

Self-Reflective RAG for Biomedical Question Answering with Retrieve-Critique-Revise Loops

Tao Huang¹, Ashley Lewis², Hui Zhang³, Steven Walker⁴, Ling Gao⁵

¹John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

²Department of Computer Science, Duke University, Durham, NC, USA

³Department of Computer Science, Rice University, Houston, TX, USA

⁴Department of Computer Science, Northwestern University, Evanston, IL, USA

⁵Department of Computer Science, Brown University, Providence, RI, USA

taohuang_cs@yahoo.com

Abstract

Biomedical yes/no/maybe question answering requires a model to distinguish affirmative findings from negative results and genuine evidential uncertainty. This study evaluates a self-reflective retrieval-augmented architecture on the 1,000 expert-labeled PubMedQA PQA-L instances. The canonical protocol was reconstructed with 500 development cases, ten stratified development folds, and a separate 500-case test set. Three systems were compared: question-only classification, section-aware retrieval-augmented classification, and a Retrieve-Critique-Revise loop. The retriever selected result-bearing abstract sections, the answer head used word and character TF-IDF features with a linear support vector classifier, and the critic combined 79 cross-fitted answer, retrieval, lexical, structural, confidence, and disagreement signals. The correction gate was fixed on development data under two constraints: no more than 10% correction coverage and at least 60% decisive correction precision. On the held-out test set, question-only QA reached 55.4% accuracy and 0.309 macro-F1. Section-aware retrieval increased accuracy to 58.8% and macro-F1 to 0.335. The full loop reached 60.0% accuracy, 0.381 macro-F1, and 0.405 balanced accuracy. It changed 49 answers, corrected 26 initial errors, harmed 20 correct answers, and produced six net additional correct decisions. The strongest effect was improved no-class recall, which rose from 17.8% under retrieval alone to 33.1% after critique. Paired testing showed that the full loop exceeded question-only QA by 4.6 percentage points with a 95% bootstrap interval of 1.2 to 8.2 points and an exact McNemar p-value of 0.019. The results show that selective evidence criticism can improve biomedical QA without indiscriminate second-pass changes, while the unresolved maybe class remains the central limitation.

Keywords: Biomedical Question Answering; Retrieval-Augmented Generation; Self-Reflection; Evidence Critique; Pubmedqa; Selective Correction; Uncertainty.

Introduction

Biomedical question answering maps a research question and its source article to yes, no, or maybe [1]. These labels distinguish affirmative findings, negative findings, and unresolved evidence rather than mere confidence levels. Related biomedical QA and reading-comprehension benchmarks [23]-[27] show that domain knowledge alone is insufficient: a system must locate the result that actually resolves the question.

Pretrained language models improved biomedical and scientific representation, from BERT [8] to BioBERT, SciBERT, and PubMedBERT [9]-[11], while text-to-text transfer broadened task coverage [12]. Yet long abstracts can dilute result-bearing sentences with background and methods, so evidence selection remains necessary even when the correct article is already paired with the question.

Retrieval-augmented generation separates parametric prediction from retrieved evidence [2]. DPR and REALM showed the value of learned retrieval [4], [5], and Fusion-in-Decoder integrates multiple passages [13]. Sparse BM25 [6] and MMR reranking [7] remain strong when exact terminology matters. PubMedQA provides a controlled setting because retrieval occurs within a structured abstract rather than across the open web.

Retrieval alone cannot guarantee correctness: an answerer may over-weight mechanism, miss a null result, or interpret uncertainty as weak affirmation. ReAct, Reflexion, chain-of-thought, self-consistency, and Self-Refine use actions or feedback to expose such errors [14]-[18]. Self-RAG applies the same idea to retrieval [3], but any second pass must remain selective because it can replace a correct answer with a plausible alternative.

PQA-L combines three difficulties: yes is the majority class, abstracts can contain positive motivation beside null findings, and changing yes to no creates asymmetric risk. A reflective system must therefore report class-level behavior, identify exactly which cases change, and constrain intervention frequency rather than rely on a single aggregate score. This study evaluates that principle with a fixed Retrieve-Critique-Revise loop. It asks whether result-focused retrieval improves held-out accuracy, whether a learned critic can safely correct affirmative errors, and how depth, context length, and coverage affect the correction trade-off. Compact linear heads isolate evidence and policy effects, while calibration-aware gating [19], [20] and paired testing make every intervention auditable.

METHOD

A. Task, data, and evaluation protocol

Each instance contains a research question, a list of abstract sections, section-type labels, MeSH descriptors, publication year, a long-form conclusion, and the expert final decision. The prediction input was restricted to the question and abstract sections. The long-form conclusion and final decision were never included in model features. The task was three-class classification with the fixed class order no, maybe, and yes. The complete PQA-L set contains 1,000 cases: 552 yes, 338 no, and 110 maybe. Abstracts contain an average of 3.36 sections and 206.7 words, while questions average 13.0 words.

The canonical PubMedQA split was reconstructed with seed 0 by stratifying each class into equal 500-case development and held-out halves. The development half was further divided into ten stratified folds. All model settings, critic parameters, and the correction threshold were selected from cross-fitted development predictions; held-out labels were used once for final evaluation. Each half contains 276 yes, 169 no, and 55 maybe cases (Table 1).

Table 1. PQA-L composition and canonical experimental split.

Split	N	Yes	No	Maybe
All PQA-L	1000	552	338	110
Development	500	276	169	55
Held-out test	500	276	169	55

B. Retrieve and answer stages

The question-only baseline represents each question with the union of word TF-IDF features over one- to three-grams and character TF-IDF features over three- to five-grams. TF-IDF follows the classical vector-space retrieval tradition [29]. A linear support vector classifier maps the sparse vector to no, maybe, or yes; linear margin classifiers are well suited to high-dimensional text representations [28]. The baseline therefore measures how much label information can be recovered from question wording alone, including constructions such as 'does,' 'is associated,' or 'can.'

