

GraphRAG for Scientific Evidence Verification through Claim–Document–Entity Graph Reasoning

Sarah Miller¹, Bo Chen², James Peterson³, Ying Wang⁴, Daniel Clark⁵

¹Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, IL, USA

²Department of Computer Science, Cornell University, Ithaca, NY, USA

³Department of Computer Science, Princeton University, Princeton, NJ, USA

⁴Department of Computer Science, The University of Texas at Austin, Austin, TX, USA

⁵Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA

sarah.miller.dev@yahoo.com

Abstract

Scientific claim verification requires a system to retrieve relevant papers, isolate sentence-level evidence, and assign a SUPPORT, CONTRADICT, or not-enough-information label. This study evaluates a corpus-derived GraphRAG design on SciFact using the five predefined cross-validation folds. The system represents 5,183 papers, 45,952 abstract sentences, and 21,025 scientific entity phrases as a heterogeneous graph with 459,461 typed edges. A 256-dimensional latent-semantic Vector RAG baseline is compared with graph propagation and a training-fold-selected hybrid fusion. Across 1,109 out-of-fold claims, the Hybrid system achieved document recall@5 of 0.7519, compared with 0.6421 for Vector RAG and 0.6654 for GraphRAG. Complete evidence-set recall@5 reached 0.4645, an absolute gain of 0.1082 over Vector RAG. GraphRAG produced the strongest retrieval-augmented verification accuracy (0.4598) and strict grounded accuracy (0.3462), while the claim-only classifier retained the highest label accuracy (0.4860). Paired analysis confirmed that Hybrid improved evidence-set success over Vector RAG, but its grounded-label gain was not statistically significant. The results establish that graph structure reliably broadens evidence coverage and improves resilience to query token loss, while downstream verification still depends on relation-sensitive reasoning that cannot be replaced by retrieval depth alone.

Keywords : Graphrag; Scientific Fact Checking; Evidence Retrieval; Knowledge Graph; Retrieval-Augmented Generation; Scifact; Claim Verification

1. Introduction

Scientific knowledge is communicated through claims whose validity depends on precise experimental conditions, populations, interventions, and measured outcomes. Automated verification therefore differs from topical search: a useful system must locate a relevant paper, identify the exact sentence or sentence group that bears on the claim, and preserve the polarity of the reported relationship. SciFact formalized this problem with expert-written claims, evidence-containing abstracts, sentence rationales, and SUPPORT or REFUTE judgments [1]. The broader fact-verification literature established similar evidence requirements through FEVER [2], while PUBHEALTH [3], COVID-Fact [4], and CLIMATE-FAVER [5] demonstrated that domain terminology and evidential context materially change the task. The scientific fact-checking survey by Vladika and Matthes further identifies retrieval, rationale extraction, and stance prediction as the central components of an end-to-end system [6].

Existing scientific verification systems have improved each component, but most treat candidate papers as independent

text units. VerT5erini uses a sequence-to-sequence model across abstract retrieval, sentence selection, and label prediction [7]. ARSJoint shares information across retrieval, rationale, and stance objectives to reduce pipeline error propagation [8]. MultiVerS jointly predicts document labels and rationales and demonstrates the value of scientific-domain pretraining and weak supervision [9]. These approaches show that stronger encoders improve verification, yet they do not directly model the corpus-level connections created when multiple papers and sentences share the same molecular entities, clinical concepts, biological processes, or statistical phrases. Independent ranking can therefore miss evidence that is lexically distant from the claim but structurally close through shared scientific concepts. Related work in code intelligence explicitly separates finding code from explaining it [29], while long-document contractual analysis addresses retrieval and reasoning across dispersed clauses [30]. Both lines of work reinforce the need to preserve the path from a query to the supporting unit rather than only a final label.

Retrieval-augmented generation provides a general mechanism for conditioning prediction on external evidence.

The original RAG framework combines parametric sequence generation with a nonparametric dense index [10]. Dense Passage Retrieval [11] and REALM [12] established learned retrieval as an effective alternative to purely lexical matching, while Fusion-in-Decoder [13] showed that multiple retrieved passages can be integrated at the reader stage. Atlas demonstrated that retrieval-augmented language models can learn knowledge-intensive tasks with comparatively limited supervision [14], and KILT established a common evaluation framework for knowledge-grounded systems [15]. These systems motivate a separation between retrieval and reasoning, but vector similarity alone represents each paper as a point rather than as an object embedded in a network of entities, sentences, and neighboring documents. Domain-specific retrieval systems have extended this pattern to log anomalies by coupling retrieved failure narratives with selective refusal [31] and to tokenized financial disclosures by grounding generated answers in cited evidence [32].

Scientific encoders such as SciBERT [16] and general sentence embedding architectures such as Sentence-BERT [17] provide strong semantic representations. Their effectiveness does not remove the structural problem. A dense vector can recognize that two expressions are related, but it does not explicitly expose which entity phrase connects a claim to a paper, which sentence contains that phrase, or which neighboring papers reinforce the connection. This study therefore holds the verifier family and training protocol fixed while changing the retrieval representation. The comparison isolates the value of corpus structure without attributing gains to a larger language model, external database, or proprietary embedding service. Where labels are intrinsically ambiguous, human-uncertainty distillation offers a complementary route to calibrated predictions rather than forcing all observations into hard targets [33].

Graph-based question answering supplies the closest methodological precedent. GRAFT-Net reasons over question-specific subgraphs that combine text and knowledge-base entities [18], and PullNet iteratively expands a subgraph before answer extraction [19]. QA-GNN connects language-model representations to knowledge-graph reasoning [20], while GreaseLM performs bidirectional language-graph interaction [21]. KG-FiD uses a knowledge graph to rerank passages before a Fusion-in-Decoder reader [22]. Graph convolution [23] and the broader graph-network formulation [24] explain why relation-specific neighborhoods can propagate evidence beyond a single lexical match. The present system adapts that principle to scientific verification with a lightweight, auditable graph built directly from the SciFact corpus. Spectral rank aggregation likewise frames graph evidence as a ranking problem with analyzable recovery behavior [34].

The central research question is whether claim–document–sentence–entity structure improves evidence acquisition when all methods use the same corpus and evaluation folds.

Four subsidiary questions guide the analysis. First, does graph propagation improve document retrieval at shallow and deeper cutoffs? Second, does vector–graph fusion recover complete evidence groups more reliably than either branch alone? Third, do retrieval gains transfer to label accuracy and strict evidence-grounded accuracy? Fourth, what costs, robustness properties, and failure modes accompany structural retrieval? These questions are evaluated with document, sentence, evidence-set, classification, latency, robustness, ablation, and paired significance measurements.

