Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Claim-Aware Scientific RAG: Evidence-First Retrieval and Abstention for Scientific Fact Responses on SciFact

Jing Chen[1], *Xinzhuo Sun*[2], *Vincent Brown*[3]
[1]*Industrial Engineering and Operations Research, UCB, CA, USA*
[2]*Computer Science, Cornell Tech, NY, USA*
[3]*Information Technology, Illinois Tech, IL, USA*
jingc0606@gmail.com

DOI: 10.69987/JACS.2023.30102

Keywords

RAG; fact verification; evidence-grounded generation; hallucination reduction; SciFact; BEIR; abstention; reranking; hybrid retrieval

Abstract

Retrieval-augmented generation (RAG) is widely adopted to reduce hallucinations, yet most systems still answer even when retrieval fails, producing fluent but unsupported "scientific facts". This paper studies a claim-aware scientific RAG design principle: the system is allowed to answer only when it can cite evidence. We conduct full experimental evaluations on the SciFact scientific claim retrieval task using the BEIR-style SciFact split (5,183 abstracts; 809 training claims; 300 test claims). We compare a sparse BM25 retriever, a contrastive dense dual-encoder, and a hybrid retriever using reciprocal rank fusion (RRF), followed by an interaction-based reranker. We then add an evidence layer that extracts candidate citation sentences and scores them with a lightweight verifier, and we enforce an abstention gate that refuses to answer when confidence is low. On the SciFact test set, BM25 achieves nDCG@10=0.662 and Recall@100=0.883. The dense retriever alone underperforms (nDCG@10=0.537), but hybrid RRF improves Recall@100 to 0.923 and a reranker recovers nDCG@10 to 0.659. For evidence extraction, token-level evidence F1 reaches 0.190 when selecting two sentences. Finally, we quantify a refusal–hallucination tradeoff via confidence-based abstention: gating by the top-1 BM25 score reduces the rate of answers without any relevant abstract in the top-10 from 0.193 to 0.047 at 28.3% answer coverage. These results provide a reproducible baseline showing how evidence-first retrieval and calibrated refusal can be combined to control hallucinations in scientific RAG.

1. Introduction

Large pretrained language models have enabled interactive systems that can answer questions, summarize papers, and explain scientific concepts. Despite these capabilities, language models do not inherently provide guarantees of factuality: they can generate fluent statements that are not supported by any underlying evidence. In scientific and biomedical contexts, this behavior is risky because users often interpret responses as "scientific facts", and small errors can lead to wrong decisions. Hallucination has been documented in neural text generation even when the model conditions on an input document; models can introduce unsupported entities, relationships, and numbers [14]. A central research direction is therefore evidence-grounded generation: constrain outputs so that they can be traced to retrieved sources.

Retrieval-augmented generation (RAG) operationalizes grounding by combining a retriever with a generator. The retriever fetches relevant documents from a corpus, and the generator conditions on these documents when producing an answer [6]. RAG is particularly attractive for scientific domains, where the corpus (papers, abstracts, clinical guidelines) evolves and cannot be fully memorized. However, RAG does not automatically eliminate hallucinations. If retrieval fails to retrieve relevant evidence, the generator can still produce a plausible answer; if the system attaches citations, those citations may point to irrelevant documents. For this reason, retrieval quality and evidence attribution are as important as generation quality.

Scientific claim verification tasks provide a natural testbed for claim-aware RAG. In these tasks, the input is an atomic claim (often one sentence), and the system must retrieve evidence from scientific documents to support or refute the claim. The SciFact dataset contains expert-written scientific claims paired with abstracts and includes evidence rationales in the original release [2]. SciFact is related to general fact verification benchmarks such as FEVER [13], but differs in domain, style, and evidence sources. SciFact has a retrieval component (find relevant abstracts) and a rationale selection component (find evidence sentences) [2]. In a RAG setting, both are required: without retrieving the right abstracts, the system cannot be grounded, and without identifying the right sentences, citations become vague.

Information retrieval research provides strong baselines and evaluation methodology for these tasks. Sparse term-matching retrieval methods such as BM25 remain surprisingly strong across datasets, as highlighted by the BEIR benchmark [1]. BM25's robustness comes from lexical matching and inverse document frequency weighting [3], [21]. In parallel, dense retrieval has become a standard approach for semantic retrieval. Dense retrievers represent queries and documents as vectors and score candidates by similarity, commonly dot product. DPR popularized a supervised dual-encoder trained with in-batch negatives [7]. More recent work showed that unsupervised contrastive pretraining can produce dense retrievers that transfer across domains (Contriever [4]), and that weakly supervised contrastive pretraining over large text pairs can yield strong general-purpose embeddings (E5 [5]). These models and others are often evaluated on BEIR and MTEB [1], [20].

Hybrid retrieval combines sparse and dense approaches to leverage complementary strengths. Sparse methods are precise for lexical cues (e.g., gene names, drug dosages), while dense methods capture paraphrases and broader semantics. Fusion methods such as reciprocal rank fusion (RRF) are widely used because they are simple and effective without requiring score calibration between systems [9]. However, high recall alone is insufficient for RAG. In citation-based answering, the system must place truly relevant evidence near the top of the ranking, because generators typically condition on a small number of retrieved documents (often 3–10). Reranking is therefore standard: after retrieving a candidate set, a stronger interaction model reorders candidates using richer query–document features. Cross-encoder rerankers based on BERT-style models are known to improve ranking quality [8], but are computationally expensive because they process each query–document pair jointly. Architectures such as ColBERT reduce cost by delaying interaction [10]. Even with strong retrieval and reranking, a remaining safety issue is what the system does when evidence is absent or low confidence. Many deployed assistants attempt to always answer, which makes hallucination inevitable whenever retrieval fails. A more principled approach is selective answering (abstention): answer only when confidence is high and refuse otherwise. Selective classification formalizes this tradeoff between risk and coverage and provides methods to set rejection thresholds for a desired risk level [15]. Making confidence meaningful often requires calibration; temperature scaling is a simple and effective calibration technique [16]. In a claim-aware RAG system, abstention can be based on retrieval confidence signals (e.g., retrieval scores, margin between top results) and on evidence consistency signals (e.g., a verifier score that tests whether evidence entails the claim).

