compose these cues into entailment. NEI creates another boundary because topically related text may still fail to establish the claim. Escalation and abstention improved NEI behavior, but cautious decisions could also shift true SUPPORT cases toward NEI.

For deployment, the results favor a tiered verification service rather than one fixed RAG configuration. The fast path handles most claims with a compact evidence bundle; escalation allocates additional retrieval only when confidence is low; arbitration prevents a larger context from overriding a stronger compact prediction; and the final threshold sends the least reliable cases to review. Thresholds are operating points, not universal constants. Across decision-support domains, IoT traffic anomaly detection [55], financial risk dashboards [56], C5.0 bank credit screening [57], traffic-sign explanation cards [58], and predictive DDoS mitigation [59] likewise separate model scoring from evidence display, escalation, or action. These analogies support the architecture pattern, not direct cross-domain validity.

Transfer must still be evaluated for each user and task. Bilingual CS education research provides a technical setting in which language and presentation affect engagement [60]; AI-driven financial-service analysis [61] and counterfactual ad ranking [62] expose policy consequences; and e-commerce recommendation [63], DenseNet cancer-image classification [64], and data-driven marketing decisions [65] span markedly different error costs. A production scientific verifier should therefore test whether reviewers understand the retrieved rationale,

calibrate thresholds to the application, and log every escalation and abstention.

4. Limitations

The first limitation is model capacity. The dense branch used 256-dimensional LSA, and the verifier used sparse lexical features with logistic classification. DPR [22], SciBERT [24], SPECTER [26], Sentence-BERT [27], or cross-encoder rerankers [43] can provide richer semantics, but they also introduce model-weight provenance, hardware, latency, and tuning variables absent from this compact comparison.

Second, context tokens measure only the retrieved evidence passed to the verifier. They exclude a generative decoder's prompt, output tokens, caching, concurrency, and provider-specific billing. The reported 42.3% reduction is therefore an input-context reduction, not a universal monetary saving. Production latency also requires separate measurements of retrieval, sentence selection, inference, storage, and load.

Third, SciFact is a closed corpus of 5,183 abstracts and 1,109 labeled claims. The five-fold design uses all labeled claims efficiently, but the data are smaller and cleaner than web-scale scientific search. Some claims require full text, tables, supplements, or evidence distributed across papers. The released test split has no labels, so it cannot provide another independent score. SciFact-Open [7] and long-document retrieval settings [13] are necessary extensions, while domain-bridging methods [25] offer a separate route to robustness.

Fourth, selective accuracy remains 54.4% and CONTRADICTION F1 remains 0.180. Abstention reduces emitted errors but does not turn a weak verifier into an autonomous judge of scientific truth. The appropriate role is evidence-linked decision support with human escalation. The deterministic, mutually exclusive error taxonomy is also diagnostic rather than causal: retrieval misses can coexist with difficult negation, and sentence-selection misses can coexist with an NEI boundary.

5. Conclusion

This study compared fixed and adaptive retrieval-augmented verification on all 1,109 labeled SciFact claims. BM25 was the strongest retriever, with 87.9% gold-document and 54.7% gold-sentence recall at five units; dense LSA struggled with terminology-sensitive matching, and unconditional fusion did not improve classification. The adaptive system used a compact BM25 path, escalated 28.5% of claims, accepted the full posterior for 20.1%, and abstained on 4.3%. It achieved 54.4% selective accuracy at 95.7% coverage, reduced context by 42.3%, and concentrated 68.8% of abstentions on otherwise wrong predictions. Its forced gain over BM25 was not significant, but its gain over dense RAG was. Adaptive evidence allocation and calibrated abstention therefore improve the efficiency-safety profile of scientific RAG, while stronger contradiction reasoning and sentence selection remain the main needs.