The paired abstract is treated as a small evidence collection. BM25 supports depth ablations, while the primary retriever selects sections labeled RESULT, CONCLUSION, FINDING, or CASE, falling back to the final section. The question and evidence are encoded with word and character TF-IDF and classified by a class-balanced linear SVM ($C = 0.05$). Full-abstract, last-sentence, BM25 top-k, and MMR variants remain ablations rather than components of the primary score.

The implementation also trains ten evidence views for the critic: question-only SVM and Complement Naive Bayes heads; BM25 top-1 and top-7 SVM heads; a top-1 Naive Bayes head; result-section SVM and Naive Bayes heads; a full-abstract SVM; a last-two-sentence SVM; and an MMR top-5 SVM. Each view produces three class scores. Development scores are generated out of fold, and every final view is fitted on all 500 development cases before it scores the held-out set. The three systems compared in the main analysis are defined in Table 2, and Fig. 1 shows the end-to-end loop.

Ten-fold cross-fitting generates one out-of-fold score vector for every development case, and the critic is evaluated on the same fold structure. After selecting the operating point, all answer views and the critic are refit on the 500 development cases and applied once to the held-out 500. Neither test features nor labels influence model construction, parameter choice, or threshold selection.

The evidence pack and revision gate follow a broader reliability pattern in which retrieval, interpretation, and abstention remain separable. Retrieved failure narratives and selective refusal have been used to distinguish actionable HDFS anomalies from ambiguous cases [30], complementing the retrieve-and-critique logic of Self-RAG [3]. Domain-specific representation learning has likewise improved cancer-image classification [31], while evidence-grounded financial-disclosure QA ties answers to traceable source passages [32]. Bilingual computing instruction further shows that the organization of technical material can affect comprehension [33]. These findings motivate the compact, result-focused evidence pack used here. Drift-aware rollback policies [34] and human-uncertainty distillation [35] also support separating confidence estimation from intervention, consistent with established calibration methods [19], [20].

Table 2. System definitions and fixed decision rules.

System	Retrieval	Answer head	Critic	Rule
Baseline QA	nan	Word and character TF-IDF + linear SVM	nan	Argmax over no/maybe/yes
Section-aware RAG	RESULT/CONCLUSION/FINDING/CASE sections; final section fallback	Word and character TF-IDF + class-balanced linear SVM (C=0.05)	nan	Argmax over no/maybe/yes
RCR	Same result-focused evidence pack	Same initial answer	79-feature no-verifier; balanced logistic regression (C=0.03)	Change yes to no when $p(\text{no}) \geq 0.70$

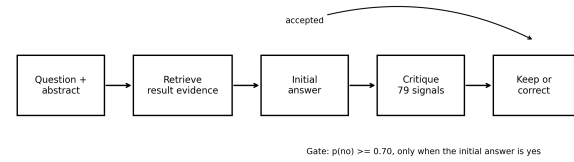


Fig. 1. Retrieve-Critique-Revise architecture evaluated on PQA-L.

C. Critique features and selective correction

The critic estimates whether negative evidence is strong enough to challenge an initial yes answer. Its 79 inputs comprise 30 standardized class scores, 25 retrieval, lexical, and structural features, ten margins, ten entropies, three vote counts, and one disagreement count. Features capture BM25 strength, question-evidence overlap, evidence length, section structure, MeSH count, publication year, polarity cues, conclusion cues, and explicit p-value expressions.

A balanced logistic regression ($C = 0.03$) predicts no versus not-no. Development analysis did not support a precise maybe branch, so the gate acts only when the initial answer is yes. Thresholds from 0.40 to 0.95 are evaluated on cross-fitted predictions; an eligible point must change at most 10% of cases and achieve at least 60% decisive correction precision. The best eligible point is 0.70.

The selective policy operationalizes the central self-reflection hypothesis: a critic should act as a verifier rather than as an unrestricted second answerer. An unrestricted second pass can improve class balance while lowering total accuracy, particularly when the positive class is common. The coverage and precision constraints make the intervention auditable. They also separate the critic's discrimination quality from its downstream value, which is essential when probability estimates are imperfect [19], [20].

The critic is therefore treated as a constrained fusion layer rather than an unrestricted replacement model. Comparable fusion problems arise in autonomous-driving perception [36] and clinical minimal-residual-disease prediction [37], whereas risk-constrained bidding systems make the operating rule explicit [38]. Statistical analyses of synchronization estimators [39] and calibrated credit-risk decisions [40] emphasize that a score becomes actionable only when its uncertainty and threshold are interpretable. Spectral rank aggregation [41], confidence-aware late fusion for 3-D perception [42], and transfer through approximately shared features [43] offer related ways to reconcile heterogeneous signals. In the present pipeline, logistic fusion combines the standardized answer views, while the 0.70 gate preserves a conservative revision policy [19], [20].

The revision is deterministic: yes changes to no only when the critic probability is at least 0.70. Outcomes are logged as unchanged, corrected, harmed, or wrong-to-wrong. This accounting distinguishes raw self-correction counts from net accuracy gain and makes the operating policy reproducible.

D. Metrics and statistical analysis

Primary evaluation uses exact three-class accuracy, with macro-F1, balanced accuracy, and per-class scores. Self-correction is reported as trigger success, decisive correction precision, initial-error recovery, and net gain. Accuracy intervals use 10,000 class-stratified bootstrap samples [21], and paired differences use exact McNemar tests [22]. Efficiency includes measured fit time, per-question inference time, feature count, and evidence length.

