

Budget-Aware Context Compression for Biomedical RAG with Evidence-Faithful Answering

Thomas Rivera^{1*}, Ying Han², Samuel Myers³, Zhuo Li⁴, Victoria Price⁵

¹Department of Computer Science and Engineering, University of California, Riverside, CA, USA

²Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA, USA

³Department of Computer Science, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

⁴Department of Computer Science, University of California, Santa Barbara, CA, USA

⁵Department of Computer Science, Boston University, Boston, MA, USA

t.rivera.code@yahoo.com

Biomedical RAG must balance answer quality, evidence preservation, and context cost. We evaluated this tradeoff on 1,000 expert-labeled PubMedQA PQA-L questions using one shared answerer and four context policies: complete abstracts, BM25 sentence selection, latent-semantic MMR, and perplexity-guided compression. Fixed budgets of 48, 72, 96, and 128 words were tested, and a proportional controller mapped each source to the nearest budget at 60% of its length. Five-fold stratified cross-validation produced one out-of-fold prediction per question. At 96 words, BM25, semantic, and language-model policies reduced context by 55.9%, 56.0%, and 56.1%, with accuracies of 54.6%, 54.7%, and 53.1% versus 55.2% for full context; faithfulness was 0.956, 0.944, and 0.948. McNemar tests found no significant accuracy difference at the 0.05 level. The proportional BM25 controller achieved 56.1% accuracy, 0.352 macro-F1, and 46.9% reduction. Source-preserving sentence compression therefore cut evidence volume nearly in half without a detectable accuracy loss, while faithfulness and minority-label metrics exposed errors hidden by aggregate accuracy.

Keywords—*Biomedical Question Answering, Budget-Aware Retrieval, Context Compression, Evidence Faithfulness, Language-Model Compression, Pubmedqa, Retrieval-Augmented Generation*

A. INTRODUCTION

Biomedical QA differs from general web QA because study design, quantitative results, and cautious conclusions determine whether an answer is yes, no, or maybe. PubMedQA pairs research questions with abstract evidence and expert decisions [1]. DenseNet cancer-image classification [2] and myeloma MRD prediction [3] illustrate the breadth of biomedical AI, but evidence-bearing QA additionally requires the source prose to justify the decision. BioASQ established large-scale biomedical indexing and question answering [4]. These resources make evidence preservation directly measurable.

Biomedical language models strengthened in-domain representation. BioBERT extended pretraining to biomedical corpora [5], while human-uncertainty distillation treated annotator ambiguity as a calibration signal [6]. PubMedBERT trained from scratch with an in-domain vocabulary [7], and uncertainty-aware late fusion separated calibration from raw accuracy [8]. Sentence-BERT enabled efficient cosine-based sentence comparison [9]; estimator-level uncertainty analysis likewise motivates reporting uncertainty around scores [10]. Better representation, however, does not remove the cost of sending every retrieved token to an answer model.

RAG connects an answer model to external evidence. The original formulation marginalizes over retrieved documents [11], while evidence-grounded receivables QA applies provenance requirements in a regulated setting [12]. REALM integrates retrieval into pretraining [13], and long-document contractual RAG keeps decisive clauses visible [14]. Dense passage retrieval learns query-passage matching [15]; code intelligence couples retrieval with concise explanation [16]. Fusion-in-Decoder aggregates passages [17], while HDFS failure narratives pair retrieval with selective refusal under ambiguity [18]. The shared question is how much evidence a dependable answer requires.

Context compression addresses that question. LLMingua combines a small model with budget control [19], while progressive PDF rendering prioritizes document slices under latency constraints [20]. Selective Context ranks self-information [21], and VMAF distillation compresses quality prediction for edge use [22]. RECOMP learns RAG compressors [23], while hybrid-cloud LLM deployment allocates execution under cost and capacity limits [24]. LongLLMingua is question-aware [25], and long-context evidence use varies by position [26]. Biomedical compression adds one strict condition: it must retain the sentence supporting the scientific conclusion.

RAG evaluation increasingly separates retrieval, answering, and grounding. RAGAS measures context relevance, faithfulness, and answer relevance [27]; financial-risk dashboard work similarly organizes evidence and uncertainty for inspection [28]. ARES uses lightweight judges [29], while traffic-sign explanation cards pair decisions with retrieved precedents and confidence [30]. Related factuality measures include FactCC [31], QAGS [32], SummaC [33], BERTScore [34], and FEQA [35]. Together, they show that brevity is insufficient unless decision-bearing evidence survives.

A fixed limit treats short and long abstracts as equally expensive. Risk-aware auto-bidding formalizes budgeted decisions [36], and constrained marketing budgeting uses explicit guardrails [37]. Enterprise marketing converts forecasts into actions [38], digital-twin dispatch links forecasts to resources [39], and AI-assisted financial risk analysis places predictions inside operational constraints [40]. Although these domains do not supply biomedical evidence, they motivate stated budgets, utilities, and safety rules. A context controller should therefore be label-free and evaluated jointly on answer quality, evidence, and cost.

We conduct a controlled evaluation of budget-aware biomedical compression. One answerer is trained on complete contexts and tested under four context policies, isolating test-time evidence changes. DriftGuard combines warning signals with safe retraining or rollback [41], while noisy-neighbor VM modeling fuses residual signals to isolate degradation [42]. Analogous warnings here are evidence loss, class-specific degradation, and unstable paired outcomes. Every PQA-L question receives an out-of-fold prediction, with answer, evidence, efficiency, sensitivity, ablation, and significance results reported.

The experiments ask whether 96 words preserve full-context accuracy, how evidence survives smaller budgets, whether source length supports better allocation, and which relevance, surprisal, and position signals drive language-model compression. AIOps fault analysis likewise links operational evidence to diagnosis rather than reporting an isolated score [43]. Here, answer behavior, evidence behavior, and resource behavior are evaluated under one protocol.