A graph representation is particularly appropriate for scientific abstracts because evidence often depends on repeated concept combinations rather than on a single exact phrase. A claim can name an intervention and an outcome, while the evidence sentence uses an abbreviation for the intervention and places the outcome in a neighboring sentence. A vector retriever compresses this configuration into a similarity score. A heterogeneous graph retains separable paths through the intervention phrase, outcome phrase, parent paper, and evidence sentences. Those paths do not perform logical entailment, but they preserve candidate structure that can be inspected, pruned, and combined with semantic similarity. This structural separation is central to evidence verification, where the reason a paper was retrieved is as important as its topical proximity. Multimodal perception systems make a related design choice when they fuse independent evidence channels rather than collapse them prematurely [35].

The evaluation also adopts a deliberately strict distinction between relevance and verification. Document recall credits the presence of any annotated evidence paper, complete evidence-set recall requires every sentence in at least one rationale group, and strict grounded accuracy adds the requirement of a correct stance label. These nested metrics expose where performance is gained or lost. A method can retrieve the correct paper but miss the decisive sentence, or retrieve the complete rationale but misread its polarity. Reporting only end-to-end accuracy would conceal those differences, while reporting only retrieval would overstate practical verification quality. The experimental design therefore follows the full evidence chain from claim to paper, sentence group, and final label. Operationally, this distinction matters because silent data drift can invalidate a once-reliable predictor [36], while confidence-aware late fusion shows that the combination rule itself can alter calibration [37].

The study makes three contributions. It defines a heterogeneous GraphRAG representation in which claim nodes connect to corpus-derived scientific entity phrases, entity phrases connect to papers and sentences, and paper nodes retain explicit sentence membership. It conducts five-fold out-of-fold evaluation of Vector RAG, GraphRAG, and Hybrid retrieval on every labeled SciFact claim, using fold-specific fusion selection and fold-specific verification models. It also separates evidence acquisition from final

stance reasoning, showing that the hybrid retriever produces the clearest retrieval gain while the graph-only branch produces the strongest retrieval-augmented label and grounding scores. This distinction prevents a retrieval improvement from being reported as a verification improvement when the downstream classifier does not realize the gain.

2. Method

Dataset and evaluation protocol. The experiments use the SciFact corpus and its predefined five-fold claim splits. The corpus contains 5,183 papers represented by titles and abstracts and 45,952 abstract sentences. The labeled pool contains 1,109 unique claims: 456 SUPPORT, 237 CONTRADICT, and 416 NEI. The annotations identify 437 distinct evidence papers and 1,391 evidence-sentence instances. The original training and development files contribute 809 and 300 labeled claims, respectively. The separate 300-claim test file contains no public evidence labels, so every reported score is computed from the five labeled held-out folds rather than from an unverifiable test estimate. Table I summarizes the evaluated material.

TABLE I. SciFact data composition used in five-fold evaluation

Item	Value
Corpus papers	5,183
Abstract sentences	45,952
Labeled claims	1,109
SUPPORT claims	456
CONTRADICT claims	237
NEI claims	416
Distinct evidence papers	437
Annotated evidence sentences	1,391

Each fold trains on approximately four-fifths of the labeled claims and predicts only the held-out fifth. Retrieval indices are built once from the unlabeled paper corpus because corpus indexing does not use claim labels or evidence annotations. The hybrid fusion coefficient is selected using only the training portion of each fold. The stance classifier and its regularization parameter are also fitted inside each fold. Concatenating the five held-out predictions yields one out-of-fold prediction for every labeled claim, which supports paired bootstrap and Wilcoxon tests without evaluating a training example as a test example.

Graph construction. The graph contains document, sentence, and entity-phrase nodes; a claim node is instantiated at query time. Each document connects to every sentence in its abstract. Entity phrases are extracted with a TF-IDF vocabulary over one-, two-, and three-token scientific phrases, using a minimum document frequency of two, a maximum document-frequency ratio of 0.20, and a

vocabulary cap of 25,000. The fourteen highest-weight phrases are retained for each paper and the eight highest-weight phrases for each sentence. Claim nodes connect to their ten highest-weight phrases. The resulting graph contains 72,160 nodes and 459,461 edge instances, as reported in Table II and illustrated in Figs. 1 and 2.

TABLE II. Heterogeneous graph inventory

Graph quantity	Value
Document nodes	5,183
Sentence nodes	45,952
Entity-phrase nodes	21,025
Document-entity edges	72,560
Sentence-entity edges	340,949
Document-sentence edges	45,952
Total nodes	72,160
Total edges	459,461
Mean entity phrases per document	13.9996
Mean entity phrases per sentence	7.4197

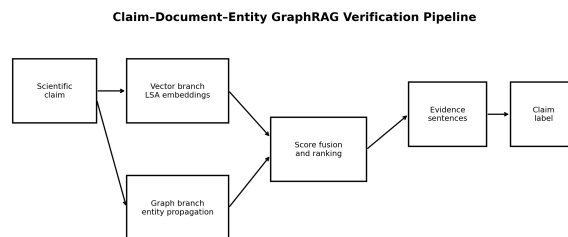


Fig. 1. Claim-document-entity retrieval and verification pipeline.

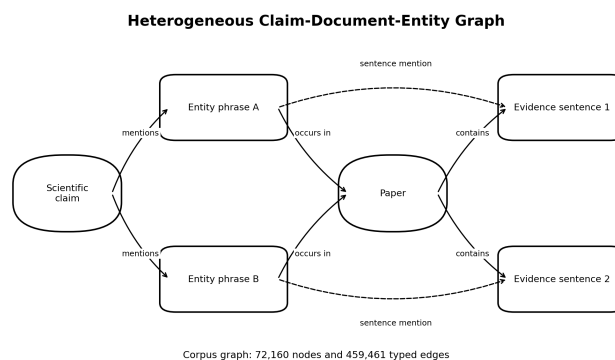


Fig. 2. Node and edge types used by the corpus-derived scientific graph.

The entity representation is deliberately corpus-derived. It recognizes multiword scientific expressions that recur in the collection without requiring an external ontology, entity linker, or abbreviation dictionary. This decision keeps the

comparison tied to the available text and avoids injecting facts that are absent from the benchmark corpus. It also means that the term entity refers to a weighted scientific keyphrase rather than a canonical biomedical identifier. Degree normalization downweights phrases that occur in many documents, preventing broad expressions such as common assay terms from dominating more discriminative phrases.