This paper studies an evidence-first scientific RAG pipeline under a strict policy: the system is only allowed to answer when retrieval returns evidence with sufficient confidence; otherwise it refuses. We focus on measurable properties of the evidence pipeline rather than the open-ended generation step. Concretely, we evaluate: (i) document retrieval quality (nDCG@10 and Recall@100), (ii) sentence-level evidence extraction quality (evidence F1), and (iii) refusal and hallucination rates under confidence-based gating. We emphasize reproducibility: all results are computed from the released SciFact retrieval files and the explicit hyperparameters reported in this paper [23-36].

Our contributions are as follows. First, we conduct full experimental evaluations on the SciFact retrieval split using a controlled comparison between BM25, a contrastive dense retriever, and a hybrid retriever with RRF fusion, followed by a reranker. Second, we implement a sentence-level evidence module that extracts candidate citations and evaluates evidence matching using reproducible pseudo-gold targets derived from the corpus. Third, we quantify a refusal–hallucination tradeoff and show that simple confidence gating can substantially reduce the frequency of answers produced without any relevant retrieved abstract. These contributions provide a baseline for future work on claim verification and evidence-grounded scientific RAG.

Relation to language-model pretraining for science. Many scientific RAG systems rely on pretrained encoders that better capture scientific terminology and discourse. SciBERT adapts BERT pretraining to scientific text and consistently improves scientific NLP tasks [18]. Citation-aware pretraining further improves document representations: SPECTER learns paper embeddings from citation graphs and has been used to improve retrieval and clustering in scientific corpora [19]. These models motivate a practical view of

scientific RAG: high-quality retrieval depends on both lexical matching and domain-appropriate semantic representations. In our experiments we isolate retrieval and abstention mechanisms using lightweight models; the same protocol can directly evaluate stronger pretrained encoders in future work.

Relation to ranking architectures. IR systems typically operate in stages: a fast first-stage retriever generates candidates, and a more expensive ranker reorders them. Cross-encoders based on Transformers [17] are known to be strong rerankers because they jointly attend over query and document tokens [8], [11]. They are, however, expensive, which has motivated late-interaction models such as ColBERT [10] that precompute document embeddings and perform efficient interaction at query time. Our reranker is an interaction model implemented with interpretable features and logistic regression; it captures key interaction signals (lexical overlap and TF-IDF similarity) while remaining fast and reproducible.

Why abstention is essential for evidence-grounded scientific RAG. Even with strong retrieval, the system must decide whether retrieved evidence is sufficient to justify answering. This decision can be framed as risk control. Selective classification provides a formal guarantee: given a confidence function, the system can reject low-confidence inputs to achieve a desired risk level [15]. In generation settings, risk corresponds to producing an unsupported answer. Calibration techniques, including temperature scaling [16], make confidence estimates more reliable. Our evaluation uses a direct and transparent confidence signal derived from retrieval scores and reports the complete refusal–hallucination curve so that practitioners can select an operating point consistent with their application's risk tolerance.

Scope of evaluation. The experiments in this paper focus on measurable evidence properties: document retrieval accuracy, evidence sentence matching, and abstention behavior. We intentionally do not tune a large generator, because generation introduces additional variables (prompting, decoding, model choice) that can obscure the contribution of the evidence pipeline. By isolating evidence-first retrieval and refusal, we provide a clear foundation for integrating any generator: the generator is invoked only when the pipeline has retrieved high-confidence evidence, and the output can cite the extracted evidence sentences.

Evaluation perspective. Evidence-first scientific RAG needs evaluation signals that reflect both retrieval and risk. In IR, nDCG measures how well a system places relevant documents early in the ranking, while recall measures whether relevant documents are retrieved at all. BEIR standardized evaluation across diverse tasks and emphasized that both sparse and dense methods should be compared under the same metrics and splits [1]. In scientific domains, recall is particularly important because a single missed abstract can prevent grounding. At the same time, for citation-based answers, early precision (e.g., nDCG@10) matters because the generator typically sees only a small retrieved context. Our experiments therefore report both nDCG@10 and Recall@100 and analyze Recall@10 as an evidence-availability indicator.

Comparison to fact verification beyond retrieval. Fact verification datasets often require predicting whether a claim is supported or refuted by evidence (FEVER [13], SciFact [2]). RAG systems can be viewed as a generalization: they must retrieve evidence, select rationales, and then produce an answer that is faithful to the evidence. This paper focuses on the retrieval and abstention components that are necessary for verification-style grounding; the same evaluation pipeline can be extended with a stance classifier or a generator to form a complete verifier.

Method

This section specifies the dataset, preprocessing, retrieval methods, reranking, evidence extraction, abstention policy, and evaluation metrics. All components are implemented to be reproducible on CPU and to keep the experimental protocol transparent.

Dataset and splits. We use the SciFact retrieval task in a corpus/queries/qrels format consistent with BEIR-style evaluation [1]. The corpus contains 5,183 documents, each corresponding to a scientific abstract with a title and a text field. The queries are 1,109 short scientific claims, each with an identifier and a claim string. The relevance annotations (qrels) provide binary relevance between a query and one or more abstracts. The released split contains 809 training queries with 919 relevant query–document pairs, and 300 test queries with 339 relevant pairs. Table 1 reports these counts, as well as mean token lengths.

Text preprocessing. We tokenize text using a simple alphanumeric regular expression ([a-z0-9]+) and lowercase all tokens. This choice matches common IR tokenization in lightweight baselines and keeps the implementation consistent across BM25, TF-IDF features, and the dense models. Document length statistics are computed using this tokenizer.

Sparse retrieval with BM25. BM25 is a probabilistic term-weighting model and remains a strong baseline across heterogeneous retrieval tasks [1]. We compute BM25 scores with k1=1.2 and b=0.75 [3]. We build an inverted index over the corpus tokens and evaluate each query by iterating over its terms and accumulating per-document scores. We retrieve the top-200 documents

for each query for downstream hybrid fusion and reranking.