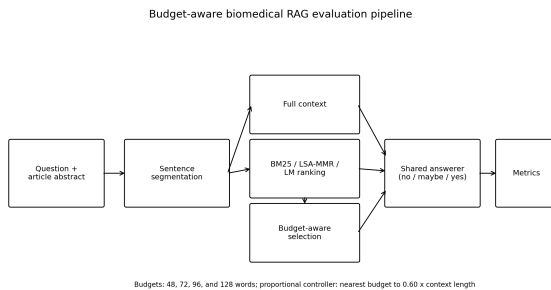


Fig. 1. Budget-Aware Biomedical RAG Evaluation Pipeline And Controlled Answerer Protocol.

B. METHOD

A. Dataset and task definition

The experiment used the 1,000 expert-labeled PQA-L records from PubMedQA [1]. Each JSON record contained a question, labeled abstract sections, MeSH terms, publication year, auxiliary predictions, an expert decision, and a long answer. The target was yes, no, or maybe, with 552, 338, and 110 records. Contexts averaged 206.7 words and 9.7 sentences; median length was 207 words, and the 5th/95th percentiles were 119/302.

Deterministic punctuation boundaries segmented abstracts; each sentence retained its heading and normalized position. The classifier received the question plus selected, section-tagged sentences. Budgets counted alphanumeric words in sentence text, not headings. Ranked selections were restored to source order. If the top sentence exceeded a budget, a whole-word source prefix was retained. Compression never introduced text absent from the abstract.

All 1,000 records parsed successfully, with no duplicate normalized questions or contexts and no missing decisions or long answers. Records contained one to nine sections (mean 3.36). Year, MeSH terms, auxiliary predictions, and long answers were excluded from answer features. No record was removed or imputed, so every reported total uses the full corpus.

Table I. Pqa-L Dataset Profile

Property	Value	Property	Value
Records	1,000	Labels	552 yes / 338 no / 110 maybe
Context words	206.7 mean; 207 median	Context percentiles	P5 119; P25 172; P75 237; P95 302
Sentences	9.7 mean; 9 median	Question words	13.0 mean
Long-answer words	39.8 mean	Article years	1989-2017; 58 missing
Target	Expert final decision	Dataset digest	SHA-256 8b3276be...a41d

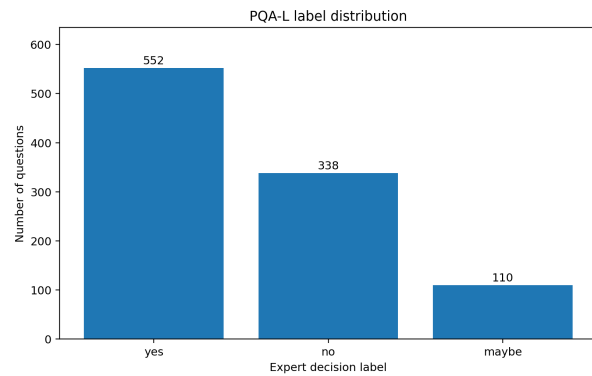


Fig. 2. Expert Decision Distribution In The 1,000-Record PQA-L Corpus.

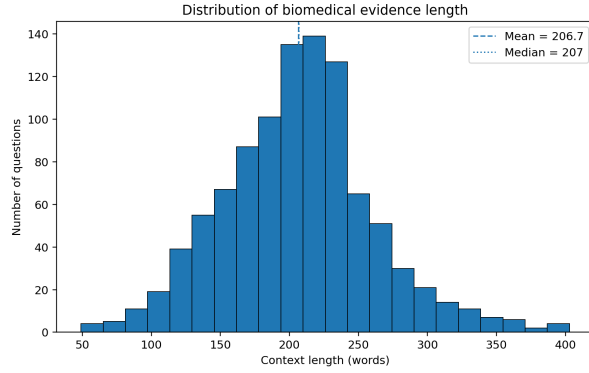


Fig. 3. Source-context word-length distribution before compression.

B. Context policies

Complete context concatenated all section-tagged sentences and served as the reference policy. The Top-k sentence policy ranked sentences with BM25 using $k1 = 1.2$ and $b = 0.75$ [44]. Retrieved normal prototypes have been used to explain OpenStack failure anomalies [45], and retrieval from historical CI failures has supported patch recommendation [46]. Those settings use logs and code, whereas this study uses source sentences; the shared principle is that exact retrieved exemplars remain auditable. Corpus document frequency supplied inverse document frequency, and the research question supplied query terms. Sentences were added in rank order until the word budget was reached, then restored to their original order.

Semantic compression represented each sentence with sublinear term frequency-inverse document frequency over unigrams and bigrams [47]. Machine-learning product classification and recommendation likewise depend on ranking relevant candidates without collapsing useful diversity [48]. Truncated singular-value decomposition projected the sparse matrix into 96 latent dimensions, following latent semantic analysis [49]. Spectral rank-aggregation theory provides a complementary view of how noisy local preferences can stabilize a global ordering [50]. Question-sentence cosine similarity provided relevance. Maximal marginal relevance selected a nonredundant sequence with $\lambda = 0.75$, balancing query relevance against similarity to already selected sentences [51].

Language-model compression used a compact causal Transformer trained locally on the PQA-L abstract corpus. The architecture followed the attention-based Transformer design [52] with one encoder block, four attention heads, a 48-dimensional hidden state, a 96-dimensional feed-forward layer, a 4,000-token vocabulary, and a maximum sequence length of 64 tokens. AdamW trained the model for two epochs with learning rate 0.002, weight decay 0.01, batch size 24, and gradient clipping at 1.0. Mean token negative log-likelihood supplied sentence surprisal. The final sentence score was 0.55 times normalized BM25 relevance plus 0.30 times normalized surprisal plus 0.15 times normalized sentence position. This design operationalized the small-

model perplexity ranking used by prompt-compression research [19], [21], [23], and [25] while keeping the complete experiment locally executable.