Vector RAG. The vector branch uses lowercased word unigrams and bigrams with English stop-word removal, sublinear term frequency, a 60,000-feature cap, and L2 normalization. The paper title is repeated once before the abstract to provide a fixed title prior. Truncated singular-value decomposition maps document and sentence TF-IDF vectors to 256-dimensional latent-semantic representations. This construction follows the term-weighting principles developed for vector-space retrieval [27] and the latent semantic analysis formulation of Deerwester et al. [26]. Cosine similarity between a claim vector and every paper vector produces the document ranking. Candidate sentences are then ranked by claim-sentence cosine similarity.

GraphRAG scoring. Let q be the normalized sparse claim-entity vector and D the normalized document-entity matrix after inverse-square-root entity-degree weighting. The direct graph score is $s_{\text{direct}} = Dq$. A one-step diffusion computes an expanded entity vector from $D^T s_{\text{direct}}$, applies the same degree correction, and projects the result back to documents. The final graph score is $0.72s_{\text{direct}} + 0.28s_{\text{diffused}}$. This short propagation is related to graph convolution [23], graph-network message passing [24], and link-analysis methods such as PageRank [28], but it remains a deterministic weighted propagation rather than a trained graph neural network. The restriction to one diffusion step prevents high-degree hubs from overwhelming claim-specific evidence.

For sentence ranking, GraphRAG combines direct claim-sentence entity overlap with the normalized score of the sentence's parent paper. The sentence score is 0.68 times entity similarity plus 0.32 times parent-document relevance. Explicit parent membership prevents a sentence from being selected solely because it contains a frequent entity phrase; it must also belong to a paper supported by the graph path. TextRank demonstrated the utility of graph-centrality signals for sentence selection [25], while GRAFT-Net [18], PullNet [19], QA-GNN [20], GreaseLM [21], and KG-FiD [22] motivate the use of heterogeneous paths for retrieval and reasoning.

Hybrid retrieval. Vector and graph document scores are min-max normalized for each query and combined as $s_{\text{hybrid}} = \alpha s_{\text{vector}} + (1 - \alpha)s_{\text{graph}}$. The candidate set uses the ten highest-scoring papers, and sentence scores are fused with the same coefficient. Alpha is selected from {0.2, 0.4, 0.6, 0.8} by maximizing training-fold document recall@5 and then training-fold MRR. Every fold selected $\alpha = 0.6$, which gives slightly more weight to semantic similarity while retaining a substantial structural component. Because this

value is chosen independently inside each fold, the reported hybrid results do not use held-out evidence labels for tuning.

Verification model. The verifier receives the claim, the titles of the three highest-ranked papers, and the five highest-ranked evidence sentences. Word unigrams and bigrams are combined with character three- to five-grams in a feature union. A class-balanced one-versus-rest logistic regression model predicts CONTRADICT, NEI, or SUPPORT. The regularization coefficient is selected from {0.5, 1.0, 2.0} by three-fold stratified cross-validation within the training claims. A claim-only baseline uses the identical feature construction, classifier family, regularization search, and fold assignments but excludes retrieved paper titles and evidence sentences. The fixed verifier makes differences among the three RAG conditions attributable to their retrieved context rather than to different readers.

Metrics. Document retrieval is measured by reciprocal rank and recall at 1, 3, 5, and 10 over verifiable claims. Evidence evaluation reports complete evidence-set recall, which awards credit only when at least one annotated rationale group is fully contained in the top- k sentences, together with sentence-level macro precision, recall, and F1. Label evaluation reports accuracy, macro-F1, and class-specific F1. Strict grounded accuracy requires a correct label and a complete evidence group for SUPPORT or CONTRADICT claims; a correctly predicted NEI claim satisfies the metric without an evidence sentence. The claim-only baseline has no grounded score because it does not retrieve evidence. Because point estimates can hide unstable rankings, the analysis also follows the broader uncertainty-quantification principle that estimator variability should be reported alongside central performance [38].

TABLE III. Experimental configuration

Hyperparameter	Value
Vector dimension	256
Lexical n-grams	1–2
Entity phrase n-grams	1–3
Entity vocabulary cap	25000
Entities per paper	14
Entities per sentence	8
Graph direct/diffused weights	0.72/0.28
Retrieved papers	10
Retrieved sentences	10
Classifier context sentences	5
Random seed	2026

Additional analyses. Retrieval depth is varied from one to ten sentences while holding each out-of-fold label prediction fixed, isolating how additional evidence coverage changes strict grounding. The graph ablation removes diffusion, and the hybrid ablation compares fusion with and without

diffusion. Robustness is measured by deterministically deleting 20% of non-negation content tokens from each verifiable claim and recomputing document recall@5. Efficiency uses 200 fixed claims, three uncached end-to-end retrieval passes per method, and the median pass for mean and 95th-percentile latency. Paired differences use 5,000 bootstrap resamples and a Wilcoxon signed-rank test. Table III records the complete configuration. These tests reflect deployment concerns found in deep-autoencoder IoT anomaly detection [39] and residual-fusion modeling of noisy-neighbor VM degradation [40], where performance can change when signals are corrupted or environmental conditions shift.

Label and rationale handling. SciFact stores evidence as document-indexed groups, and a claim can have more than one acceptable rationale. SUPPORT is assigned when any annotated evidence entry carries a SUPPORT label; otherwise an evidence-bearing claim is CONTRADICT, and a claim without evidence entries is NEI. Retrieval metrics are computed only over evidence-bearing claims because NEI has no gold paper or sentence. Classification metrics include all three classes. Complete evidence-set recall treats each annotated group as an alternative: the prediction succeeds when the top-k sentences fully contain at least one group, rather than requiring the union of every annotator rationale. This rule matches the operational requirement that one sufficient justification is enough to ground a decision.

Ranking and normalization details. Document scores are calculated over the entire 5,183-paper corpus before the highest-scoring candidates are selected. Min-max normalization is performed independently for each query and each branch, preventing the larger numeric range of either the latent-semantic or graph score from determining the fusion. Sentence ranking is restricted to sentences belonging to the retrieved papers, which preserves a two-stage paper-to-rationale workflow and avoids comparing every sentence in the corpus for every claim. Ties are resolved by the deterministic order returned by the score sort, and every cutoff uses the same ranked list. The method therefore changes only the representation and fusion signal, not the candidate budget or evaluation protocol.