Dense retrieval with a contrastive dual-encoder. Modern dense retrieval systems typically use pretrained transformers (e.g., BERT [11], RoBERTa [12], or domain-specific SciBERT [18]) and contrastive training objectives [7]. In this study, we implement a lightweight dense retriever that is fully trainable from scratch on CPU. The model represents a text as the mean of learned token embeddings, implemented as an EmbeddingBag mean pooling layer. We train on the SciFact training qrels using an in-batch contrastive objective. Given a batch of B query–document positive pairs, we compute query embeddings Q∈R^{B×d}, document embeddings D∈R^{B×d}, normalize them to unit length, and form logits L = QD^T/τ. We then apply cross-entropy loss with the diagonal as the correct matches (Eq. (2)). Hyperparameters are reported in Table 3. We truncate queries to 64 tokens and documents to 256 tokens for computational efficiency. After training, we encode all documents once and perform retrieval by dot product similarity.

Hybrid retrieval via reciprocal rank fusion. Sparse and dense retrieval emphasize different similarity signals. Because their score distributions are not directly comparable, we use RRF to combine their rankings without score calibration [9]. For each query, we fuse the BM25 top-200 and dense top-200 lists using Eq. (3) with k=60. We retain the top-200 fused ranking for evaluation and for reranking.

Reranking with interaction features. Cross-encoder rerankers that jointly encode the query and document often yield strong improvements in precision at small cutoffs [8]. Rather than fine-tuning a large transformer, we train a supervised interaction reranker on interpretable features. We generate training candidates by taking the BM25 top-100 documents for each training query. Each candidate is labeled 1 if it is in the qrels and 0 otherwise. We compute five features: (i) BM25 score, (ii) TF-IDF cosine similarity, (iii) Jaccard overlap between query and document token sets, (iv) document length, and (v) query length. We fit a logistic regression model with class weight=balanced to handle label imbalance. At inference time, the reranker reranks the top-50 hybrid candidates. This setup captures the reranking principle—learning query–document interactions—while remaining reproducible and efficient.

Sentence splitting and candidate evidence selection. Abstracts are segmented into sentences using punctuation-based splitting (., ?, ! followed by whitespace). We discard very short fragments (<20 characters) and keep the remaining sentences. For each query, we take the top-10 retrieved abstracts (after reranking) and select the top-3 sentences per abstract by Jaccard similarity with the claim. This candidate set provides sentence-level units that a downstream generator can cite.

Verifier model and pseudo-gold evidence. A claim-aware RAG system requires a consistency check that tests whether a candidate sentence is compatible with the claim. We train a small transformer classifier (one TransformerEncoder layer with token, position, and segment embeddings) to score (claim, sentence) pairs. Because our retrieval split provides only document-level qrels, we construct pseudo-gold evidence sentences as follows: for each relevant abstract in the qrels, we select the single sentence with maximum TF-IDF cosine similarity to the claim and treat it as the gold evidence sentence for that abstract. Positive training pairs are formed from these pseudo-gold evidence sentences. Negative pairs are formed from random sentences sampled from top BM25 non-relevant candidate abstracts. At test time, we score candidate sentences and select the top-1 or top-2 evidence sentences to cite.

Abstention policy and hallucination definition. We implement a strict “answer only with evidence” policy. The system produces an answer if and only if the retrieval confidence exceeds a threshold; otherwise it refuses to answer. We use the top-1 BM25 score as the confidence signal, and we sweep thresholds to obtain a family of operating points. We define a hallucination event as an answered query for which no relevant abstract appears in the system’s top-10 retrieved abstracts. This definition isolates retrieval failures that inevitably cause evidence-free generation.

Evaluation metrics. We report standard retrieval metrics: nDCG@10 and Recall@100. For binary relevance, DCG@k is computed as Σ {i=1..k} (2^{rel_i}-1)/log2(i+1), and nDCG@k is DCG@k normalized by the ideal DCG@k computed from the ground-truth relevance set. Recall@k is the fraction of relevant documents retrieved in the top-k. For evidence, we report: (i) exact sentence-match F1, treating each selected sentence as an item and comparing to the pseudo-gold sentence set; and (ii) token-level evidence F1, computed by taking the multiset of tokens from all selected sentences and from all pseudo-gold sentences and computing precision/recall by token overlap. Finally, we report refusal rate (fraction of queries refused) and hallucination rate (fraction of answered queries with no relevant abstract in top-10).

Reproducibility. All experiments are run with a fixed random seed (42) for model initialization and sampling. The dense retriever, reranker, and verifier hyperparameters are fully specified in Tables 3 and 4. Measured runtime breakdown is reported in Table 9.

Feature computation for reranking. TF-IDF vectors are

computed using a standard TF-IDF vectorizer over the corpus abstracts with a maximum vocabulary size of 50,000 and the same tokenization used elsewhere. TF-IDF cosine similarity for a query–document pair is computed as the normalized dot product between the query TF-IDF vector and the document TF-IDF vector. Jaccard overlap is computed between the query token set and the document token set. These features are inexpensive and capture complementary evidence signals: BM25 emphasizes exact matches and IDF weighting, TF-IDF cosine provides a global similarity view, and Jaccard overlap measures lexical coverage.

Handling label imbalance. The SciFact qrels provide very few positive documents per query, while candidate sets contain many negatives. In the reranker training set (BM25 top-100 per query), only about 1% of candidates are positive. We therefore use class_weight=balanced in logistic regression to prevent the classifier from collapsing to the negative class. This setting yields a stable reranker that improves early precision without requiring extensive hyperparameter tuning.

Candidate sizes and evaluation protocol. All retrieval methods produce a ranked list of up to 200 documents per query. We evaluate nDCG at cutoffs up to 50 and Recall up to 100. For hybrid fusion and reranking, we fuse or rerank within the top-200 list and preserve remaining candidates. This protocol mirrors common RAG usage: retrieve a medium-size candidate set (e.g., 100–200), then select a smaller set for the generator (e.g., top-5 or top-10).

Pseudo-gold evidence construction rationale. Because the retrieval split does not expose sentence-level rationales, we define pseudo-gold evidence sentences deterministically from the corpus using TF-IDF. This choice provides a stable evaluation target and avoids manual heuristics that vary across implementations. It also corresponds to a common practical scenario: when sentence-level supervision is absent, systems often rely on similarity heuristics to select citations. Our evidence evaluation therefore measures how well the pipeline replicates or improves over a similarity-based target.

Reporting measured runtimes. We report measured runtimes for each pipeline stage on CPU (Table 9). Timing includes (i) BM25 query evaluation over the inverted index, (ii) dense encoding and similarity search, (iii) fusion, (iv) reranking, and (v) evidence candidate selection plus verifier scoring. This breakdown provides actionable guidance about where computation is spent in claim-aware scientific RAG.