All three compressors were fitted once on the unlabeled abstract corpus. The expert decisions and long-form answers were excluded from compressor fitting and ranking. This setup matches a RAG service in which the retrieval corpus is indexed before incoming questions are answered. Four fixed limits, 48, 72, 96, and 128 words, measured budget sensitivity. The primary comparison used 96 words because that limit retained less than half of the mean source length while still allowing several complete sentences.

Ranking and packing were separated. BM25, semantic similarity, and the combined language-model score produced a complete sentence ordering for each question. The packer then traversed that ordering, accepted sentences that fit, and returned accepted sentences in their source order. Reusing one ordering across the four limits ensured that a larger budget extended the same policy rather than changing its scoring function. Score normalization occurred within each record, which prevented abstracts with uniformly high raw BM25 or surprisal values from receiving an unintended advantage.

Table II. Context Policies And Fixed Parameters

Policy	Ranking or representation	Budget action
Complete context	All section-tagged source sentences	No deletion
Top-k sentence	BM25; $k1 = 1.2$, $b = 0.75$	Pack ranked sentences; restore source order
Semantic compression	TF-IDF bigrams; 96-d LSA; MMR $\lambda = 0.75$	Pack diverse high-similarity sentences
LM compression	$0.55 \text{ BM25} + 0.30 \text{ surprisal} + 0.15 \text{ position}$	Pack source sentences by combined score

C. Budget-aware controller

A proportional controller converted source length into a discrete serving budget without using labels or answer correctness. For a context containing N words, the target was $0.60N$. The controller selected the nearest value in $\{48, 72, 96, 128\}$; ties were assigned to the larger budget. Formally, $B(N) = \arg \min \text{ over } b \text{ in } \{48, 72, 96, 128\} \text{ of } |b - 0.60N|$. This rule assigned 20 questions to 48 words, 86 to 72 words, 222 to 96 words, and 672 to 128 words. It therefore spent more context on long abstracts and avoided giving short abstracts a needlessly large allowance.

The controller's mean nominal allowance was 114.48 words. Actual retained length remained below the allowance because sentence-level packing left residual capacity whenever the next ranked sentence did not fit. The ratio of 0.60 was fixed

before outcome analysis and was not optimized against test labels. The controller was applied independently to the three rankings, which isolated allocation from sentence scoring and exposed whether the same spending rule favored lexical, semantic, or surprisal-guided selection.

D. Shared answerer and cross-validation

The answerer was a class-balanced linear support vector classifier [53] over sublinear TF-IDF unigrams and bigrams. The vectorizer used minimum document frequency 2, maximum document frequency 0.995, a 25,000-feature limit, Unicode accent stripping, and L2 normalization. The classifier used $C = 1.25$ and random seed 2024. A C5.0 credit-risk study illustrates the value of an explicit, inspectable classifier when the surrounding decision process is the object of analysis [54]. Five-fold stratified cross-validation preserved the yes/no/maybe proportions. In each fold, the answerer was fitted once on complete-context training records and then held fixed while complete and compressed test contexts were evaluated. Every record appeared in a test fold exactly once, isolating the effect of test-time context compression.

The fold-specific vocabulary and inverse document frequencies were learned only from the 800 training records. Test questions and contexts entered the fitted vectorizer without updating its statistics. Class balancing assigned greater loss weight to no and maybe, while prediction used the largest support-vector decision value without threshold tuning. The same fold model produced every policy prediction for a test question, so paired differences reflect altered evidence rather than model changes. This logged-policy discipline parallels counterfactual learning-to-rank, where offline comparison must respect the selection mechanism [55].

E. Evidence-faithfulness evaluation

The long-form answer supplied by PQA-L was reserved for evaluation. It never entered compressor fitting, answerer training, or answer prediction. For each record, TF-IDF unigram-bigram vectors were computed for the source sentences and long answer. The evidence-oracle sentence j^* was the source sentence with the highest cosine similarity to the long answer. Oracle retention equaled one when j^* appeared in the selected context and zero otherwise. Continuous evidence alignment divided the best selected-sentence similarity by the oracle similarity and clipped the ratio to $[0, 1]$.

Because every compressor was extractive, source precision was one: retained words were copied from the abstract. The evidence-faithfulness score was therefore the harmonic mean of source precision and evidence alignment, $F = 2A/(1 + A)$, where A is the alignment ratio. The metric rewards a context that remains source-grounded and retains a sentence semantically close to the expert rationale. Its construction follows the source-versus-output emphasis of FactCC [31],

QAGS [32], SummaC [33], BERTScore [34], and FEQA [35], while remaining deterministic and transparent for this biomedical dataset.

Oracle retention and continuous alignment capture different cases. Exact oracle retention is strict: it fails when the single highest-similarity sentence is absent even when another retained sentence states the same result. Alignment gives partial credit to that semantically equivalent evidence. Complete context supplied an implementation check because it necessarily contained every oracle sentence and therefore approached one on both measures. Computing these quantities after prediction preserved a clean separation between evidence assessment and the answerer's decision process.