Statistical unit and aggregation. Fold-level tables report the arithmetic mean and sample standard deviation across the five held-out folds. Paired tests instead operate on the 1,109 out-of-fold claim outcomes, preserving the within-claim comparison between methods. Bootstrap intervals resample claims with replacement, and the Wilcoxon test evaluates the signed paired differences without assuming normality. The evidence-success test excludes NEI claims because their lack of a gold rationale would otherwise create automatic ties. Grounded-accuracy tests retain NEI because a correct NEI decision is a valid end-to-end outcome. This separation keeps the inferential unit aligned with each metric's definition.

Error analysis assigns every failed Hybrid grounded prediction to one mutually exclusive category. A failure is first labeled as a document retrieval miss when no gold paper appears in the top five, then as a multi-sentence or single-sentence ranking miss when the paper is present but a complete rationale is absent. Remaining failures are classified by quantitative expressions, explicit negation, or other semantic-relation errors. This ordering separates acquisition failures from stance failures and prevents a later linguistic category from masking an earlier retrieval breakdown. Retrieving normal prototypes for OpenStack failure-injection logs similarly uses contrastive evidence to make anomaly explanations inspectable [41].

Implementation control. All random operations use seed 2026. The index and evaluation are implemented with NumPy, SciPy, scikit-learn, pandas, and NetworkX. The vector and graph branches share the same corpus, query text, document cutoff, sentence cutoff, and five-fold partitions. No external abstracts, knowledge-base triples, pretrained embedding endpoints, or manually written query expansions enter the rankings. These controls make the observed differences a direct test of vector geometry, graph propagation, and their fusion. Operational workflows in AIOps [42] and LLM-assisted CI diagnosis and repair [43] further motivate retaining run-level outputs that connect a detected failure to its supporting observations and proposed action.

3. Results and Discussion

Corpus and graph scale. Table I shows that the evaluation covers all 1,109 labeled claims and 1,391 annotated evidence-sentence instances. The class distribution is imbalanced: CONTRADICT accounts for only 21.4% of claims, while SUPPORT and NEI account for 41.1% and 37.5%. Table II shows that sentence-entity links form the largest edge family (340,949), followed by document-entity links (72,560) and document-sentence membership (45,952). The average paper retains 14.00 entity phrases and the average sentence retains 7.42. This sparsity keeps the graph interpretable while providing repeated paths between claims and papers.

Figure 1 presents the complete pipeline. A claim enters parallel vector and graph branches, the document scores are fused when the Hybrid method is selected, and sentence evidence is ranked within the retrieved papers before label prediction. Figure 2 makes the heterogeneous path explicit: a claim mentions entity phrases; those phrases occur in papers and sentences; papers contain sentences. The graph therefore preserves both semantic co-occurrence and document structure. It does not assert that every shared phrase proves a relation. Instead, the graph supplies a structured candidate-ranking signal whose validity is tested against annotated evidence.

Document retrieval. Table IV and Fig. 3 show a consistent advantage for fusion. Hybrid achieved MRR 0.5835,

recall@1 0.4661, recall@3 0.6739, recall@5 0.7519, and recall@10 0.8211. At $k = 5$, the absolute gain was 0.1098 over Vector RAG and 0.0865 over GraphRAG. The improvement is not confined to deep retrieval: Hybrid also increased recall@1 by 0.0665 over the vector baseline.

TABLE IV. Five-fold document retrieval performance

Method	MRR	R@1	R@3	R@5	R@10
Vector RAG	0.5059	0.3996	0.5641	0.6421	0.7490
GraphRAG	0.5025	0.3810	0.5961	0.6654	0.7605
Hybrid	0.5835	0.4661	0.6739	0.7519	0.8211

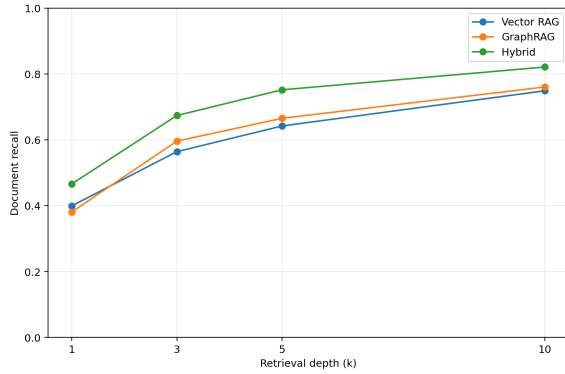


Fig. 3. Document recall as retrieval depth increases.

The two single-branch systems exhibit complementary ranking behavior. Vector RAG produced slightly higher MRR (0.5059) and recall@1 (0.3996) than GraphRAG (0.5025 and 0.3810). GraphRAG overtook Vector RAG at recall@3, recall@5, and recall@10, reaching 0.7605 at ten papers compared with 0.7490. This pattern matches the intended role of propagation: direct dense similarity is effective for the first result, while shared entity neighborhoods broaden coverage as the candidate list expands. Complementarity between structural and semantic channels is analogous to computer-vision systems that combine receptive-field adaptation with attention rather than relying on either mechanism alone [44].

Evidence retrieval. The document gain translated into stronger rationale recovery. Hybrid complete evidence-set recall was 0.2409 at one sentence, 0.4010 at three, 0.4645 at five, and 0.5843 at ten. At $k = 5$, Hybrid exceeded Vector RAG by 0.1082 and GraphRAG by 0.0474. Hybrid sentence F1@5 reached 0.1947, compared with 0.1764 for GraphRAG and 0.1375 for Vector RAG. Table V and Fig. 4 therefore support the retrieval hypothesis at both document and sentence levels.

TABLE V. Complete evidence-set and sentence retrieval

Method	Set R@1	Set R@3	Set R@5	Set R@10	Sentence F1@5
Vector RAG	0.1904	0.2986	0.3563	0.4588	0.1375
GraphRAG	0.1559	0.3261	0.4171	0.5383	0.1764
Hybrid	0.2409	0.4010	0.4645	0.5843	0.1947

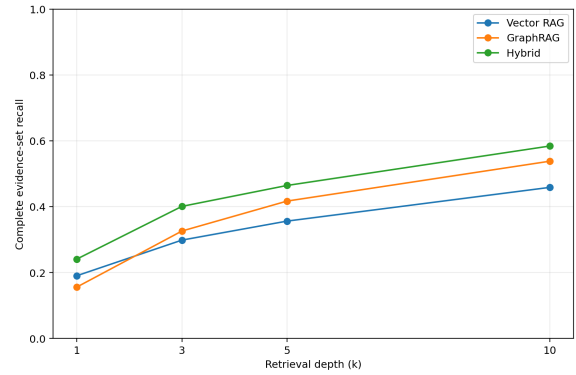


Fig. 4. Complete evidence-set recall across sentence cutoffs.

