Stout in Business and Management (SBM)
ISSN: 2979-2827
DOI: https://doi.org/10.61424/y3my3g85
Journal Homepage: https://stoutjournals.org/index.php/SBM

Trust-Calibrated Enterprise Service Copilots: Cost-Aware Agentic RAG for Customer-Support Knowledge Management

Kevin Wu
Computer Science, North Carolina State University, Raleigh, NC, USA
Corresponding Author: Kevin Wu, **E-mail**: k_wu_ncsu@outlook.com

ARTICLE INFORMATION

Submitted: March 23, 2026
Accepted: June 29, 2026
Published: August 01, 2026
Volume: 2
Issue: 1

KEYWORDS

Enterprise knowledge management; retrieval-augmented generation; customer support; selective prediction; reciprocal-rank fusion; latent semantic analysis; cost-aware routing.

ABSTRACT

Enterprise service copilots must retrieve relevant knowledge, provide traceable answers, control unsupported responses, and operate within practical latency and context budgets. This study evaluated these requirements on WixQA using the complete 6,221-article knowledge base, 6,221 Synthetic questions, 200 ExpertWritten questions, and 200 Simulated questions. The pipeline compared Okapi BM25, a 256-dimensional Dense-LSA retriever, reciprocal-rank fusion, a supervised logistic reranker, a top-1 article-match confidence gate, budget-aware stopping, and deterministic extractive evidence cards. Learned components used grouped five-fold out-of-fold evaluation on Synthetic, followed by frozen external evaluation. On the pooled external sets, Recall@5 was 0.4704 for BM25, 0.3704 for Dense-LSA, 0.4946 for hybrid retrieval, and 0.4838 after reranking. The adaptive policy answered 56.5% of ExpertWritten and 48.5% of Simulated requests; answered-case token F1 was 0.3051 and 0.2927, respectively. Median retrieved context fell to 71 and 63 analysis tokens. The Synthetic gate attained 0.0999 selective top-1 error at 0.8883 coverage, but frozen external error rose to 0.5664 and 0.6907. All tested financial scenarios produced negative ROI, although referral yielded the least negative base-case ratio. Hybrid retrieval improved external evidence recall, but synthetic confidence and benchmark efficiency did not support autonomous service. Representative calibration, explicit referral, and operational measurement remain necessary.

1. Introduction

Generative artificial intelligence has become an important component of knowledge work, but organizational value depends on how a system is integrated into actual workflows rather than on model capability alone. Field evidence from customer-support work shows that generative assistance can raise productivity, particularly for less experienced workers, while changing how expertise is distributed inside an organization (Brynjolfsson et al., 2025). Information-systems research likewise separates system quality, information quality, use, user satisfaction, and net benefits instead of treating technical accuracy as a complete success measure (DeLone & McLean, 2003). An enterprise service copilot must therefore do more than produce fluent language. It must locate the correct internal knowledge, disclose its evidence, recognize inadequate evidence, and justify the resources it consumes.

Retrieval-augmented generation (RAG) connects a language interface to an external corpus and separates current organizational knowledge from model parameters (Lewis et al., 2020). The architecture is attractive for customer support because policies, product behavior, feature availability, and troubleshooting instructions change frequently. The separation also creates linked failure points. Lexical retrieval can miss paraphrases; semantic retrieval can return conceptually related but operationally wrong documents; reranking can distort a strong first-stage list; and an overconfident acceptance policy can expose users to responses whose cited document is not the intended support evidence.

WixQA supplies a relevant test environment because its questions are linked to a fixed enterprise knowledge-base snapshot rather than to a general web corpus (Cohen et al., 2025; Wix.com, 2025). Its ExpertWritten, Simulated, and Synthetic subsets

represent materially different workloads. ExpertWritten contains real support questions with expert-authored answers, Simulated contains expert-validated questions distilled from support dialogues, and Synthetic pairs every corpus article with an automatically constructed question. The design supports abundant weak supervision and a direct test of whether learned behavior transfers to more realistic requests.

Trust depends on calibrated reliance. Confidence displays and explanations do not automatically improve decisions; they must help users discriminate between reliable and unreliable recommendations (Zhang et al., 2020). For a service copilot, this requires separating retrieval recall from service coverage. Retrieval recall measures how much annotated evidence appears in a ranked list. Service coverage measures how often the controller answers instead of referring a request. A gate can reduce coverage without controlling error when its confidence scale shifts across question sources.

Four research questions structured the evaluation. RQ1 asked how lexical, latent-semantic, fused, and reranked retrieval compared across the three WixQA question sources. RQ2 asked how retrieval choices affected extractive answer overlap and article-ID attribution. RQ3 asked whether a gate trained on Synthetic preserved its selective top-1 article error target on ExpertWritten and Simulated. RQ4 asked whether adaptive stopping improved measured resource efficiency and scenario-based net value. The study evaluates the full corpus and all questions, derives retrieval, answer, citation, confidence, latency, subgroup, and economic results from common query-level records, and treats "agentic" as a fixed controller for depth, stopping, and referral rather than an autonomous language-model agent.