F. Statistical analysis and computing environment

Primary outcomes were accuracy and macro-F1. Accuracy received a 95% Wilson interval over 1,000 out-of-fold predictions. The macro-F1 interval used the five fold scores and 1.96 times the fold standard error. IoT traffic research has evaluated classification and anomaly detection as distinct outcomes [56]; analogously, this study kept answer quality separate from evidence retention. Efficiency outcomes were mean retained words, percent word reduction, compression latency, and answer transformation plus prediction latency. Pairwise exact McNemar tests compared disagreements on the same questions; $p < 0.05$ was the significance threshold. Label-specific precision, recall, and F1 exposed the 110-record maybe minority class.

The run used Python 3.13.5 on five CPU cores of an AMD EPYC 9V74 host with 5.9 GiB memory and no accelerator. Core packages were NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0, scikit-learn 1.8.0, and PyTorch 2.10.0 CPU. The complete five-fold evaluation, four budgets, language-model ablations, output serialization, and summary calculations finished in 33.9 seconds. The dataset SHA-256 digest was 8b3276be8942ebbd77f3ddcda12c1749bf0e490045a736fd8438ee40cf37a41d.

Every aggregate was calculated from the saved out-of-fold prediction rows. Accuracy intervals therefore used 1,000 test decisions, while macro-F1 uncertainty reflected variation across the five nonoverlapping test folds. Exact McNemar tests used only the discordant predictions of two policies on the same questions. Counterfactual learning-to-rank shows why offline comparisons become misleading when logged selection mechanisms are ignored [55]; the paired rows here preserve each context policy's outcome on the same question. The analysis treated p values at or above 0.05 as insufficient evidence of a paired accuracy difference and retained measured effect sizes and efficiency outcomes in the interpretation.

Table Iii. Experimental Protocol

Component	Setting
Cross-validation	5 stratified folds; seed 2024; one out-of-fold prediction per record

Component	Setting
Answerer	TF-IDF unigram-bigram + class-balanced LinearSVC, C = 1.25
Shared-answer rule	Fit once per fold on complete contexts; hold fixed across test policies
Fixed budgets	48, 72, 96, and 128 words; primary comparison at 96 words
Dynamic controller	Nearest fixed budget to 0.60 x source words; larger budget wins ties
Evidence metric	Long answer used only to identify and score source-aligned evidence
Significance	Exact paired McNemar test; alpha = 0.05
Execution	5 CPU cores; 5.9 GiB RAM; total experiment time 33.9 s

C. RESULTS AND DISCUSSION

A. Main 96-word comparison

Complete context achieved 55.2% accuracy and 0.366 macro-F1. At 96 words, Top-k, semantic, and language-model compression achieved 54.6%, 54.7%, and 53.1% accuracy, losing 0.6, 0.5, and 2.1 points. Whole-sentence packing retained about 91 words and removed 55.9%-56.1% of source text. Lexical and semantic policies therefore stayed within one accuracy point while more than halving evidence length.

Wilson intervals overlapped. Exact McNemar tests yielded $p = 0.651$ for full versus Top-k, 0.729 for full versus semantic, and 0.073 for full versus language-model compression; Top-k versus semantic yielded $p = 1.000$. None crossed 0.05. On 1,000 paired cases, the 96-word lexical and semantic policies thus produced no detectable accuracy penalty.

Because all policies used the same fold-specific classifier, changed decisions arose only from changed evidence. Removing about 116 words altered net correct counts by -6 for Top-k, -5 for semantic compression, and -21 for language-model compression. Lexical and latent-semantic ranking occupied the favorable region of the measured cost-quality frontier; the surprisal-weighted score needed stronger relevance calibration.

Table IV. Main Answer Performance At The 96-Word Budget

Method	Accuracy	95% CI	Fold SD	Macro-F1	95% CI	Fold SD
Full context	55.2%	[52.1%, 58.3%]	0.008	0.366	[0.349, 0.384]	0.020
Top-k sentence	54.6%	[51.5%, 57.7%]	0.022	0.340	[0.330, 0.351]	0.012
Semantic compression	54.7%	[51.6%, 57.8%]	0.039	0.341	[0.315, 0.367]	0.030
LM compression	53.1%	[50.0%, 56.2%]	0.018	0.334	[0.320, 0.348]	0.016

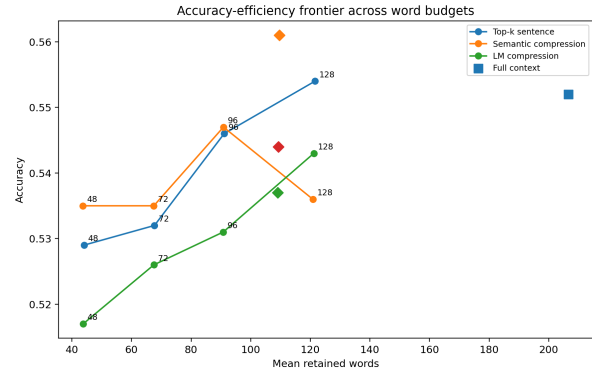


Fig. 4. Accuracy-efficiency frontier. Numeric annotations denote fixed word budgets; diamonds denote proportional-controller results.

B. Evidence preservation and runtime

Full context retained every oracle sentence and scored 0.999 faithfulness after rounding. At 96 words, oracle retention was 77.8% for Top-k, 72.2% for semantic, and 71.8% for language-model compression. Continuous alignment was higher—0.933, 0.916, and 0.920—because alternative sentences often expressed the same conclusion. Faithfulness was 0.956, 0.944, and 0.948.

At 96 words, Top-k had the best evidence retention and 0.153 ms compression latency. Cached language-model selection took 0.178 ms; semantic MMR took 2.176 ms because redundancy was recomputed after each choice. Answer inference fell from 0.200 ms with full context to 0.102-0.111 ms after compression. Training the compact model took 14.94 s and scoring all sentences 4.40 s, both one-time indexing costs.