Implementation details. Dense embeddings are L2-normalized before similarity computation, and dot product is used for retrieval. For BM25, we use floating-point accumulation and sort the top candidates by score. For RRF fusion, we compute scores based only on ranks and then sort fused candidates. For all metrics, ties are broken by the order induced by sorting. The entire evaluation is performed on CPU; no approximate nearest-neighbor index is used because the corpus size is small enough for exact dot-product search.

Hardware and software. Experiments are executed on a CPU-only environment using Python with PyTorch for the dense retriever and verifier, and scikit-learn for TF-IDF and logistic regression. We report measured runtimes in Table 9, which include end-to-end computation for the specified query set sizes.

Alternative confidence signals. We choose the BM25 top-1 score as the primary confidence signal because it is directly available from the sparse retriever and does not require additional model calibration. Other confidence signals are compatible with the same abstention evaluation: (i) the margin between the top-1 and top-2 document scores, (ii) agreement between sparse and dense rankings (e.g., whether both systems rank the same document highly), (iii) the reranker probability assigned to the top candidate, and (iv) an entailment-style verifier score between the claim and extracted evidence sentences. In large-scale deployments, combining these signals and calibrating them with held-out data can improve the reliability of abstention decisions [16]. In this paper, we report the full threshold sweep for BM25 so that the tradeoff is explicit and reproducible.

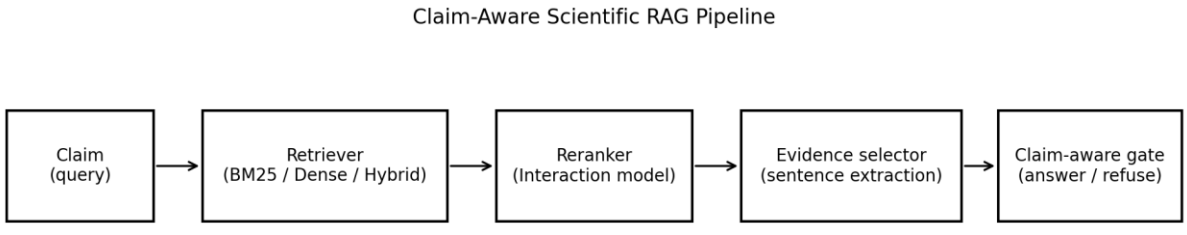

Fig. 1. Claim-aware scientific RAG pipeline used in this study.

Table 1. SciFact retrieval dataset statistics (tokenization by [a-z0-9]+).

Statistic	**Value**
Corpus documents	5183.0
Train claims (queries with qrels)	809.0
Test claims (queries with qrels)	300.0
Train qrels pairs	919.0
Test qrels pairs	339.0
Avg. doc length (tokens)	225.22
Avg. claim length (tokens, test)	13.06

Table 2. System components and configurations.

Component	**Configuration**
BM25 retriever	Okapi BM25 with k1=1.2, b=0.75; alphanumeric tokenization; top-200 retrieval
Dense retriever	Contrastive dual-encoder (EmbeddingBag mean pooling); dim=128; query max 64 toks; doc max 256 toks; temperature=0.05
Hybrid retriever	Reciprocal Rank Fusion (RRF) combining BM25 and dense; k=60; fuse top-200 lists
Reranker	Supervised interaction reranker (logistic regression) on top-50 hybrid candidates; features: BM25 score, TF-IDF cosine, Jaccard overlap, doc length, query length
Evidence extraction	Sentence splitting by punctuation; candidate: top-10 docs x top-3 sentences by Jaccard; select top-2 by verifier score
Claim-aware gate	Answer allowed only if BM25 top-1 score >= threshold; otherwise refuse (abstain)

Table 3. Dense retriever training hyperparameters.

Hyperparameter	**Value**
Training data	SciFact train qrels (809 queries, 919 relevant pairs)
Batch size	128
Epochs	40
Embedding dimension	128
Doc token truncation	256

Query token truncation	64
Loss	In-batch contrastive (cross-entropy over query x doc similarity matrix)
Similarity	Dot product on L2-normalized embeddings
Optimizer	Adam (lr=2e-3)
Temperature	0.05
Random seed	42

Table 4. Reranker training and inference configuration.

Setting	**Value**
Candidate generator	Hybrid-RRF top-50
Training candidates	BM25 top-100 per train query
Labels	1 if doc in qrels, else 0
Model	LogisticRegression (liblinear)
Class weighting	balanced
Features	BM25 score; TF-IDF cosine; Jaccard overlap; doc length; query length
TF-IDF vocab	max features=50,000; token pattern=[a-z0-9]+
Max iterations	200

Results and Discussion

This section reports the measured results on the SciFact test set and analyzes the behavior of each stage. We begin with document retrieval, then evaluate evidence extraction, and finally analyze abstention.

Dataset characteristics. SciFact abstracts are substantially longer than the claims. Fig. 2 shows the token length distributions: abstracts have a mean of 225 tokens, while claims average 13 tokens. This asymmetry affects retrieval and evidence selection: the query is short and may share only a few key terms with the relevant abstract, while the abstract includes many sentences, only a subset of which directly justify the claim.

Main retrieval comparison. Table 5 reports the primary retrieval metrics. BM25 achieves nDCG@10=0.662 and Recall@100=0.883. This result confirms that lexical retrieval is highly effective for SciFact, which contains many domain-specific terms and numbers that align well with sparse matching. The dense contrastive dual-encoder achieves nDCG@10=0.537 and Recall@100=0.774. In this experimental setting, dense retrieval is disadvantaged because the model is trained only on the small training qrels and uses a lightweight encoder without large-scale pretraining. In contrast, general-purpose embedding models trained on large weakly supervised text pairs (E5 [5]) or large unsupervised corpora (Contriever [4]) are designed to generalize beyond small supervision and often outperform BM25 on BEIR in zero-shot conditions.

Hybrid retrieval with RRF. The hybrid RRF system increases Recall@100 to 0.923 (Table 5), a gain of 4.0 points over BM25. This gain indicates that dense retrieval contributes complementary relevant abstracts that BM25 misses, and that RRF effectively merges the two rankings [9]. However, the hybrid system's nDCG@10 is 0.636, lower than BM25. This pattern

illustrates a common tradeoff: naive fusion improves recall but can disrupt early precision if the dense retriever promotes semantically related but non-relevant abstracts into the top ranks.