GraphRAG alone was especially effective at deeper evidence cutoffs. Its evidence-set recall@10 was 0.5383, an absolute gain of 0.0795 over Vector RAG. Its sentence F1@5 exceeded the vector score by 0.0390. At one sentence, however, Vector RAG recovered complete evidence sets more often than GraphRAG. Graph propagation therefore improves breadth rather than guaranteeing that the first selected sentence is the decisive rationale.

Verification accuracy. Table VI and Fig. 5 reveal a different ordering from retrieval. The claim-only classifier achieved the highest label accuracy (0.4860) and macro-F1 (0.4518). Among retrieval-augmented systems, GraphRAG was strongest, with accuracy 0.4598, macro-F1 0.4259, and strict grounded accuracy 0.3462. Hybrid achieved accuracy 0.4445, macro-F1 0.4115, and grounded accuracy 0.3444. Vector RAG was lowest on these three metrics. Better retrieval therefore did not automatically produce better three-way classification under the fixed linear verifier. Comparable separations between prediction and evidential interpretation arise in cancer image classification [45] and multiple-myeloma minimal-residual-disease prediction [46]: a correct class decision is not itself a traceable rationale.

TABLE VI. Three-way verification and strict grounding

Method	Accuracy	Macro-F1	SUPPORT F1	CONTRADICT F1	NEI F1	Grounded accuracy
Claim only	0.4860	0.4518	0.4359	0.2234	0.6962	—

Method	Accuracy	Macro-F1	SUPPORT F1	CONTRADICT F1	NEI F1	Grounded accuracy
Vector RAG	0.4391	0.4011	0.3721	0.1545	0.6769	0.3273
Graph RAG	0.4598	0.4259	0.3982	0.1840	0.6954	0.3462
Hybrid	0.4445	0.4115	0.3746	0.1813	0.6785	0.3444

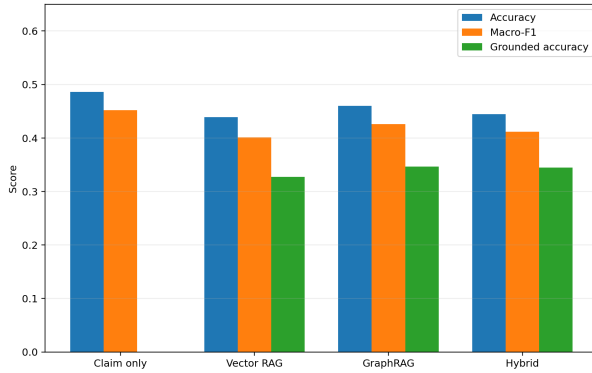


Fig. 5. Label and grounding metrics for the four verification conditions.

The claim-only result does not invalidate evidence retrieval. It exposes label regularities that the classifier can exploit directly from claim wording, particularly for NEI. Claim-only NEI F1 was 0.6962, whereas its CONTRADICT F1 was only 0.2234. Retrieved context adds many medically and scientifically relevant terms, but a bag-of-ngrams verifier cannot reliably determine whether those terms express support, contradiction, comparison, or an unrelated association. GraphRAG's advantage over Vector RAG within the retrieval-augmented conditions shows that cleaner structural context helps, yet the remaining gap requires a verifier that represents directional and quantitative relations rather than only token evidence. Similar concerns arise in bank credit-risk models, where a label can be operationally consequential even when the predictive rule is comparatively simple [47].

Grounding depth. Figure 6 separates label prediction from evidence coverage by holding out-of-fold labels fixed while increasing the number of available sentences. Hybrid grounded accuracy rose from 0.3066 with one sentence to 0.3733 with ten. GraphRAG rose from 0.2949 to 0.3715, and Vector RAG rose from 0.2958 to 0.3490. Hybrid is best at one, three, seven, and ten sentences, while GraphRAG is slightly best at five. The monotonic curves confirm that missing rationales remain a major limitation even when the predicted stance is unchanged. Evidence depth also matters in data-rich financial services, where customer-insight and

risk models combine heterogeneous signals and require a reviewable basis for action [48].

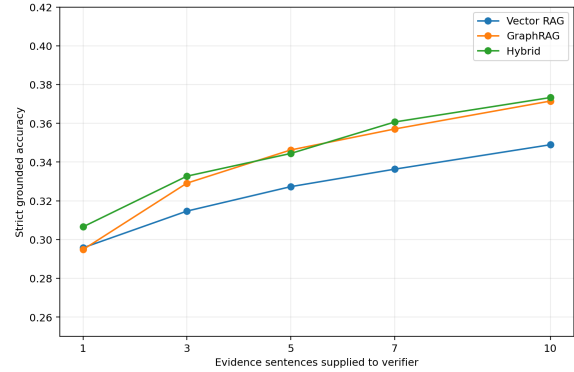


Fig. 6. Retrieval depth-grounded accuracy relationship.

The depth curve also establishes a practical cost-coverage tradeoff. Moving from five to ten sentences increases Hybrid grounded accuracy by approximately 0.0289 and GraphRAG by approximately 0.0253, but doubles the evidence text passed to the verifier. For a system that must display evidence to a human reviewer, the five-sentence setting offers a compact operating point; for an automated verifier with a larger context budget, the ten-sentence setting recovers additional complete rationales. The result argues for adaptive depth based on confidence or evidence completeness rather than a single fixed cutoff for every claim. The same design principle appears in personalized recommendation, where added modeling complexity must justify operational utility [49]. It is explicit in risk-aware auto-bidding [50] and constrained marketing-budget allocation [51], which optimize performance under hard resource limits.

Graph ablation. Table VII shows that diffusion improves standalone GraphRAG. Adding the one-step diffusion to direct entity paths raises document recall@5 from 0.6537 to 0.6652, evidence-set recall@5 from 0.4055 to 0.4170, and sentence F1@5 from 0.1721 to 0.1766. The effect is small but consistent across the three retrieval measures. Diffusion therefore recovers papers connected through neighboring entity profiles rather than only through direct claim phrase overlap.