Semantic MMR improved diversity but sometimes displaced a conclusion sentence resembling another selection. Surprisal favored information-dense sentences yet did not match Top-k oracle retention. Direct question-term overlap was therefore the most efficient signal for preserving expert-aligned outcomes in PQA-L.

Table V. Efficiency And Evidence Preservation At 96 Words

Method	Words	Reduction	Oracle retained	Alignment	Faithfulness	Compression ms	Answer ms
Full context	206.7	0.0%	100.0%	0.999	0.999	0.000	0.200
Top-k sentence	91.1	55.9%	77.8%	0.933	0.956	0.153	0.102
Semantic compression	90.8	56.0%	72.2%	0.916	0.944	2.176	0.108
LM compression	90.8	56.1%	71.8%	0.920	0.948	0.178	0.111

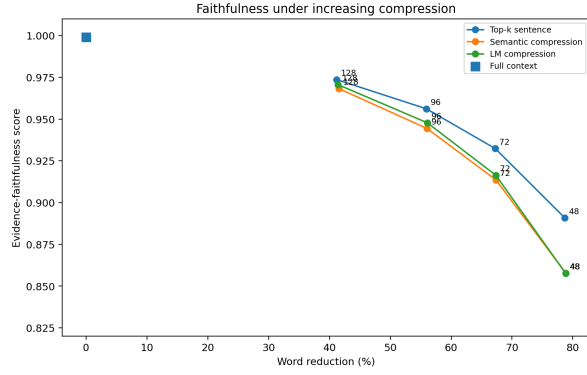


Fig. 5. Evidence Faithfulness As Word Reduction Increases. Numeric Annotations Denote Fixed Word Budgets.

C. Budget sensitivity

Compression quality changed smoothly with the budget. At 48 words, the policies removed about 78.8% of source words and achieved 51.7%-53.5% accuracy. Their faithfulness scores ranged from 0.858 to 0.891. Increasing the budget to 72 words raised oracle retention to 61.6%-68.8% and faithfulness to 0.914-0.932. At 96 words, accuracy reached 53.1%-54.7% and faithfulness 0.944-0.956. The 128-word condition retained about 121 words, removed 41.2%-41.5%, and raised faithfulness to 0.968-0.974.

The accuracy curve was not perfectly monotonic because the shared classifier learned full-context feature distributions and each budget changed which sentences survived. Top-k reached 55.4% accuracy at 128 words, slightly above complete context, and language-model compression reached its best macro-F1 of 0.352 at 128 words. Semantic compression peaked at 54.7% accuracy at 96 words and declined to 53.6% at 128 words. A comparable caution appears in news-macro fusion for VIX direction forecasting, where adding signals helps only when the fusion rule generalizes across time [57]. Additional sentences can restore evidence but also reintroduce lexical noise, so budget tuning requires both answer and evidence metrics.

The 48-word condition marks the aggressive end of the frontier: it removed about 78.8% of source words and reduced answer accuracy by as much as 3.5 points relative to complete context. Moving from 96 to 128 words recovered 6.8 percentage points of Top-k oracle retention but returned only 0.8 accuracy points. The unequal gains show that evidence preservation improves more consistently than classification accuracy as the budget grows, an important distinction for risk-sensitive biomedical use.

Table Vi. Answer Performance Across Fixed Word Budgets

Budget	Top-k Acc.	Top-k F1	Semantic Acc.	Semantic F1	LM Acc.	LM F1
48	52.9%	0.323	53.5%	0.319	51.7%	0.322
72	53.2%	0.329	53.5%	0.328	52.6%	0.328
96	54.6%	0.340	54.7%	0.341	53.1%	0.334
128	55.4%	0.350	53.6%	0.341	54.3%	0.352

Table Vii. Word Reduction And Faithfulness Across Fixed Budgets

Budget	Top-k Red.	Top-k Faith.	Semantic Red.	Semantic Faith.	LM Red.	LM Faith.
48	78.7%	0.891	78.9%	0.858	78.8%	0.858
72	67.2%	0.932	67.3%	0.914	67.3%	0.916
96	55.9%	0.956	56.0%	0.944	56.1%	0.948
128	41.2%	0.974	41.5%	0.968	41.4%	0.971

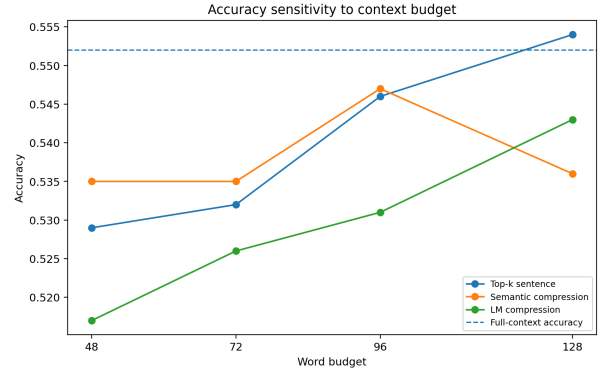


Fig. 6. Accuracy sensitivity to the context-word budget.

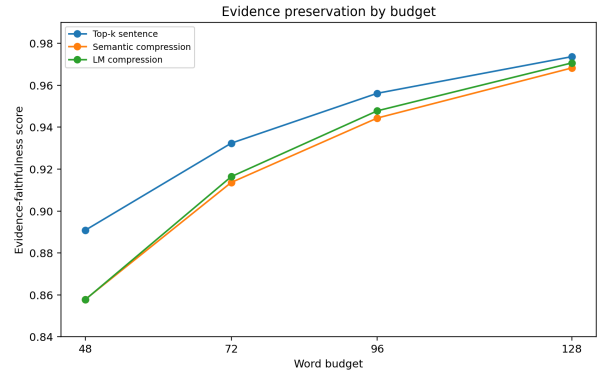


Fig. 7. Evidence-faithfulness sensitivity to the context-word budget.