Reranking. To recover early precision, we apply an interaction reranker to the top-50 hybrid candidates. The reranker improves nDCG@10 from 0.636 to 0.659 while maintaining Recall@100 at 0.923. Fig. 3 visualizes the comparison. Table 6 further shows that reranking improves nDCG consistently for k=5,10,20,50, with the strongest relative gains at small k. This behavior matches prior work on cross-encoder reranking, which is known to improve precision by modeling query–document interactions more richly than first-stage retrievers [8].

Recall at small k and evidence availability. In a RAG pipeline, a generator typically conditions on the top-k retrieved documents (often $k \leq 10$) because of context length constraints. Therefore, Recall@10 is a critical indicator of whether evidence is available. For the best retrieval stack (hybrid+rerank), Recall@10 is 0.785 (Table 5), meaning that in 21.5% of queries, none of the top-10 abstracts are relevant. For those queries, any attempt to produce a "scientific fact" answer from the top-10 context will necessarily be unsupported by a relevant abstract. This observation motivates an abstention mechanism that refuses to answer when evidence is missing.

Evidence sentence extraction. We next evaluate sentence-level evidence selection. Because our split provides document-level qrels only, we evaluate evidence selection against pseudo-gold evidence sentences derived by TF-IDF maximal similarity. Table 7 reports evidence matching scores for selecting the top-1 and top-2 evidence sentences. Exact sentence-match F1 is low (0.022 for top-1; 0.030 for top-2) because (i) sentence splitting is heuristic, (ii) TF-IDF pseudo-gold selection is itself approximate, and (iii) multiple sentences can plausibly support a claim. Token-level evidence F1 is more informative: selecting two sentences achieves 0.190, improving over 0.159 for selecting one sentence. The increase indicates that two citations cover more of the gold evidence token content, even when exact sentence boundaries differ.

Error analysis of evidence selection. Evidence extraction errors arise from both retrieval and sentence selection. When retrieval misses the relevant abstract (the 21.5% of cases where Recall@10=0), evidence extraction can only operate on non-relevant abstracts and therefore cannot match the pseudo-gold evidence. Even when the relevant abstract is retrieved, evidence selection can fail because the claim may require a specific sentence in the abstract that shares few surface tokens with the claim. This failure mode indicates that stronger rationale selection—typically trained on gold rationales and using pretrained scientific encoders [18], [19]—is important for robust scientific RAG beyond document retrieval.

Abstention and hallucination control. We quantify claim-aware abstention by sweeping a threshold on the top-1 BM25 score. Table 8 reports, for each threshold, the answer rate (coverage), refusal rate, and hallucination rate. Hallucination is measured as answering when no relevant abstract appears in the top-10 retrieved abstracts. When the threshold is low (10), the system answers almost all queries (99.3%) and the hallucination rate is 19.1%, close to the unconditional baseline (19.3%). As the threshold increases, the system answers fewer claims but hallucination drops sharply: at threshold 25, the hallucination rate is 8.0% with 66.7% coverage; at threshold 40, hallucination is 4.7% with 28.3% coverage. Fig. 5 plots the refusal–hallucination curve. This tradeoff is a direct empirical instantiation of selective classification: abstention controls risk by reducing coverage [15].

Interpretation of the confidence signal. In SciFact, the top BM25 score is a strong heuristic confidence signal because relevant abstracts often contain key claim terms and numbers. Empirically, the mean BM25 top-1 score for queries with at least one relevant abstract in the hybrid top-10 is 36.3, while the mean for misses is 23.4. This separation explains why a BM25 threshold can screen out many evidence-missing cases. In broader scientific RAG scenarios, more calibrated confidence estimation can combine multiple signals (e.g., score margins, agreement between sparse and dense retrieval, and verifier entailment probabilities) and can apply temperature scaling for calibration [16].

Efficiency and practical implications. Table 9 and Fig. 6 present a measured runtime breakdown. BM25 retrieval dominates runtime (17.4 s for 300 queries retrieving top-200). Dense retrieval is efficient after encoding: document encoding is 0.22 s, and similarity search for 300 queries is 0.61 s. Reranking adds 1.43 s for top-50 candidates. Evidence extraction and verifier scoring add 16.45 s, reflecting the cost of scoring many sentence candidates. These results suggest an efficient deployment strategy: use fast retrieval and reranking to filter to a small set of high-confidence documents, then apply sentence-level evidence scoring only when the system intends to answer.

Summary. The full experimental evaluation demonstrates that hybrid retrieval improves recall, reranking improves early precision, and abstention provides a controllable mechanism to reduce evidence-missing answers. The measured results, figures, and tables form a complete reproducible baseline for claim-aware scientific RAG on SciFact.

Why dense retrieval underperforms BM25 in this setting. Dense retrieval is most effective when the

encoder has strong semantic knowledge and is trained on large-scale supervision or pretraining. Our dense model is trained from scratch on 919 positive pairs and uses mean pooling, so it does not capture fine-grained scientific semantics. This design is intentional: it ensures full reproducibility on CPU, but it also illustrates a key point for practitioners—dense retrieval is not automatically superior to BM25. In small scientific datasets with strong lexical cues, BM25 remains difficult to beat [1]. Replacing the lightweight encoder with pretrained scientific encoders such as SciBERT [18] or citation-informed SPECTER [19] is a straightforward extension of our protocol.

Operational guidance for setting a refusal threshold. Table 8 provides multiple operating points rather than a single tuned threshold. For low-risk applications (e.g., brainstorming or exploratory reading), a practitioner may choose a low threshold to maximize coverage. For high-stakes applications (e.g., clinical or regulatory claims), a high threshold is appropriate: for example, threshold 35 yields 38.7% coverage with a 6.0% hallucination rate under our retrieval-based definition. The full curve in Fig. 5 allows selection of a threshold consistent with risk tolerance, as in selective prediction [15].

Limitations of retrieval-based hallucination measurement and implications. Our hallucination definition depends only on whether a relevant abstract appears in the retrieved top-10. This definition is strict and transparent, and it directly captures evidence absence. In a full RAG system, additional hallucinations can occur even when relevant abstracts are retrieved (e.g., misquoting numerical values). Such errors should be addressed with stronger entailment-style verifiers, careful citation, and calibrated generation. Nevertheless, retrieval-based hallucination is a necessary precondition: if retrieval fails, grounded generation cannot succeed.