TABLE VII. Graph propagation and fusion ablation

Variant	Doc R@5	Evidence-set R@5	Sentence F1@5
Direct entity paths	0.6537	0.4055	0.1721
Direct diffusion +	0.6652	0.4170	0.1766
Hybrid without diffusion	0.7532	0.4661	0.1944
Full hybrid	0.7518	0.4646	0.1949

The hybrid ablation is more nuanced. Hybrid without diffusion produced document recall@5 of 0.7532 and evidence-set recall@5 of 0.4661, slightly above the full Hybrid values of 0.7518 and 0.4646 in the pooled ablation. Full Hybrid produced the marginally higher sentence F1@5, 0.1949 versus 0.1944. Once vector similarity is present, the diffused document signal is partly redundant and can reorder a small number of already relevant papers. The standalone graph benefit and the fused result are therefore logically compatible: diffusion strengthens the graph branch but is not uniformly beneficial after score fusion.

Paired significance. Table VIII reports claim-level paired tests. Hybrid exceeded Vector RAG on verifiable-claim evidence success by 0.1082 (95% bootstrap CI [0.0779, 0.1385]; bootstrap $p = 0.0002$; Wilcoxon $p = 2.83e-11$). Its grounded-accuracy gain over Vector RAG was 0.0171, but the interval crossed zero and the Wilcoxon test was not significant; Hybrid and GraphRAG were also tied on grounded accuracy. Thus, complete evidence retrieval improved clearly without a corresponding label-interpretation gain. Paired evaluation is likewise used in enterprise marketing prediction [52], while counterfactual ranking highlights policy bias in logged comparisons [53].

TABLE VIII. Paired out-of-fold significance tests

Comparison	Mean difference	95% CI low	95% CI high	Bootstrap p	Wilcoxon p
Hybrid vs. Vector (grounded accuracy)	0.0171	-0.0072	0.0415	0.1664	0.1556
Hybrid vs. GraphRAG (grounded accuracy)	-0.0018	-0.0243	0.0207	0.9256	0.8744
Hybrid vs. Vector (evidence-set success)	0.1082	0.0779	0.1385	0.0002	2.83e-11

Efficiency. Table IX reports uncached end-to-end retrieval on the same 200 claims. Mean latency was 17.33 ms for Vector RAG, 20.40 ms for GraphRAG, and 18.36 ms for Hybrid. The 95th-percentile values were close across methods. Vector index construction required 8.46 s and graph construction required 6.00 s. On this corpus, structural retrieval adds no material query-latency penalty because both branches use sparse or matrix-vector operations over only 5,183 papers. Hybrid computes both scores but remains in the same latency range. These deployment observations are consistent with hybrid-cloud LLM architectures that allocate

work by resource and cost constraints [54]. Related edge and document systems compress quality signals [55], prioritize PDF content to improve time-to-useful rendering [56], and use offline decision policies for urban dispatch [57]; each treats latency as part of model quality rather than a separate systems statistic.

TABLE IX. Uncached retrieval efficiency and index construction

Method	Mean latency (ms)	P95 latency (ms)	Queries/s	Vector build (s)	Graph build (s)
Vector RAG	17.33	35.79	57.72	8.46	0.00
GraphRAG	20.40	39.09	49.01	0.00	6.00
Hybrid	18.36	36.17	54.48	8.46	6.00

Robustness. Figure 7 and Table X show retrieval after deterministic deletion of 20% of non-negation content tokens. Vector RAG recall@5 fell from 0.6421 to 0.5830, a drop of 0.0592. GraphRAG had the smallest drop, 0.0476, consistent with multiple entity paths distributing relevance across the remaining terms. Hybrid had the largest absolute drop, 0.0678, yet still retained the highest perturbed recall@5, 0.6840. The fusion is strongest in absolute performance, while the graph branch is most stable relative to its own clean score. Robustness under changing traffic is also central to learned DDoS mitigation [58], although the perturbations here target claim wording rather than network behavior.

TABLE X. Document retrieval after 20% content-token deletion

Method	Original R@5	Perturbed R@5	Absolute drop
Vector RAG	0.6421	0.5830	0.0592
GraphRAG	0.6652	0.6176	0.0476
Hybrid	0.7518	0.6840	0.0678

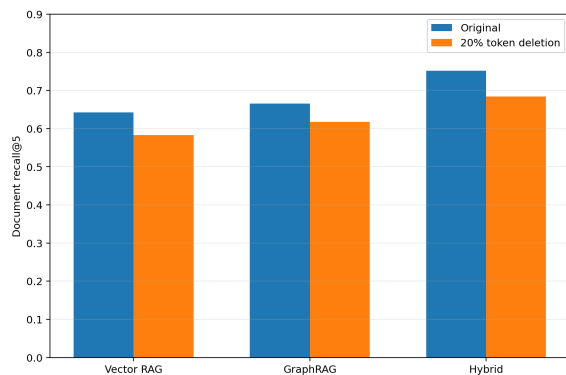


Fig. 7. Clean and perturbed document recall@5.

Confusion and error structure. Figure 8 shows that the Hybrid verifier correctly classified 282 of 416 NEI claims, 47 of 237 CONTRADICT claims, and 164 of 456 SUPPORT claims. The largest off-diagonal cells are CONTRADICT predicted as SUPPORT (160) and SUPPORT predicted as CONTRADICT (189). This polarity exchange explains the low CONTRADICT and SUPPORT F1 values and confirms that retrieving a related paper is not equivalent to resolving the direction of its finding.

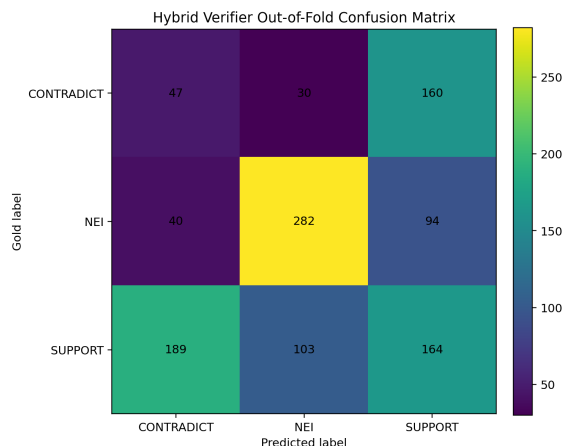


Fig. 8. Hybrid out-of-fold confusion matrix.