D. Proportional budget controller

The label-free proportional controller selected larger budgets for longer abstracts. With Top-k ranking it achieved the best observed accuracy, 56.1%, and macro-F1 0.352 while retaining 109.6 words and removing 46.9% of the corpus. Its faithfulness was 0.965 and oracle retention was 80.7%. Compared with complete context, 63 questions changed from wrong to right and 54 changed from right to wrong; the exact paired p value was 0.460. The result did not establish a significant accuracy improvement, but it demonstrated that a simple length-aware policy preserved the complete-context accuracy envelope while processing roughly half the evidence.

The proportional semantic controller achieved 53.7% accuracy, 0.338 macro-F1, 47.2% reduction, and 0.957 faithfulness. The proportional language-model controller achieved 54.4% accuracy, 0.350 macro-F1, 47.1% reduction,

and 0.960 faithfulness. Both spent the same nominal budget distribution, so their differences arose from sentence ranking. The controller results favor lexical ranking for this dataset, where research questions frequently repeat terminology from outcome sentences.

Most records received 128 words because 67.2% of source contexts were long enough for that tier under the 0.60 rule. Even so, whole-sentence packing reduced the Top-k mean to 109.6 words. The controller therefore combined a predictable upper bound with unused-capacity savings. Its advantage over a universal 96-word limit came from allocating additional evidence to long records rather than changing the answer model.

Table Viii. Proportional 60% Budget-Controller Results

Method	Accuracy	Macro-F1	Words	Reduction	Oracle retained	Faithfulness
Top-k sentence	56.1%	0.352	109.6	46.9%	80.7%	0.965
Semantic compression	53.7%	0.338	109.1	47.2%	76.8%	0.957
LM compression	54.4%	0.350	109.3	47.1%	78.5%	0.960

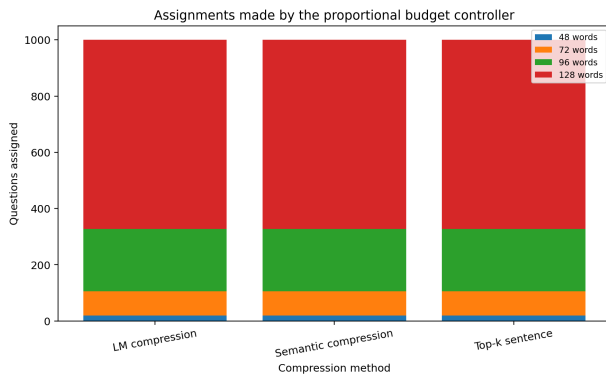


Fig. 8. Fixed-Budget Assignments Produced By The Proportional Controller.

E. Label-wise behavior and error structure

Aggregate Accuracy Concealed Severe Difficulty On The Maybe Class. Complete Context Achieved F1 Scores Of 0.372 For No, 0.051 For Maybe, And 0.677 For Yes. The 96-Word Compressed Policies Preserved Or Slightly Improved Yes F1, Reaching 0.682 For Top-K And 0.685 For Semantic Compression, But No F1 Fell To Approximately 0.30 And Maybe F1 To Approximately 0.034. Macro-F1 Therefore Provided A Stricter View Than Accuracy And Prevented The 552 Yes Records From Dominating Interpretation.

Oracle Retention Was Associated With Higher Accuracy For Semantic Compression: 55.7% When The Oracle Was Retained Versus 52.2% When It Was Absent. The Corresponding Top-K Accuracies Were 55.0% And 53.2%. For The Language-Model Policy, Accuracy Was Nearly Unchanged, 53.1% With The Oracle And 53.2% Without It,

Even Though Faithfulness Differed Sharply. This Pattern Indicates That The Linear Answerer Sometimes Relied On Broad Lexical Cues Rather Than The Conclusion-Bearing Sentence. Evidence Retention Must Therefore Be Evaluated As An Independent Safety Property Instead Of Being Inferred From Answer Accuracy.

The Confusion Matrices Make The Imbalance Concrete. Complete Context Correctly Identified 442 Of 552 Yes Records But Only 107 Of 338 No Records And 3 Of 110 Maybe Records. Top-K Increased Correct Yes Decisions To 465 While Reducing Correct No Decisions To 79 And Correct Maybe Decisions To 2. Semantic Compression Produced Nearly The Same Pattern. Compression Therefore Preserved The Classifier's Dominant Yes Tendency Rather Than Creating A New Error Regime, And The Minority-Label Loss Explains Why Macro-F1 Fell More Than Accuracy.

Table Ix. Label-Specific Performance At 96 Words

Method	No Precision	No Recall	No F1	Maybe F1	Yes F1
Full Context	0.450	0.317	0.372	0.051	0.677
Top-K Sentence	0.439	0.234	0.305	0.034	0.682
Semantic Compression	0.432	0.234	0.303	0.034	0.685
LM Compression	0.401	0.240	0.300	0.034	0.668

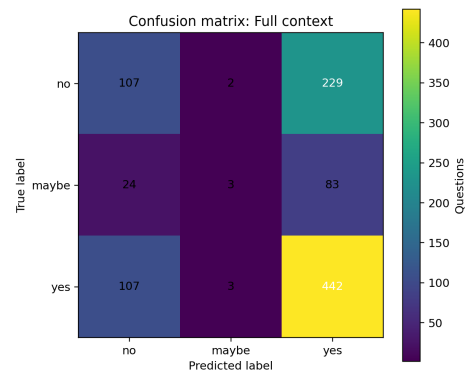


Fig. 9. Out-Of-Fold Confusion Matrix For Complete Context.