RESULTS AND DISCUSSION

A. Data profile and main comparison

The dataset is strongly imbalanced toward yes, as shown in Fig. 2. A majority strategy would therefore reach 55.2% on the full set, a property already noted in the original benchmark [1]. The measured question-only system reached 55.4% on the held-out half, confirming that question wording added little beyond the class prior. Its predictions contained no maybe decisions, and 440 of 500 cases were assigned yes. This behavior is important because an apparently reasonable overall score can conceal near-total failure on negative and uncertain evidence.

Section-aware RAG increased held-out accuracy from 55.4% to 58.8%, a gain of 3.4 percentage points. Macro-F1 rose from 0.309 to 0.335, and balanced accuracy rose from 0.354 to 0.378. The result-focused evidence pack improved yes recall from 90.9% to 95.7% and no precision from 43.3% to 71.4%, but no recall remained only 17.8%. The retrieval stage was therefore conservative: when it predicted no, it was usually right, yet it continued to classify most negative and uncertain cases as yes.

The full RCR loop reached 60.0% accuracy, 0.381 macro-F1, and 0.405 balanced accuracy. Relative to retrieval alone, the loop added 1.2 accuracy points and 4.7 macro-F1 points. Relative to question-only QA, it added 4.6 accuracy points and 7.2 macro-F1 points. Table 3 reports the aggregate metrics and bootstrap intervals, while Fig. 3 visualizes the accuracy and macro-F1 progression. The accuracy interval for RCR was 56.8% to 63.2%. The overlap between method-specific intervals does not replace a paired test, because all systems answered the same cases; paired comparisons are reported later.

The divergence between accuracy and macro-F1 is central to the interpretation. Most of the absolute accuracy remains attributable to correct yes decisions because yes is the largest class. Retrieval raises accuracy while changing relatively few predictions, whereas critique has a larger effect on macro-F1 because it redistributes decisions toward no. Balanced accuracy follows the same direction, increasing by 5.1 points from the question-only system to RCR. These three metrics therefore answer different questions: accuracy measures the probability of a correct benchmark decision under the observed label distribution, macro-F1 tests whether performance is shared across labels, and balanced accuracy summarizes sensitivity to each class. The full loop improves all three, but the gap between 60.0% accuracy and 40.5%

balanced accuracy shows that class coverage remains substantially uneven.

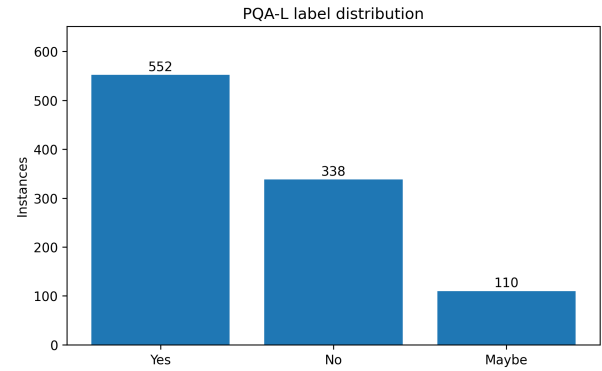


Fig. 2. Label counts in the 1,000 expert-labeled PQA-L instances.

Table 3. Held-out aggregate performance with stratified bootstrap accuracy intervals.

Method	Accuracy	95% CI	Macro-F1	Balanced acc.
Baseline QA	55.4%	52.8-58.0%	0.309	0.354
Section-aware RAG	58.8%	56.4-61.2%	0.335	0.378
RCR	60.0%	56.8-63.2%	0.381	0.405

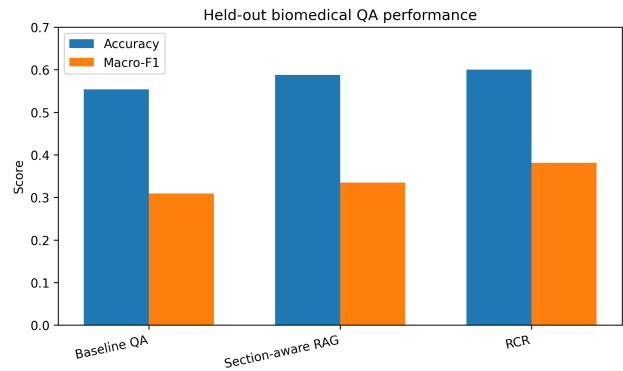


Fig. 3. Accuracy and macro-F1 for the three primary systems.

B. Class-conditional behavior

The class-conditional results in Table 4 and Fig. 4 show that the critique layer primarily improved recognition of negative findings. No recall nearly doubled from 17.8% under section-aware RAG to 33.1% under RCR, and no F1 increased from 0.284 to 0.431. The cost was a reduction in yes recall from 95.7% to 88.4%. The yes F1 remained stable because precision increased from 57.8% to 59.8% while recall decreased. This pattern is consistent with the gate's design: it trades a controlled amount of positive-class coverage for better recovery of negative evidence.

Maybe remained unsolved. None of the three primary systems correctly predicted a maybe case, and the RCR policy did not attempt to create that label. This result is not hidden by the aggregate score. It indicates that uncertainty is not merely an intermediate confidence level between yes and no. In PQA-L, maybe can arise from insufficient power, qualified subgroup findings, incomplete follow-up, competing mechanisms, or an abstract that does not resolve the title question. A binary no verifier cannot model these patterns, and the 55 maybe cases in each split provide limited supervision for a high-precision correction branch. The failure supports the broader observation that medical QA benchmarks test several reasoning types rather than a single factual mapping [24]-[26].