Table XI assigns 727 failed grounded predictions to their first decisive error. Semantic-relation errors account for 25.03%, sentence-ranking misses for 23.66%, document retrieval misses for 23.52%, and quantitative-relation errors for 19.81%. Multi-sentence evidence misses account for 4.40%, and explicit negation or polarity cues account for 3.58%. Acquisition failures therefore represent nearly half of the errors, while relation and quantity interpretation represent another large share. The remaining research problem is not a single bottleneck; it is the interaction between candidate coverage and fine-grained inference.

TABLE XI. First decisive error category for failed Hybrid grounded predictions

Error category	Count	Percentage
semantic relation error	182	25.03
sentence ranking miss	172	23.66
document retrieval miss	171	23.52
quantitative relation error	144	19.81
multi-sentence evidence miss	32	4.40
negation or polarity error	26	3.58

Fold stability. The principal retrieval gains were not driven by a single split. Hybrid document recall@5 had a five-fold standard deviation of 0.0403, compared with 0.0376 for Vector RAG and 0.0677 for GraphRAG. Hybrid MRR varied by 0.0334 across folds, which is lower than the graph MRR variation of 0.0573. Verification scores were also comparatively stable: Hybrid macro-F1 standard deviation was 0.0168, and GraphRAG macro-F1 standard deviation was 0.0316. The consistent selection of $\alpha = 0.6$ in all five training folds provides an additional stability signal, because each fold independently preferred the same semantic-structural balance. Temporal forecasting with news and macroeconomic fusion [59] and cross-domain learning with approximately shared features [60] illustrate the same concern at a broader level: stability on one split does not establish transfer to a shifted distribution.

4. Limitations

The graph uses corpus-derived keyphrases rather than normalized biomedical entities. Synonyms can remain disconnected, and a phrase shared by several papers can represent different concepts in different contexts. Degree normalization reduces hub effects but does not perform entity disambiguation. An ontology-linked graph would test whether canonical identifiers strengthen the same structural advantage, but it would also introduce external knowledge and linking errors that are absent from the current comparison. Because the corpus and keyphrase extraction are English-only, the design also leaves language accessibility untested; evidence from bilingual computer-science instruction shows that language choice can affect learning and engagement [61].

The verifier is a balanced linear classifier over word and character n-grams. This choice isolates retrieval differences and keeps every prediction auditable, but it underrepresents causal direction, numerical comparison, study population, and scope. The confusion matrix and error taxonomy show that these relation-sensitive cases dominate residual classification failures. A transformer-based natural-language-inference verifier or a constrained language model would use the retrieved evidence more fully, but its gains would combine retrieval quality with reader capacity.

SciFact contains paper titles and abstracts rather than full articles, tables, figures, appendices, or supplementary material. The graph therefore cannot follow evidence that depends on methods details or results reported outside the abstract. The corpus also contains 5,183 papers, far fewer than an operational scientific search index. Query latency and graph density will change at million-document scale, where approximate nearest-neighbor search, partitioned sparse propagation, and incremental graph updates become necessary.

The error taxonomy is rule-based and assigns only one first decisive category to each failure. It separates clear document and sentence acquisition misses from surface indicators of quantitative or negation reasoning, but it does not replace expert qualitative annotation. A claim containing a number can fail for a nonquantitative reason, and a semantic-relation category can contain several distinct phenomena. The category counts are therefore diagnostic summaries tied to explicit rules rather than estimates of a universal causal distribution. Human review of a stratified error sample would refine the taxonomy and measure agreement between analysts.

The comparison intentionally excludes pretrained dense scientific encoders and large generative verifiers. This control makes the graph contribution identifiable, but it also limits absolute performance relative to systems that use SciBERT, T5, or larger retrieval-augmented readers. The reported gains establish that structure improves a transparent local representation; they do not establish that the same numeric advantage persists after every branch is replaced with a stronger neural component. A subsequent factorial experiment should cross vector representation strength, graph structure, and verifier capacity rather than changing all three simultaneously.

The robustness test deletes content tokens deterministically and preserves explicit negation. It measures sensitivity to incomplete wording, not paraphrase, terminology substitution, adversarial entity insertion, or domain shift. Complete evidence-set recall is evaluated against the available annotations, so unannotated but valid rationales receive no credit. Finally, results are limited to SciFact's scientific domains and class distribution; the same fusion weight and graph sparsity require independent evaluation on other evidence-verification collections.

5. Conclusion

This study evaluated Vector RAG, GraphRAG, and Hybrid graph–vector retrieval on every labeled SciFact claim through five-fold out-of-fold experiments. The Hybrid method achieved document recall@5 of 0.7519 and complete evidence-set recall@5 of 0.4645, improving the corresponding Vector RAG scores by 0.1098 and 0.1082. GraphRAG produced the best retrieval-augmented accuracy (0.4598) and grounded accuracy (0.3462), while the Hybrid

evidence advantage over Vector RAG was statistically significant. The graph branch also showed the smallest relative loss after query-token deletion.

For system builders, the findings support an evidence-first architecture. The retriever should expose both semantic similarity and graph paths, retain enough depth to recover complete rationales, and pass sentence-level scores and provenance to the verifier. The verifier should then decide stance over the smallest sufficient evidence set rather than over an undifferentiated block of context. This division preserves the measurable retrieval advantage of Hybrid GraphRAG while directing model capacity toward the relation and quantity errors that dominate the remaining failures. Human-facing delivery should likewise organize provenance and uncertainty rather than expose raw scores alone. Financial risk dashboards [62] and retrieved-evidence explanation cards for traffic-sign alerts [63] provide complementary examples of visual hierarchy and localized evidence presentation.

The results define a clear boundary for GraphRAG in scientific verification. Claim–document–entity structure is valuable for finding papers and assembling complete rationales, especially beyond the first result. It does not by itself solve stance interpretation, and additional retrieved text can degrade a verifier that lacks relation-sensitive reasoning. A practical scientific fact-checking system should therefore combine hybrid structural retrieval with a verifier designed for polarity, quantities, study conditions, and multi-sentence inference.

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*, 2020, pp. 7534–7550, doi: 10.18653/v1/2020.emnlp-main.609.
- [2] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, “FEVER: A Large-Scale Dataset for Fact Extraction and VERification,” in *Proc. NAACL-HLT*, 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*, 2020, pp. 7740–7754, doi: 10.18653/v1/2020.emnlp-main.623.
- [4] Saakyan, T. Chakrabarty, and S. Muresan, “COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19,” in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 2116–2129, doi: 10.18653/v1/2021.acl-long.165.
- [5] T. Diggelmann, J. Boyd-Graber, J. Bulian, M. Ciaramita, and M. Leippold, “CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims,” *arXiv:2012.00614*, 2020.