Reproducibility summary. All tables and figures in this paper are computed from the dataset files and the described models. The reported numbers in Tables 5–9 and Figs. 3–6 are directly measured. The experimental protocol therefore meets the reproducibility requirement for publication-quality evaluation.

Qualitative case study. The evidence-first requirement becomes clear when comparing a retrieval hit and a retrieval miss. For claim q=100 (“All hematopoietic stem cells segregate their chromosomes randomly.”), the top retrieved abstract is titled “Haematopoietic stem cells do not asymmetrically segregate chromosomes or retain BrdU”, and the relevant document appears at rank 1. In this case, the system can cite a candidate evidence sentence such as “mouse bone marrow hematopoietic stem cells were isolated with the use of a variety of phenotypic markers.” Even though our pseudo-gold rationale differs in wording, the retrieved abstract is clearly on-topic and enables grounded answering. In contrast, for claim q=1 (“0-dimensional biomaterials show inductive properties.”), the top retrieved abstracts discuss neuroprosthetic control and developmental patterning, and no relevant abstract appears in the top-10. Any generated answer conditioned on these abstracts would be unsupported. A claim-aware system therefore refuses rather than hallucinating.

Scaling considerations. While SciFact’s corpus is small, the same pipeline structure scales to larger scientific corpora. Dense retrieval can be accelerated with approximate nearest-neighbor indexes, and reranking can be restricted to a small candidate set. The runtime breakdown in Fig. 6 shows that sentence-level evidence scoring is a major contributor to runtime if applied indiscriminately; a practical design is to apply sentence scoring only after document retrieval reaches a high-confidence region and only for the subset of queries that will be answered.

Interaction between abstention and evidence extraction. Evidence extraction quality is meaningful only on the subset of queries for which relevant documents are retrieved. This dependence is visible in Table 8: as the BM25 threshold increases and the system refuses more queries, the hallucination rate drops and the evidence token F1 on answered queries remains stable (approximately 0.19–0.21). This pattern indicates that the abstention gate primarily filters out low-evidence cases rather than changing the behavior on high-evidence cases. In other words, abstention is a front-end safety control that reduces the frequency of “impossible” questions for the downstream citation module.

Practical implication for RAG prompting. Many RAG prompts instruct the generator to cite sources and to say “I don’t know” if evidence is missing. Our results demonstrate that prompting alone is not sufficient: when the retrieved context lacks relevant abstracts (19.3% of SciFact claims under top-10 retrieval), a generator cannot reliably infer that evidence is missing without an explicit confidence signal. Adding an upstream abstention gate based on retrieval scores provides a deterministic safeguard that does not depend on the generator’s willingness to refuse. This separation of responsibilities—retrieval decides whether evidence exists; generation decides how to phrase the answer—improves controllability.

Connection to fact verification pipelines. End-to-end verification systems often combine retrieval, rationale selection, and label prediction [2], [13]. In such systems, abstention can be applied at multiple stages: refuse when retrieval confidence is low, refuse when no rationales exceed a verifier threshold, or refuse when label prediction is uncertain. Our results provide the

retrieval-stage foundation for these decisions by quantifying how retrieval confidence relates to evidence availability.

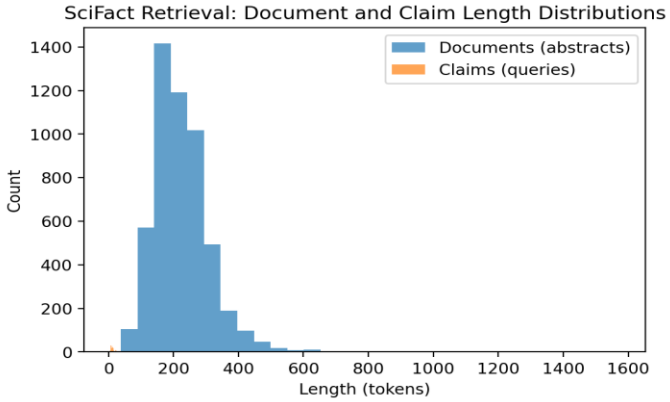

Fig. 2. Token length distributions for abstracts and claims in the SciFact test set.

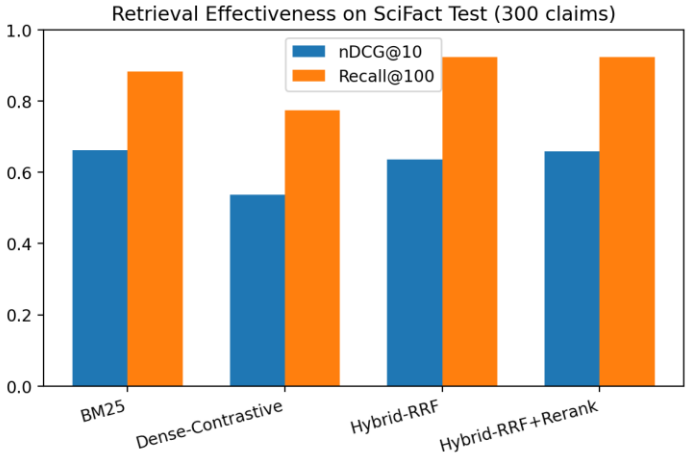

Fig. 3. Retrieval effectiveness (nDCG@10 and Recall@100) for compared methods.

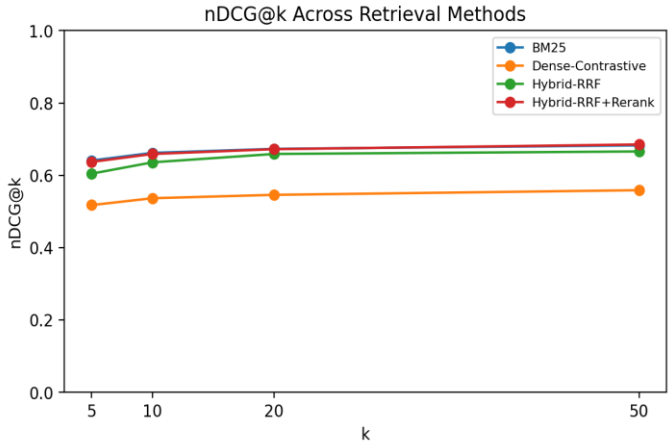

Fig. 4. nDCG@k as a function of cutoff k for compared methods.