Table 4. Per-label precision, recall (class-conditional accuracy), and F1.

Method	Label	Precision	Recall	F1	Support
Baseline QA	no	0.433	0.154	0.227	169
Baseline QA	maybe	0.000	0.000	0.000	55
Baseline QA	yes	0.570	0.909	0.701	276
Section-aware RAG	no	0.714	0.178	0.284	169
Section-aware RAG	maybe	0.000	0.000	0.000	55
Section-aware RAG	yes	0.578	0.957	0.720	276
RCR	no	0.615	0.331	0.431	169
RCR	maybe	0.000	0.000	0.000	55
RCR	yes	0.598	0.884	0.713	276

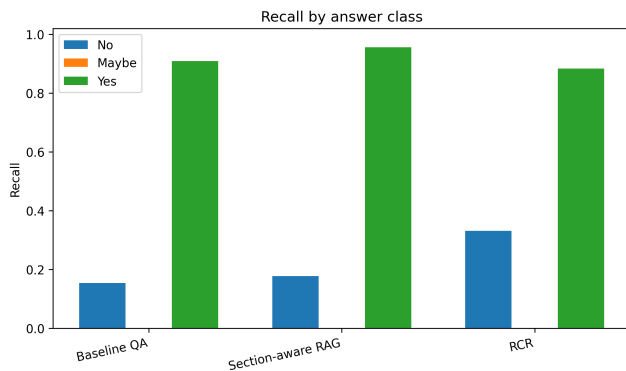


Fig. 4. Recall for no, maybe, and yes across the three primary systems.

Table 5. Confusion matrices in predicted no/maybe/yes order for each gold row.

Method	Gold no	Gold maybe	Gold yes
Baseline QA	26/0/143	9/0/46	25/0/251
Section-aware RAG	30/0/139	1/0/54	11/1/264
RCR	56/0/113	4/0/51	31/1/244

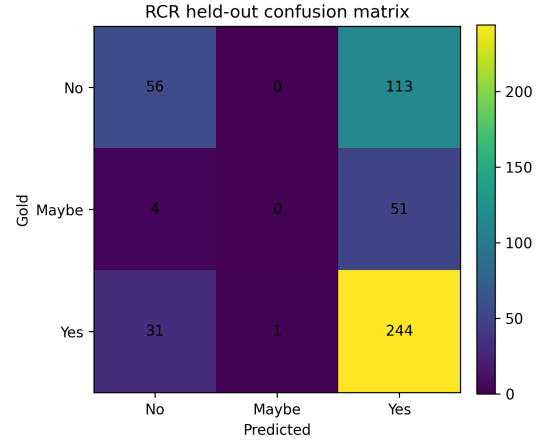


Fig. 5. Held-out confusion matrix for the RCR system.

C. Retrieval depth, evidence selection, and context length
Sentence-level BM25 retrieval did not match section-aware selection. Across $k = 1$ to 7, held-out accuracy ranged from 51.2% to 54.8% (Table 6 and Fig. 6). The best BM25 depth was seven sentences, but it remained four points below the result-section retriever. MMR top-5 reached 55.2%, last-two-sentence selection reached 54.6%, and the full abstract reached 57.2%. Result-section retrieval was best at 58.8%. The ranking suggests that biomedical decision evidence is often distributed across several result sentences, while pure lexical similarity can elevate background statements that repeat terms from the question. Conversely, the full abstract retains too much non-decisive material. Section labels provide a compact structural prior that is particularly effective in this paired-abstract setting.

The depth curve is non-monotonic rather than a simple more-context-is-better relationship. Test accuracy fell to 51.2% at one sentence, rose to 53.6% at two, dropped at three, and reached its BM25 maximum of 54.8% at seven. Development accuracy followed a different local pattern, peaking earlier and declining at larger depths. This instability reflects the small number of labeled cases as well as a retrieval mismatch: lexical rank identifies sentences that echo the question, while the answer may depend on a later comparison, a confidence interval, or a sentence stating that an expected association was absent. MMR reduces repeated evidence but cannot infer discourse function. The result-section rule succeeds because it encodes that function directly, selecting a broader evidence unit without retaining the entire abstract.

The context-length analysis in Table 12 provides a complementary view. For the shortest quartile, RCR achieved 57.1% compared with 53.2% for question-only QA. In the third quartile, RCR reached 60.0% while both earlier systems

remained below 56%. Section-aware RAG was strongest for the longest abstracts at 68.0%, showing that structural filtering becomes more valuable as context grows. RCR reached 65.6% in that quartile because some long-context affirmative answers were changed incorrectly. This is a reminder that a critic should not be judged solely by average improvement: its error profile changes with evidence density and should be monitored by context regime.

The retrieval findings also fit a broader family of evidence-selection systems. Long-document RAG for contractual clauses separates passage finding from clause interpretation [44], and retrieval-summary integration performs a similar division in code intelligence [45]. Retrieved normal prototypes have been used to explain OpenStack fault-injection anomalies [46], while CI diagnosis pipelines connect failure evidence to repair recommendations [47]. Deployment studies add a resource dimension: hybrid-cloud LLM architectures [48], layout-aware PDF slice prioritization [49], and edge-deployable video-quality proxies [50] all require useful evidence to arrive under latency or compute constraints. Explanation-oriented financial dashboards [51] and traffic-sign evidence cards [52] likewise place selected evidence, prediction, and confidence in one traceable presentation. These patterns help explain why the compact section-aware pack outperformed longer BM25 contexts [6], [7], [13].