References

- [1] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, and H. Hajishirzi, "Fact or Fiction: Verifying Scientific Claims," in Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 7534–7550, doi: 10.18653/v1/2020.emnlp-main.609.
- [2] J. Thorne, A. Vlachos, O. Cocarascu, 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: Human Language Technologies, vol. 1, 2018, pp. 809–819, doi: 10.18653/v1/N18-1074.
- [3] N. Kotonya and F. Toni, "Explainable Automated Fact-Checking for Public Health Claims," in Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 7740–7754, doi: 10.18653/v1/2020.emnlp-main.623.
- [4] A. Saakyan, T. Chakrabarty, and S. Muresan, "COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic," in Proc. 59th Annual Meeting of the Association for Computational Linguistics and 11th Int. Joint Conf. Natural Language Processing, vol. 1, 2021, pp. 2116–2129, doi: 10.18653/v1/2021.acl-long.165.
- [5] M. Sarroui, A. Ben Abacha, Y. Mrabet, and D. Demner-Fushman, "Evidence-based Fact-Checking of Health-related Claims," in Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 3499–3512, doi: 10.18653/v1/2021.findings-emnlp.297.
- [6] 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.
- [7] D. Wadden, K. Lo, B. Kuehl, A. Cohan, I. Beltagy, L. L. Wang, and H. Hajishirzi, "SciFact-Open: Towards Open-Domain Scientific Claim Verification," in Findings of the Association for Computational Linguistics: EMNLP 2022, 2022, pp. 4719–4734, doi: 10.18653/v1/2022.findings-emnlp.347.
- [8] 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.
- [9] R. Pradeep, X. Ma, R. Nogueira, and J. Lin, "Scientific Claim Verification with VerT5erini," in Proc. 12th Int. Workshop on Health Text Mining and Information Analysis, 2021, pp. 94–103, doi: 10.18653/v1/2021.louhi-1.11.
- [10] 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.
- [11] 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, doi: 10.18653/v1/2023.findings-acl.387.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] H. 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.
- [17] 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, doi: 10.1561/15000000019.
- [18] 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, doi: 10.1145/361219.361220.
- [19] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391-407, 1990, doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-AS11>3.0.CO;2-9.
- [20] G. V. Cormack, C. L. A. Clarke, and S. Büttcher, "Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods," in *Proc. 32nd Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, 2009, pp. 758-759, doi: 10.1145/1571941.1572114.
- [21] 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/imaia/iaae020.
- [22] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6769-6781, doi: 10.18653/v1/2020.emnlp-main.550.
- [23] 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: Human Language Technologies*, vol. 1, 2019, pp. 4171-4186, doi: 10.18653/v1/N19-1423.
- [24] I. 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 Int. Joint Conf. Natural Language Processing*, 2019, pp. 3615-3620, doi: 10.18653/v1/D19-1371.
- [25] Z. S. Zhong, X. Pan, and Q. Lei, "Bridging domains with approximately shared features," in *Proc. 28th Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, vol. 258, 2025, pp. 559-567.
- [26] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld, "SPECTER: Document-level Representation Learning using Citation-informed Transformers," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2270-2282, doi: 10.18653/v1/2020.acl-main.207.
- [27] 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 Int. Joint Conf. Natural Language Processing*, 2019, pp. 3982-3992, doi: 10.18653/v1/D19-1410.
- [28] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proc. 34th Int. Conf. Machine Learning*, vol. 70, 2017, pp. 1321-1330.
- [29] 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.
- [30] S. Desai and G. Durrett, "Calibration of Pre-trained Transformers," in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 295-302, doi: 10.18653/v1/2020.emnlp-main.21.
- [31] Z. S. Zhong and S. Ling, "Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization," *arXiv:2408.05944*, Aug. 2024.
- [32] M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, and M. Lucic, "Revisiting the Calibration of Modern Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 15682-15694.
- [33] Q. Xin, "Uncertainty-aware late fusion for 3D perception: Confidence calibration and fusion rule learning," *J. Technol. Informatics Eng.*, vol. 4, no. 1, pp. 215-238, Apr. 2025, doi: 10.51903/jtie.v4i1.485.
- [34] Y. Geifman and R. El-Yaniv, "Selective Classification for Deep Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4885-4894.
- [35] 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.
- [36] Y. Geifman and R. El-Yaniv, "SelectiveNet: A Deep Neural Network with an Integrated Reject Option," in *Proc. 36th Int. Conf. Machine Learning*, vol. 97, 2019, pp. 2151-2159.
- [37] A. Kamath, R. Jia, and P. Liang, "Selective Question Answering under Domain Shift," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5684-5696, doi: 10.18653/v1/2020.acl-main.503.
- [38] D. Hendrycks and K. Gimpel, "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks," in *Proc. Int. Conf. Learning Representations*, 2017.
- [39] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in *Advances in*

- Neural Information Processing Systems, vol. 30, 2017, pp. 6402–6413.
- [40] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning,” in Proc. 33rd Int. Conf. Machine Learning, vol. 48, 2016, pp. 1050–1059.
- [41] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [42] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, “Optimization of autonomous driving image detection based on RFACnv and triplet attention,” in Proc. 2nd Int. Conf. Software Engineering and Machine Learning (SEML), 2024, pp. 210–217, doi: 10.54254/2755-2721/77/2024MA0067.
- [43] R. Nogueira and K. Cho, “Passage Re-ranking with BERT,” arXiv:1901.04085, 2019.
- [44] 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.
- [45] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993.
- [46] 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.
- [47] 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.
- [48] L. 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.
- [49] 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.
- [50] 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.
- [51] 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.
- [52] 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.
- [53] 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.
- [54] 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.
- [55] Q. Xin, Z. Xu, L. Guo, F. Zhao, and B. Wu, “IoT traffic classification and anomaly detection method based on deep autoencoders,” in Proc. 6th Int. Conf. Computing and Data Science (CDS), 2024, pp. 64–70, doi: 10.54254/2755-2721/69/20241511.
- [56] 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. Graphic Des.*, vol. 3, no. 1, pp. 196–210, May 2025, doi: 10.51903/ijgd.v3i1.3715.
- [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] 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. Graphic Des.*, vol. 3, no. 2, pp. 455–473, Oct. 2025, doi: 10.51903/ijgd.v3i2.3991.
- [59] B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, “Predictive optimization of DDoS attack mitigation in distributed systems using machine learning,” in Proc. 6th Int. Conf. Computing and Data Science (CDS), 2024, pp. 89–94, doi: 10.54254/2755-2721/64/20241350.
- [60] A. G. S. Raj et al., “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] 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.
- [62] 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.
- [63] K. Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, “Intelligent classification and personalized recommendation of e-commerce products based on machine learning,” in Proc. 6th Int. Conf. Computing and Data Science (ICCDs), 2024, pp. 143–149, doi: 10.54254/2755-2721/64/20241365.
- [64] 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.
- [65] 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.