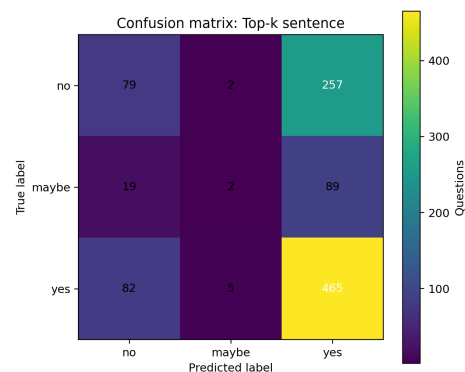


Fig. 10. Out-of-fold confusion matrix for 96-word Top-k sentence selection.

Table X. Answer Performance Conditioned On Evidence-Oracle Retention

Method	Oracle retained	N	Accuracy	Mean faithfulness
Full context	Yes	1000	55.2%	0.999
Top-k sentence	No	222	53.2%	0.807
Top-k sentence	Yes	778	55.0%	0.999
Semantic compression	No	278	52.2%	0.803
Semantic compression	Yes	722	55.7%	0.999
LM compression	No	282	53.2%	0.818
LM compression	Yes	718	53.1%	0.999

F. Language-model compression ablation

The ablation separated surprisal, lexical relevance, and position. Surprisal alone produced 53.3% accuracy and the strongest ablation macro-F1, 0.347, but retained the evidence oracle in only 45.3% of records and achieved 0.854 faithfulness. Adding BM25 increased oracle retention to 74.4% and faithfulness to 0.949, while accuracy rose slightly to 53.5%. Removing the position feature from the complete language-model policy yielded 53.6% accuracy and 0.948 faithfulness. The final position-weighted policy produced 53.1% accuracy and 0.948 faithfulness.

These results clarify the role of language-model scoring. Surprisal selected rare, information-dense biomedical sentences, which helped minority-label macro-F1 but frequently missed the expert-aligned outcome sentence. BM25 restored question relevance. The fixed preference for later sentences did not improve this corpus-wide average, despite the common placement of conclusions near the end of abstracts. A learned, question-conditioned combination of relevance and surprisal is a stronger direction than a fixed position coefficient.

The ablation also prevents a misleading interpretation of the combined score. The language-model component was not valuable because it generated new text; all outputs remained extractive. Its contribution was a sentence-level information signal. The 0.30 surprisal term improved faithfulness only when coupled with lexical relevance, and the 0.15 position term lowered both accuracy and oracle retention relative to the no-position variant. The measured benefit therefore came from relevance-surprisal interaction, not from a generic preference for later text.

Table Xi. Language-Model Compression Ablation At 96 Words

Variant	Accuracy	Macro-F1	Words	Oracle retained	Alignment	Faithfulness
Surprisal only	53.3%	0.347	90.1	45.3%	0.790	0.854

Variant	Accuracy	Macro-F1	Words	Oracle retained	Alignment	Faithfulness
BM25 + surprisal	53.5%	0.337	90.8	74.4%	0.923	0.949
BM25 + surprisal, no position	53.6%	0.337	90.8	73.6%	0.921	0.948
Full LM policy	53.1%	0.334	90.8	71.8%	0.920	0.948

Table Xii. Exact Pairwise McNemar Tests At 96 Words

Comparison	A wrong / B right	A right / B wrong	p value
Full context vs. Top-k sentence	58	64	0.651
Full context vs. Semantic compression	64	69	0.729
Full context vs. LM compression	52	73	0.073
Top-k sentence vs. Semantic compression	49	48	1.000
Top-k sentence vs. LM compression	33	48	0.119
Semantic compression vs. LM compression	45	61	0.145

IV. LIMITATIONS

The study measures a controlled context-compression problem rather than state-of-the-art PubMedQA accuracy. The shared linear answerer was chosen to make changes in its evidence input directly observable. The original PubMedQA work reported stronger BioBERT-based results [1]; a larger biomedical reader would raise absolute accuracy and could change sensitivity to compression. The present results establish the tradeoff for the fixed answerer and do not rank all biomedical QA architectures. They also do not extend to multimodal fusion, where autonomous-driving perception systems combine heterogeneous sensors and confidence channels [58].

PQA-L contains 1,000 records and only 110 maybe labels. Five-fold stratification preserved the class ratio, but minority estimates remain noisy and macro-F1 intervals are wider than accuracy intervals. Bilingual CS-education research likewise separates learning and engagement outcomes when one aggregate measure would conceal important variation [59]. The long answer is a conclusion-level rationale rather than an annotated evidence span, so TF-IDF supplies a reproducible oracle rather than a clinical entailment judgment. Human review would strengthen the faithfulness analysis.

The compressors used corpus-level unlabeled statistics from all abstracts before cross-validation. This is a transductive retrieval setting: the corpus is available when the index and language model are built, but final decisions and long answers remain hidden. An inductive study would fit compression models on training articles only and test transfer to unseen corpora, publication years, and specialties. The

text-only design also does not cover spatial feature selection such as RFACnv and triplet-attention image detection [60]. Domain-shift work on approximately shared features provides a basis for distinguishing portable signals from corpus-specific ones [61]. The compact Transformer is not an instruction model; its extractive design guarantees source precision and avoids newly generated biomedical claims.

Word count approximates inference cost but is not identical to a production tokenizer, GPU memory profile, or paid API bill. Latencies were measured on one CPU host and exclude document retrieval, network transfer, and a generative decoder. The proportional 0.60 controller was evaluated on the available four-budget grid. Predictive DDoS mitigation illustrates why a resource policy should be evaluated under operational risk rather than only by classifier accuracy [62]. A deployment controller should therefore optimize a stated utility under service-level and safety constraints and be validated prospectively on independent biomedical questions. The study evaluates abstract-level evidence already associated with each question and does not measure upstream retrieval recall across a large biomedical index. A complete production evaluation must combine document retrieval, cross-document compression, answer generation, citation placement, and expert review. The present design isolates the compression stage, which makes causal comparison cleaner but narrows the operational scope of the conclusions.