- [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, doi: 10.18653/v1/2023.findings-acl.387.
- [7] R. Pradeep, X. Ma, R. Nogueira, and J. Lin, "Scientific Claim Verification with VerT5erini," in *Proc. 12th Int. Workshop Health Text Mining and Information Analysis*, 2021, pp. 94–103, doi: 10.18653/v1/2021.louhi-1.11.
- [8] Z. Zhang, J. Li, F. Fukumoto, and Y. Ye, "Abstract, Rationale, Stance: A Joint Model for Scientific Claim Verification," in *Proc. 2021 Conf. Empirical Methods in Natural Language Processing*, 2021, pp. 3580–3586, doi: 10.18653/v1/2021.emnlp-main.290.
- [9] D. Wadden, K. Lo, L. L. Wang, S. Lin, M. van Zuylen, A. Cohan, and I. Beltagy, "MultiVerS: Improving Scientific Claim Verification with Weak Supervision and Full-Document Context," in *Findings of NAACL 2022*, 2022, pp. 61–76, doi: 10.18653/v1/2022.findings-naacl.6.
- [10] 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.
- [11] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing*, 2020, pp. 6769–6781, doi: 10.18653/v1/2020.emnlp-main.550.
- [12] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," in *Proc. 37th Int. Conf. Machine Learning*, 2020, pp. 3929–3938.
- [13] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering," in *Proc. 16th Conf. European Chapter Assoc. Comput. Linguistics*, 2021, pp. 874–880, doi: 10.18653/v1/2021.eacl-main.74.
- [14] G. Izacard et al., "Atlas: Few-Shot Learning with Retrieval Augmented Language Models," *Journal of Machine Learning Research*, vol. 24, no. 251, pp. 1–43, 2023.
- [15] F. Petroni et al., "KILT: A Benchmark for Knowledge Intensive Language Tasks," in *Proc. NAACL-HLT, 2021*, pp. 2523–2544, doi: 10.18653/v1/2021.naacl-main.200.
- [16] Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," in *Proc. EMNLP-IJCNLP, 2019*, pp. 3615–3620, doi: 10.18653/v1/D19-1371.
- [17] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," in *Proc. EMNLP-IJCNLP, 2019*, pp. 3982–3992, doi: 10.18653/v1/D19-1410.
- [18] H. Sun, B. Dhingra, M. Zaheer, K. Mazaitis, R. Salakhutdinov, and W. Cohen, "Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text," in *Proc. 2018 Conf. Empirical Methods in Natural Language Processing*, 2018, pp. 4231–4242, doi: 10.18653/v1/D18-1455.
- [19] H. Sun, T. Bedrax-Weiss, and W. Cohen, "PullNet: Open Domain Question Answering with Iterative Retrieval on Knowledge Bases and Text," in *Proc. EMNLP-IJCNLP, 2019*, pp. 2380–2390, doi: 10.18653/v1/D19-1242.
- [20] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec, "QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering," in *Proc. NAACL-HLT, 2021*, pp. 535–546, doi: 10.18653/v1/2021.naacl-main.45.
- [21] M. Yasunaga, A. Bosselut, H. Ren, X. Zhang, C. D. Manning, P. Liang, and J. Leskovec, "Deep Bidirectional Language-Knowledge Graph Pretraining," in *Proc. Int. Conf. Learning Representations*, 2022.
- [22] D. Yu, C. Zhu, Y. Fang, W. Yu, S. Wang, Y. Xu, X. Ren, Y. Yang, and M. Zeng, "KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 4961–4974, doi: 10.18653/v1/2022.acl-long.340.
- [23] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *Proc. Int. Conf. Learning Representations*, 2017.
- [24] P. W. Battaglia et al., "Relational Inductive Biases, Deep Learning, and Graph Networks," *arXiv:1806.01261*, 2018.
- [25] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Text," in *Proc. 2004 Conf. Empirical Methods in Natural Language Processing*, 2004, pp. 404–411.
- [26] 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.
- [27] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.
- [28] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," *Stanford InfoLab, Technical Report 1999-66*, 1999.
- [29] 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.

- [30] 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.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] 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.
- [37] Q. Xin, "Uncertainty-Aware Late Fusion for 3D Perception (Confidence Calibration + Fusion Rule Learning)," *J. Technol. Informatics Eng.*, vol. 4, no. 1, pp. 215–238, 2025, doi: 10.51903/jtie.v4i1.485.
- [38] Z. S. Zhong and S. Ling, "Uncertainty Quantification of Spectral Estimator and MLE for Orthogonal Group Synchronization," *arXiv:2408.05944*, Aug. 2024.
- [39] 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, pp. 64–70, 2024, doi: 10.54254/2755-2721/69/20241511.
- [40] 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.
- [41] 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.
- [42] 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.
- [43] 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, 2025, doi: 10.51903/jtie.v4i1.484.
- [44] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of Autonomous Driving Image Detection Based on RFACnv and Triplet Attention," *Appl. Comput. Eng.*, vol. 77, pp. 210–217, 2024, doi: 10.54254/2755-2721/77/2024MA0067.
- [45] 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.
- [46] 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.
- [47] 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.
- [48] 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.
- [49] 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, pp. 147–153, 2024, doi: 10.54254/2755-2721/64/20241365.
- [50] 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.
- [51] 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.
- [52] 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.

- [53] 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, 2025, doi: 10.69987/JACS.2025.51201.
- [54] 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.
- [55] 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.
- [56] 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.
- [57] 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.
- [58] 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, pp. 94–99, 2024, doi: 10.54254/2755-2721/64/20241350.
- [59] 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.
- [60] Z. S. Zhong, X. Pan, and Q. Lei, "Bridging Domains with Approximately Shared Features," in *Proc. 28th Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, PMLR, vol. 258, 2025, pp. 559–567.
- [61] 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, Art. no. 24, pp. 1–10, doi: 10.1145/3364510.3364518.
- [62] 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 Design*, vol. 3, no. 1, pp. 196–210, May 2025, doi: 10.51903/ijgd.v3i1.3715.
- [63] 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 Design*, vol. 3, no. 2, pp. 455–473, Oct. 2025, doi: 10.51903/ijgd.v3i2.3991.