Table 6. BM25 sentence-retrieval depth sensitivity.

k	Dev accuracy	Dev macro-F1	Test accuracy	Test macro-F1
1	0.558	0.338	0.512	0.299
2	0.540	0.322	0.536	0.304
3	0.530	0.284	0.528	0.292
4	0.534	0.290	0.536	0.294
5	0.544	0.297	0.528	0.284
6	0.550	0.306	0.532	0.294
7	0.568	0.316	0.548	0.305

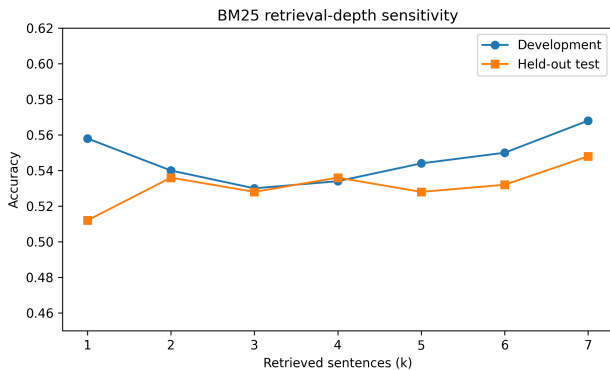


Fig. 6. Development and held-out accuracy as BM25 retrieval depth increases.

D. Critic threshold and self-correction

The no-verifier achieved 0.640 development AUROC, 0.501 average precision, and a Brier score of 0.233. These values indicate useful but incomplete error separation. Figure 7 shows why a gate is necessary. Low thresholds change many affirmative answers and sharply reduce accuracy. A threshold of 0.50 changed 35.6% of development cases and produced 55.2% accuracy. Development accuracy peaked at 59.8% near 0.60, but the decisive correction precision was only 57.3% and correction coverage was 20.0%, violating both safety constraints. At 0.70, coverage fell to 7.8%, decisive correction precision rose to 61.1%, and development accuracy was 58.8%. Table 7 reports the central threshold range used to select the operating point.

On the held-out set, the 0.70 gate changed 49 of 500 answers (9.8%). Twenty-six changes converted an error to the correct no label, twenty changed a correct yes answer to no, and three moved one wrong answer to another wrong label. The success rate among all triggered cases was 53.1%; the decisive correction precision was 56.5%; and the loop recovered 12.6% of the 206 initial retrieval errors. Corrections exceeded harms by six cases, producing the 1.2-point accuracy gain. Table 8 gives the complete accounting, and Fig. 8 makes the small net margin visible. The held-out decisive precision was below the 60% development floor, but the correction coverage remained within the 10% cap. This shift illustrates why a development constraint controls risk without guaranteeing identical test precision.

These rates have distinct denominators and should not be conflated. The 53.1% success rate asks how often any triggered change ends at the gold label. Decisive correction precision removes the three wrong-to-wrong transitions and compares only beneficial changes with harmful changes, yielding 26 divided by 46. Initial-error recovery instead asks how much of the original error pool is repaired, yielding 26 divided by 206. The six-case net gain is the quantity that directly changes aggregate accuracy. A critic can therefore have an apparently useful discrimination score and still contribute little if it acts too broadly, while a narrow gate can improve accuracy even though it repairs only a minority of initial errors. The measured loop occupies the latter regime: limited coverage, moderate precision, and a small positive net effect.

The confusion matrix in Fig. 5 shows the resulting trade-off. Correct no decisions increased from 30 under retrieval alone to 56 under RCR. At the same time, correct yes decisions decreased from 264 to 244. The number of maybe cases assigned no increased from one to four, while 51 of 55 maybe cases remained yes. The critic therefore corrected a genuine polarity failure but did not become a general uncertainty reasoner. This narrow behavior is desirable for auditability, yet it also limits the ceiling of the current loop.

Operational anomaly-detection research provides a useful analogue for the critic's intervention record. Noisy-neighbor VM degradation models combine residual signals before acting [53]; deep-autoencoder IoT classifiers [54] and

predictive DDoS mitigation models [55] similarly separate anomaly evidence from downstream response. AIOps fault-analysis systems [56], attention-enhanced road-scene detectors [57], and digital-twin dispatching with offline reinforcement learning [58] extend the same pattern to more complex control settings. Here, the 0.70 gate plays the corresponding control role: it converts critic probability into a bounded, auditable change policy instead of allowing every disagreement to overwrite the initial answer [19], [20].

Table 7. Development threshold trade-off around the selected operating point.

Threshold	Accuracy	Correction coverage	Corrections	Harms	Decision precision
0.550	0.582	26.0%	59	54	52.2%
0.575	0.596	22.6%	56	44	56.0%
0.600	0.598	20.0%	51	38	57.3%
0.625	0.588	16.8%	43	35	55.1%
0.650	0.586	13.2%	34	27	55.7%
0.675	0.578	10.8%	27	24	52.9%
0.700	0.588	7.8%	22	14	61.1%
0.725	0.586	6.2%	18	11	62.1%
0.750	0.584	5.2%	16	10	61.5%
0.775	0.580	4.0%	12	8	60.0%
0.800	0.574	3.4%	9	8	52.9%

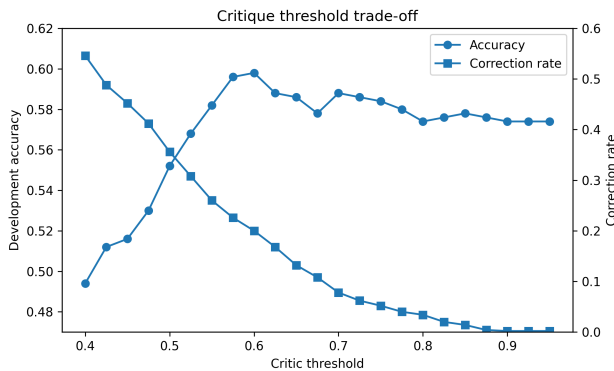


Fig. 7. Development accuracy and correction coverage across critic thresholds.