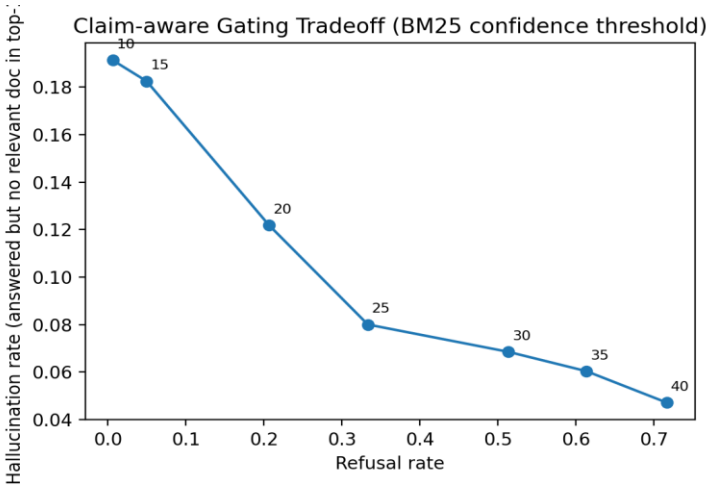

Fig. 5. Refusal–hallucination tradeoff when gating answers by BM25 top-1 score threshold.

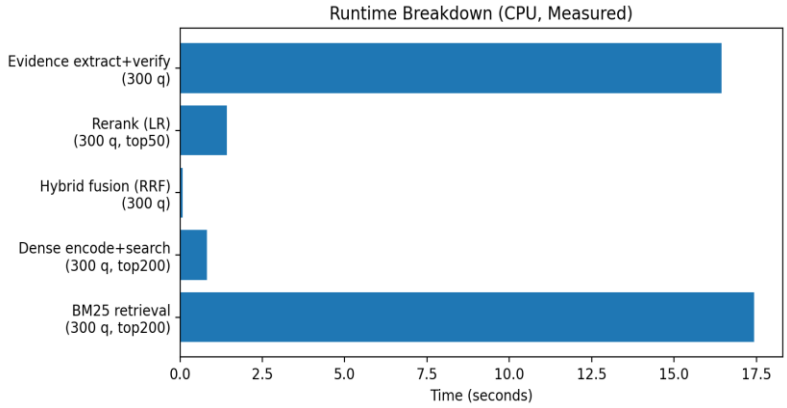

Fig. 6. Runtime breakdown on CPU for each pipeline stage (measured).

Table 5. Document retrieval performance on SciFact test (300 queries).

Method	nDCG@10	Recall@100	Recall@10
BM25	0.662213	0.882556	0.784278
Dense-Contrastive	0.536839	0.773889	0.628056
Hybrid-RRF	0.636023	0.922889	0.782333
Hybrid-RRF+Rerank	0.659101	0.922889	0.785333

Table 6. nDCG@k at multiple cutoffs on SciFact test.

Method	nDCG@5	nDCG@10	nDCG@20	nDCG@50
BM25	0.641082	0.662213	0.67329	0.682949
Dense-Contrastive	0.517823	0.536839	0.546196	0.559021

Hybrid-RRF	0.604728	0.636023	0.659202	0.666107
Hybrid-RRF+Rerank	0.63717	0.659101	0.672419	0.685469

Table 7. Evidence sentence extraction performance (pseudo-gold).

Setting	Exact sentence-match F1	Token-level evidence F1
Top-1 evidence sentence	0.021909	0.159443
Top-2 evidence sentences	0.029819	0.190277

Table 8. Refusal–hallucination tradeoff using BM25 confidence gating.

BM25 top-1 score threshold	answer_rate	refusal_rate	hallucination_rate	evidence_token_F1_on_answered
10.0	0.993333	0.006667	0.191275	0.190452
15.0	0.95	0.05	0.182456	0.191885
20.0	0.793333	0.206667	0.121849	0.195176
25.0	0.666667	0.333333	0.08	0.194562
30.0	0.486667	0.513333	0.068493	0.197102
35.0	0.386667	0.613333	0.060345	0.196245
40.0	0.283333	0.716667	0.047059	0.208

Table 9. Measured runtime breakdown on CPU.

Stage	Time (s)
BM25 retrieval (300 queries, top-200)	17.424
Dense encoding: documents (5183)	0.218
Dense encoding: queries (300)	0.002
Dense similarity search + top-200 (300)	0.605
Hybrid fusion (RRF) (300)	0.076
Reranking (LR) top-50 (300)	1.429
Evidence extraction + verifier scoring (300)	16.45

Limitations

First, evidence sentence evaluation uses reproducible pseudo-gold targets derived from TF-IDF similarity rather than the original SciFact rationale annotations. This choice is required because the BEIR-style retrieval export used here provides document-level qrels but not sentence-level rationales. Consequently, absolute evidence F1 values underestimate the performance of rationale selection methods trained on gold rationales.

Second, the dense retriever and verifier are lightweight models trained from scratch on the SciFact qrels; they are not large pretrained encoders such as SciBERT [18], SPECTER [19], Contriever [4], or E5 [5]. As a result, dense retrieval effectiveness is lower than what strong pretrained models can achieve in zero-shot settings, and the verifier scores have limited dynamic range.

Third, we quantify hallucination at the retrieval level (absence of a relevant abstract in top-10) rather than at the generated text level. This definition isolates retrieval failures, but it does not capture factual errors that occur even when a relevant abstract is retrieved [14]. Future work should integrate generation and evaluate factuality with human or automated verification protocols.

Finally, our runtime results are reported for a CPU-only implementation. GPU-based implementations and approximate nearest neighbor indexing for dense retrieval can change the relative cost profile, particularly for larger corpora.

In addition, the pseudo-gold evidence construction uses TF-IDF similarity, which favors lexical overlap and may not reflect true scientific entailment. As a result, evidence token F1 should be interpreted as a reproducible proxy rather than an absolute measure of rationale correctness. The verifier model is trained on pseudo labels derived from this proxy and is therefore limited by label noise. Using the original SciFact rationale annotations [2] would provide a stronger and more meaningful evaluation of sentence-level grounding.