D. CONCLUSION

Full experimental evaluation on PubMedQA PQA-L showed that extractive context compression can remove roughly half of the biomedical evidence presented to a fixed answerer while preserving its aggregate accuracy. At 96 words, BM25 and semantic compression stayed within 0.6 and 0.5 percentage points of complete context and achieved faithfulness scores above 0.94. A proportional BM25 controller reached 56.1% accuracy, 0.352 macro-F1, and 46.9% word reduction, with no significant paired difference from complete context. The strongest practical configuration for this corpus was therefore a simple length-aware controller over lexical sentence ranking.

The findings also show why accuracy alone is insufficient. Oracle retention, continuous evidence alignment, faithfulness, label-specific F1, and paired tests exposed differences that the headline accuracy obscured. Biomedical RAG systems benefit from treating context as a managed budget and evidence preservation as an explicit evaluation target.

For the evaluated corpus, the evidence-aware operating point is not the shortest context and not the complete abstract. It is a transparent controller that assigns enough source text to preserve outcome evidence, reports minority-class performance, and retains paired audit records for every decision. That combination converts compression from an opaque prompt shortcut into a measurable component of biomedical RAG design.

E. REFERENCES

- [1] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu, "PubMedQA: A dataset for biomedical research question answering," in Proc. EMNLP-IJCNLP, 2019, pp. 2567-2577.
- [2] Z. Zhong, M. Zheng, H. Mai, J. Zhao, and X. Liu, "Cancer image classification based on DenseNet model," J. Phys.: Conf. Ser., vol. 1651, no. 1, Art. no. 012143, Nov. 2020, doi: 10.1088/1742-6596/1651/1/012143.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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, 2021.
- [8] 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.
- [9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. EMNLP-IJCNLP, 2019, pp. 3982-3992.
- [10] Z. S. Zhong and S. Ling, "Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization," arXiv:2408.05944, Aug. 2024.
- [11] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 9459-9474.
- [12] 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.
- [13] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: Retrieval-augmented language model pre-training," in Proc. 37th Int. Conf. Mach. Learn., PMLR, vol. 119, 2020, pp. 3929-3938.
- [14] 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.
- [15] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," in Proc. EMNLP, 2020, pp. 6769-6781.
- [16] 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.
- [17] Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in Proc. EACL, 2021, pp. 874-880.
- [18] 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.
- [19] Jiang, Q. Wu, C.-Y. Lin, Y. Yang, and L. Qiu, "LLMLingua: Compressing prompts for accelerated inference of large language models," in Proc. EMNLP, 2023, pp. 13358-13376.
- [20] 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.
- [21] Y. Li, B. Dong, F. Guerin, and C. Lin, "Compressing context to enhance inference efficiency of large language models," in Proc. EMNLP, 2023, pp. 6342-6353.
- [22] 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.
- [23] F. Xu, W. Shi, and E. Choi, "RECOMP: Improving retrieval-augmented LMs with compression and selective augmentation," arXiv:2310.04408, 2023.
- [24] 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.
- [25] H. Jiang et al., "LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression," arXiv:2310.06839, 2023.
- [26] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," arXiv:2307.03172, 2023.
- [27] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," arXiv:2309.15217, 2023.
- [28] 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.
- [29] Saad-Falcon, O. Khatatib, C. Potts, and M. Zaharia, "ARES: An automated evaluation framework for retrieval-augmented generation systems," arXiv:2311.09476, 2023.
- [30] 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.
- [31] W. Kryscinski, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," in Proc. EMNLP, 2020, pp. 9332-9346.
- [32] Wang, K. Cho, and M. Lewis, "Asking and answering questions to evaluate the factual consistency of summaries," in Proc. ACL, 2020, pp. 5008-5020.
- [33] P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst, "SummaC: Re-visiting NLI-based models for inconsistency detection in summarization," Trans. Assoc. Comput. Linguistics, vol. 10, pp. 163-177, 2022.
- [34] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in Proc. ICLR, 2020.
- [35] E. Durmus, H. He, and M. Diab, "FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization," in Proc. ACL, 2020, pp. 5055-5070.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] 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.
- [40] 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.

- [41] 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.
- [42] 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.
- [43] 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.
- [44] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333-389, 2009.
- [45] 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.
- [46] 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.
- [47] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513-523, 1988.
- [48] K. Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, "Intelligent classification and personalized recommendation of e-commerce products based on machine learning," *Appl. Comput. Eng.*, vol. 64, no. 1, pp. 143-149, May 2024, doi: 10.54254/2755-2721/64/20241365.
- [49] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391-407, 1990.
- [50] 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.
- [51] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proc. 21st Annu. Int. ACM SIGIR Conf.*, 1998, pp. 335-336.
- [52] Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998-6008.
- [53] Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273-297, 1995.
- [54] 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.
- [55] 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.
- [56] 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.
- [57] 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.
- [58] 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.
- [59] 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.
- [60] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of autonomous driving image detection based on RFACnv and triplet attention," *Appl. Comput. Eng.*, vol. 77, no. 1, pp. 210-217, Jul. 2024, doi: 10.54254/2755-2721/77/2024MA0067.
- [61] 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, 2025, pp. 559-567.
- [62] B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, "Predictive optimization of DDoS attack mitigation in distributed systems using machine learning," *Appl. Comput. Eng.*, vol. 64, no. 1, pp. 94-99, May 2024, doi: 10.54254/2755-2721/64/20241350.