Table 8. Held-out self-correction accounting.

Quantity	Cases	Associated rate
Initial retrieval errors	206	100.0%
Answers changed by the gate	49	9.8%
Error corrected	26	53.1%
Correct answer harmed	20	40.8%
Wrong-to-wrong change	3	6.1%
Initial-error recovery	26	12.6%
Net additional correct cases	6	1.2%

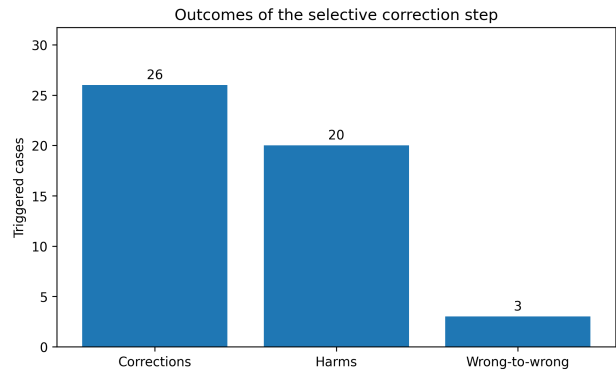


Fig. 8. Outcomes among the 49 held-out answers changed by the critic.

E. Ablations, efficiency, and paired inference

Table 9 confirms that the gain depends on both evidence structure and selective gating. BM25 top-1 scored 51.2%, BM25 top-7 scored 54.8%, MMR top-5 scored 55.2%, and the full abstract scored 57.2%. Result-section retrieval scored 58.8%. Applying the critic with a 0.50 threshold reduced held-out accuracy to 55.0% because the gate acted too often. A 0.60 threshold reached 58.2%, 0.70 reached 60.0%, and 0.80 reached 59.0%. Thus, stronger intervention did not monotonically improve performance. The best held-out score occurred at the preselected risk-controlled threshold rather than at the lower threshold with the best unconstrained development accuracy.

The evidence pack was also smaller than the full abstract. Section-aware RAG used an average of 97.5 evidence words, compared with 209.0 words for full-context reading, a 53.3% reduction. Including the question, the average input was 110.6 words rather than 222.1. Table 11 reports measured CPU timings from the complete run. The section-aware answerer required 0.231 seconds for ten-fold development fitting plus final fitting and 0.578 milliseconds per held-out question. The full-context ablation required 0.623 seconds and 1.213 milliseconds per question. The final 79-feature critic added 0.034 seconds of fitting and approximately 0.001 milliseconds per question after its features were available. The evidence-view bank is the principal extra cost of critique; the logistic gate itself is negligible.

Paired statistical results appear in Table 10. Section-aware RAG exceeded question-only QA by 3.4 points; the paired bootstrap interval was 0.0 to 6.8 points, and exact McNemar $p = 0.064$. RCR exceeded section-aware RAG by 1.2 points with an interval of -1.2 to 3.8 points and $p = 0.461$. The incremental correction gain is therefore not individually significant at the 0.05 level. In contrast, RCR exceeded question-only QA by 4.6 points with an interval of 1.2 to 8.2 points and $p = 0.019$. Fifty-six cases were improved and 33 regressed in that paired comparison. The appropriate conclusion is that the complete retrieve-and-critique system outperformed the question-only baseline on this split, while a larger evaluation is needed to isolate the smaller second-pass contribution with high power.

From an operational perspective, the strongest result is the combination of a smaller evidence footprint and an explicit intervention record. The primary retriever removes slightly more than half of the abstract words while improving accuracy over both question-only input and the sentence-retrieval variants. The critic then changes a bounded set of cases that can be inspected directly in the prediction file together with its probability, original answer, final answer, and gold label. This design does not eliminate errors, but it converts reflection from an unobserved generation step into a measurable decision policy. The same audit can be repeated at another threshold to quantify how many additional corrections would be purchased at the cost of additional harms.

Evaluation should therefore follow the complete decision policy, not only the predictor that feeds it. E-commerce recommendation pipelines illustrate how classification quality affects downstream selection [59], while uncertainty-aware VIX forecasting shows why macroeconomic and textual signals must be fused before a directional decision [60]. Enterprise marketing models [61] and constrained budgeting systems [62] make similar links between prediction and allocation. Financial-service risk analytics adds the need for interpretable customer and risk signals [63], and counterfactual learning-to-rank evaluates policies from logged interactions when online trials are expensive [64]. The paired bootstrap and McNemar analyses used here [21], [22] apply the same principle at benchmark scale by testing a fixed end-to-end policy on held-out cases.

Table 9. Held-out ablation study.

Configuration	Accuracy	Macro-F1	Balanced acc.
Question only	0.554	0.309	0.354
BM25 top-1	0.512	0.299	0.332
BM25 top-7	0.548	0.305	0.350
MMR top-5	0.552	0.264	0.339
Last two sentences	0.546	0.359	0.379
Full abstract	0.572	0.328	0.369
Result-section retrieval	0.588	0.335	0.378
RCR threshold 0.50	0.550	0.384	0.415
RCR threshold 0.60	0.582	0.392	0.414
RCR threshold 0.70	0.600	0.381	0.405
RCR threshold 0.80	0.590	0.352	0.386

Table 10. Paired held-out comparisons.