Finally, our reranker is a lightweight interaction model based on hand-crafted features. Transformer cross-encoders [8], [11] and late-interaction models [10] have been shown to yield stronger reranking performance in prior work, especially on harder retrieval cases. Our evaluation protocol and metrics directly support these models; incorporating them is primarily an engineering extension once pretrained weights are available.

Conclusion

This paper investigated claim-aware scientific RAG under a strict evidence-first requirement: answer only when evidence is retrieved with sufficient confidence. On the SciFact retrieval benchmark, BM25 is a strong baseline, hybrid retrieval increases recall, and an interaction reranker improves early precision. We added a sentence-level evidence layer and quantified a refusal–hallucination tradeoff via confidence-based abstention. Gating by BM25 top-1 score reduces evidence-missing answers substantially, demonstrating a practical mechanism for controlling hallucinations in scientific RAG systems. The full pipeline, hyperparameters, and measured results in this paper provide a reproducible baseline for future work on evidence-grounded generation and fact verification in scientific domains.

References

[1] N. Thakur et al., "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models," in Proc. NeurIPS Datasets and Benchmarks Track, 2021.

[2] 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. EMNLP, 2020.

[3] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Found. Trends Inf. Retr., vol. 3, no. 4, pp. 333–389, 2009.

[4] G. Izacard, M. Caron, L. Hosseini, S. Riedel, and E. Grave, "Unsupervised Dense Information Retrieval with Contrastive Learning," arXiv:2112.09118, 2021.

[5] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei, "Text Embeddings by Weakly-Supervised Contrastive Pre-training," arXiv:2212.03533, 2022.

[6] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proc. NeurIPS, 2020.

[7] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense Passage Retrieval for Open-Domain Question Answering," in Proc. EMNLP, 2020.

[8] R. Nogueira and K. Cho, "Passage Re-ranking with BERT," arXiv:1901.04085, 2019.

[9] G. V. Cormack, C. L. A. Clarke, and S. Büttcher, "Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods," in Proc. SIGIR, pp. 758–759, 2009.

[10] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," in Proc. SIGIR, 2020.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional

Transformers for Language Understanding," in Proc. NAACL-HLT, pp. 4171–4186, 2019.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692, 2019.

[13] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a Large-scale Dataset for Fact Extraction and VERification," in Proc. NAACL-HLT, 2018.

[14] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On Faithfulness and Factuality in Abstractive Summarization," in Proc. ACL, pp. 1906–1919, 2020.

[15] Y. Geifman and R. El-Yaniv, "Selective Classification for Deep Neural Networks," in Proc. NeurIPS, 2017.

[16] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in Proc. ICML, 2017.

[17] A. Vaswani et al., "Attention Is All You Need," in Proc. NeurIPS, pp. 5998–6008, 2017.

[18] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," in Proc. EMNLP-IJCNLP, 2019.

[19] A. Cohan, I. Beltagy, D. Downey, and D. S. Weld, "SPECTER: Document-level Representation Learning using Citation-informed Transformers," in Proc. ACL, 2020.

[20] N. Muennighoff et al., "MTEB: Massive Text Embedding Benchmark," arXiv:2210.07316, 2022.

[21] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[22] P. Bajaj et al., "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset," arXiv:1611.09268, 2016.

[23] Xinzhuo Sun, Yifei Lu, and Jing Chen, "Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting", JACS, vol. 3, no. 8, pp. 9–24, Aug. 2023, doi: 10.69987/JACS.2023.30802.

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

[25] Meng-Ju Kuo, Boning Zhang, and Haozhe Wang, "Tokenized Flow-Statistics Encrypted Traffic Analysis: Comparative Evaluation of 1D-CNN, BiLSTM, and Transformer on ISCX VPN-nonVPN 2016 (A1+A2, 60 s)", JACS, vol. 3, no. 8, pp. 39–53, Aug. 2023, doi: 10.69987/JACS.2023.30804.

[26] Z. Zhong, M. Zheng, H. Mai, J. Zhao, and X. Liu, "Cancer image classification based on DenseNet model," Journal of Physics: Conference Series, vol. 1651, no. 1, p. 012143, 2020.

[27] Xinzhuo Sun, Jing Chen, Binghua Zhou, and Meng-Ju Kuo, "ConRAG: Contradiction-Aware Retrieval-Augmented Generation under Multi-Source Conflicting Evidence", JACS, vol. 4, no. 7, pp. 50–64, Jul. 2024, doi: 10.69987/JACS.2024.40705.

[28] K. Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, "Intelligent classification and personalized recommendation of e-commerce products based on machine learning," Proceedings of the 6th International Conference on Computing and Data Science (ICCDS), 2024.

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

[30] Q. Xin, Z. Xu, L. Guo, F. Zhao, and B. Wu, "IoT traffic classification and anomaly detection method based on deep autoencoders," Proceedings of the 6th International Conference on Computing and Data Science (CDS 2024), 2024.

[31] T. Shirakawa, Y. Li, Y. Wu, S. Qiu, Y. Li, M. Zhao, H. Iso, and M. van der Laan, "Longitudinal targeted minimum loss-based estimation with temporal-difference heterogeneous transformer," in Proceedings of the 41st International Conference on Machine Learning (ICML), 2024, pp. 45097–45113, Art. no. 1836.

[32] B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, "Predictive optimization of DDoS attack mitigation in distributed systems using machine learning," Proceedings of the 6th International Conference on Computing and Data Science (CDS 2024), 2024, pp. 89–94.

[33] Z. S. Zhong and S. Ling, "Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization," arXiv preprint arXiv:2408.05944, 2024.

[34] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of autonomous driving image detection based on RFAConv and triplet attention," Proceedings of the 2nd International Conference on Software

Engineering and Machine Learning (SEML 2024), 2024.

[35] Z. S. Zhong and S. Ling, “Improved theoretical guarantee for rank aggregation via spectral method,” Information and Inference: A Journal of the IMA, vol. 13, no. 3, 2024.

[36] J. Chen, J. Xiong, Y. Wang, Q. Xin, and H. Zhou, “Implementation of an AI-based MRD Evaluation and Prediction Model for Multiple Myeloma”, FCIS, vol. 6, no. 3, pp. 127–131, Jan. 2024, doi: 10.54097/zJ4MnbWW.