Comparison	Accuracy change	Bootstrap 95% CI	Improved	Regression	McNemar p
Section-aware RAG vs. Baseline QA	+3.4 pp	+0.0 to +6.8 pp	46	29	0.064
RCR vs. Section-aware RAG	+1.2 pp	-1.2 to +3.8 pp	26	20	0.461
RCR vs. Baseline QA	+4.6 pp	+1.2 to +8.2 pp	56	33	0.019

Table 11. Evidence footprint and measured CPU efficiency.

Component	Evidence words	Total input words	Fit time (s)	Inference (ms/q)	Features
Baseline QA	0.0	13.1	0.067	0.091	11875
Section-aware RAG	97.5	110.6	0.231	0.578	37723
Full-context ablation	209.0	222.1	0.623	1.213	47672
Critic and gate	97.5	110.6	0.034	0.001	79

Table 12. Accuracy by held-out abstract-length quartile.

Length group	N	Mean words	Baseline	RAG	RCR
Q1 shortest	126	141.4	0.532	0.556	0.571
Q2	124	192.3	0.581	0.565	0.573
Q3	125	223.3	0.504	0.552	0.600
Q4 longest	125	279.2	0.600	0.680	0.656

LIMITATIONS

The evaluation has five important limits. First, PQA-L contains only 1,000 expert-labeled questions, and the held-out comparison contains 500 cases. The paired test detects the full-loop gain over question-only QA, but it does not provide enough power to establish the 1.2-point difference between retrieval and RCR. Results may also vary under another fixed split even when class proportions are preserved. The fixed release split was retained to prevent post hoc test selection, but repeated external evaluations would be needed to estimate between-split variability and to determine whether the same 0.70 operating point transfers without retuning.

Second, this is paired-abstract retrieval rather than unrestricted literature search. Every question arrives with its source abstract, and the primary retriever uses section labels that may be unavailable in unstructured documents. The findings therefore support result-focused evidence selection inside structured biomedical abstracts; they do not establish open-corpus retrieval quality, citation accuracy, or robustness to irrelevant articles.

Third, the answerer is a sparse linear classifier, not a generative biomedical language model. This choice makes the loop fast and exposes the effect of evidence selection, but it limits semantic reasoning, numerical comparison, coreference resolution, and synthesis across sentences. Domain-specific encoders such as BioBERT, SciBERT, or PubMedBERT [9]-[11] could strengthen each evidence view, while Self-RAG-style training [3] could integrate critique signals directly into generation. Those extensions may change both the baseline and the marginal value of an external critic.

Fourth, the maybe class remains effectively unmodeled. Its 11% prevalence, heterogeneous semantics, and low precision in development prevented activation of a maybe correction branch. A future system should treat maybe as a structured uncertainty decision, potentially using explicit study-quality, power, contradiction, and sufficiency features rather than a third class attached to the same linear boundary.

Fifth, the critic improves average benchmark accuracy but is not a clinical safety mechanism. It changes 20 correct held-out answers to incorrect ones, and its test correction precision is lower than its development constraint. The system does not assess evidence quality beyond the abstract, publication bias, retractions, guideline applicability, or patient-specific contraindications. Its outputs should therefore be interpreted as benchmark decisions, not medical advice.

CONCLUSION

A selective Retrieve-Critique-Revise loop improved biomedical yes/no/maybe question answering on the canonical PQA-L split. Result-section retrieval raised held-out accuracy from 55.4% to 58.8%, and the constrained critic raised it to 60.0%. The loop's main benefit was better recognition of negative findings: no recall increased from 17.8% to 33.1%, and no F1 increased from 0.284 to 0.431. The critic changed fewer than one in ten answers, recovered 26 errors, introduced 20 errors, and added six net correct cases. Paired analysis supported the complete system over question-only QA, while the smaller correction gain over retrieval alone did not reach statistical significance. The evidence therefore favors selective reflection over automatic second-pass rewriting. It also identifies the next research priority clearly: uncertainty must be modeled as evidence sufficiency and study qualification if the maybe class is to improve without sacrificing reliable yes and no decisions.

REFERENCES

- [1] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "PubMedQA: A dataset for biomedical research question answering," in Proc. EMNLP-IJCNLP, 2019, pp. 2567-2577.
- [2] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 9459-9474.
- [3] Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," arXiv:2310.11511, 2023.
- [4] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," in Proc. EMNLP, 2020, pp. 6769-6781.
- [5] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "Retrieval augmented language model pre-training," in Proc. ICML, PMLR 119, 2020, pp. 3929-3938.
- [6] S. E. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," Found. Trends Inf. Retr., vol. 3, no. 4, pp. 333-389, 2009.
- [7] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in Proc. ACM SIGIR, 1998, pp. 335-336.
- [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171-4186.
- [9] J. Lee et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234-1240, 2020.
- [10] Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in Proc. EMNLP-IJCNLP, 2019, pp. 3615-3620.
- [11] Y. Gu et al., "Domain-specific language model pretraining for biomedical natural language processing," ACM Trans. Comput. Healthcare, vol. 3, no. 1, Art. no. 2, pp. 1-23, 2022.
- [12] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," J. Mach. Learn. Res., vol. 21, no. 140, pp. 1-67, 2020.
- [13] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in Proc. EACL, 2021, pp. 874-880.
- [14] S. Yao et al., "ReAct: Synergizing reasoning and acting in language models," in Proc. ICLR, 2023.
- [15] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," in Adv. Neural Inf. Process. Syst., vol. 36, 2023, pp. 8634-8652.
- [16] Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in Adv. Neural Inf. Process. Syst., vol. 35, 2022, pp. 24824-24837.

- [17] X. Wang et al., "Self-consistency improves chain of thought reasoning in language models," in Proc. ICLR, 2023.
- [18] Madaan et al., "Self-Refine: Iterative refinement with self-feedback," in Adv. Neural Inf. Process. Syst., vol. 36, 2023.
- [19] Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. ICML, PMLR 70, 2017, pp. 1321-1330.
- [20] Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc. ICML, 2005, pp. 625-632.
- [21] Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman & Hall, 1993.
- [22] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153-157, 1947.
- [23] G. Tsatsaronis et al., "An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition," BMC Bioinformatics, vol. 16, Art. no. 138, 2015.
- [24] Pal, L. K. Umapathi, and M. Sankarasubbu, "MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering," in Proc. CHIL, PMLR 174, 2022, pp. 248-260.
- [25] Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? A large-scale open domain question answering dataset from medical exams," Appl. Sci., vol. 11, no. 14, Art. no. 6421, 2021.
- [26] Singhal et al., "Large language models encode clinical knowledge," Nature, vol. 620, pp. 172-180, 2023.
- [27] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in Proc. EMNLP, 2016, pp. 2383-2392.
- [28] Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, pp. 273-297, 1995.
- [29] G. Salton and M. J. McGill, Introduction to Modern Information Retrieval. New York, NY, USA: McGraw-Hill, 1983.
- [30] J. Nie and D. Zheng, "Ambiguity-aware HDFS log anomaly detection with retrieval-augmented failure narratives and selective refusal," J. Adv. Comput. Syst., vol. 3, no. 1, pp. 66-80, Jan. 2023, doi: 10.69987/JACS.2023.30105.
- [31] 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, 2020, doi: 10.1088/1742-6596/1651/1/012143.
- [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] G. S. Raj et al., "Impact of bilingual CS education on student learning and engagement in a data structures course," in Proc. 19th Koli Calling Int. Conf. Comput. Educ. Res., 2019, pp. 1-10, doi: 10.1145/3364510.3364518.
- [34] 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.
- [35] 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.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] Z. S. Zhong and S. Ling, "Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization," arXiv:2408.05944, Aug. 2024.
- [40] 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.
- [41] 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.
- [42] 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, Apr. 2025, doi: 10.51903/jtie.v4i1.485.
- [43] Z. S. Zhong, X. Pan, and Q. Lei, "Bridging domains with approximately shared features," in Proc. 28th Int. Conf. Artif. Intell. Statist. (AISTATS), PMLR, vol. 258, pp. 559-567, 2025.
- [44] 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.
- [45] 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.
- [46] 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.
- [47] H. Zhang, "LLM-driven CI failure diagnosis and automated repair: From GitHub Actions logs to patch recommendation," *J. Technol. Informatics Eng.*, vol. 4, no. 1, pp. 190–214, Apr. 2025, doi: 10.51903/jtie.v4i1.484.
- [48] 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.
- [49] H. Wang, Y. Ren, and X. Chang, "Layout-aware progressive PDF rendering: AI prioritization of PDF slices to reduce time-to-functional-first-frame on FUNSD," *J. Technol. Informatics Eng.*, vol. 4, no. 2, pp. 425–446, Aug. 2025, doi: 10.51903/jtie.v4i2.523.
- [50] 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.
- [51] Z. Li, S. Zhou, and Z. Zhou, "Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench," *Int. J. Graph. Des.*, vol. 3, no. 1, pp. 196–210, May 2025, doi: 10.51903/ijgd.v3i1.3715.
- [52] R. Ma, L. Zhang, and T. Song, "Computer-vision-informed visual explanation cards for autonomous-driving traffic-sign alerts: Localization, classification, and retrieved evidence on GTSDb," *Int. J. Graph. Des.*, vol. 3, no. 2, pp. 455–473, Oct. 2025, doi: 10.51903/ijgd.v3i2.3991.
- [53] J. Nie and D. Zheng, "Noisy-neighbor-aware VM degradation risk modeling with unsupervised residual fusion," *J. Adv. Comput. Syst.*, vol. 4, no. 4, pp. 112–123, Apr. 2024, doi: 10.69987/JACS.2024.40409.
- [54] Q. Xin, Z. Xu, L. Guo, F. Zhao, and B. Wu, "IoT traffic classification and anomaly detection method based on deep autoencoders," *Appl. Comput. Eng.*, vol. 69, no. 1, pp. 64–70, Jul. 2024, doi: 10.54254/2755-2721/69/20241511.
- [55] B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, "Predictive optimization of DDoS attack mitigation in distributed systems using machine learning," in *Proc. 6th Int. Conf. Computing and Data Science*, 2024, pp. 89–94, doi: 10.54254/2755-2721/64/20241350.
- [56] 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.
- [57] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of autonomous driving image detection based on RFACONV and triplet attention," *Appl. Comput. Eng.*, vol. 77, no. 1, pp. 210–217, Jul. 2024, doi: 10.54254/2755-2721/77/2024MA0067.
- [58] 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.
- [59] K. Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, "Intelligent classification and personalized recommendation of e-commerce products based on machine learning," in *Proc. 6th Int. Conf. Computing and Data Science*, 2024, doi: 10.54254/2755-2721/64/20241365.
- [60] 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.
- [61] J. Wang, Q. Xin, Y. Liu, J. Wang, and T. Yang, "Predicting enterprise marketing decision making with intelligent data-driven approaches," *J. Ind. Eng. Appl. Sci.*, vol. 2, no. 3, pp. 12–19, Jun. 2024, doi: 10.5281/zenodo.11357252.
- [62] 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.
- [63] 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.
- [64] H. Zhang, "Counterfactual learning-to-rank for ads: Off-policy evaluation on the Open Bandit Dataset," *J. Adv. Comput. Syst.*, vol. 5, no. 12, pp. 1–11, Dec. 2025, doi: 10.69987/JACS.2025.51201.