2. Literature Review

2.1 Enterprise and Document-Grounded Retrieval

RAG combines parametric language knowledge with nonparametric evidence for knowledge-intensive tasks (Lewis et al., 2020). Enterprise use adds document governance, versioning, access controls, latency, and audit requirements. RAGOps frames these concerns as lifecycle-management problems extending beyond offline answer accuracy (X. Xu et al., 2025). Document-grounded dialogue benchmarks established the importance of authoritative evidence: Doc2Dial modeled goal-oriented conversations grounded in documents (Feng et al., 2020), and MultiDoc2Dial extended the setting to conversations that cross document boundaries (Feng et al., 2021). MultiHop-RAG further showed that multi-step requests may require several documents rather than one semantically similar passage (Tang & Yang, 2024). WixQA brings these issues into a single-company support corpus with article-level annotations (Cohen et al., 2025; Wix.com, 2025).

Adjacent operations research reinforces the need for escalation and monitoring. Software-operations frameworks combine domain knowledge, retrieval, and workflow-specific evaluation (J. He et al., 2026), while AI-driven DevOps research emphasizes adaptation across heterogeneous cloud conditions (Khan et al., 2025). These tasks are broader than support retrieval, yet they motivate lifecycle controls instead of one-time benchmark optimization.

2.2 Sparse, Dense, Hybrid, and Reranked Retrieval

BM25 remains a strong sparse baseline because probabilistic term weighting balances document frequency, term-frequency saturation, and document length (Robertson & Zaragoza, 2009). This is valuable in product support, where feature names, interface labels, error strings, and settings are highly discriminative. Dense retrieval can bridge vocabulary differences, but dense representation is not synonymous with a particular pretrained encoder. Latent semantic analysis projects term-document structure into a lower-dimensional space with truncated singular-value decomposition (Deerwester et al., 1990). E5 instead learns transferable neural embeddings through weakly supervised contrastive pretraining (Wang et al., 2022). The experiment used the former approach; E5 provides the neural-embedding context for future comparison.

Sparse and dense rankings often contain complementary evidence. Reciprocal-rank fusion (RRF) combines positions without normalizing incomparable scores (Cormack et al., 2009). Candidate reranking then applies query-document interaction signals after broad retrieval. Cross-encoder work demonstrates the value of joint query-passage processing (Nogueira & Cho, 2019). Here, logistic regression over retrieval, overlap, length, and document-type features preserved low local resource requirements while testing whether Synthetic supervision improved ordering.

2.3 Adaptive Control, Evaluation, and Trust

Adaptive systems vary retrieval effort by request. Self-RAG integrates retrieval and self-critique within generation (Asai et al., 2024); Adaptive-RAG routes questions according to estimated complexity (Jeong et al., 2024); and interleaved retrieval can improve multi-step evidence collection when later searches depend on earlier findings (Trivedi et al., 2023). ReAct alternates reasoning and environment actions (Yao et al., 2023), whereas τ-bench evaluates agents that interact with tools, users, and policies over multiple turns (Yao et al., 2024). The present controller is narrower: it selects one document, three documents, or referral without iterative model reasoning or external actions.

Cost-aware routing provides a complementary management lens. FrugalGPT uses cascades to balance model quality and expense (L. Chen et al., 2024); Hybrid LLM estimates quality-cost trade-offs (Ding et al., 2024); and RouteLLM learns routing from preference data (Ong et al., 2025). Those studies route among models, whereas this study routes evidence depth and referral. Both reject the assumption that every request requires the same computational path.

RAG evaluation is necessarily modular. RAGAS evaluates answer relevance, faithfulness, and context quality (Es et al., 2024); ARES uses learned judges trained with synthetic supervision (Saad-Falcon et al., 2024); and RAGChecker decomposes retrieval and response errors (Ru et al., 2024). Citation research emphasizes whether sources support responses rather than merely counting citation markers (Gao et al., 2023). The present metrics are deliberately narrower: normalized answer overlap follows the SQuAD token-F1 tradition (Rajpurkar et al., 2016), and citations are unique article IDs compared with dataset annotations.

Calibration concerns whether predicted confidence corresponds to observed correctness (Guo et al., 2017). Selective classification converts confidence into a risk-coverage decision by abstaining on lower-confidence cases (Geifman & El-Yaniv, 2017). In customer support, abstention becomes referral. The decisive issue is whether a threshold learned from article-derived questions preserves its target after the language, evidence multiplicity, and construction process change.

2.4 Knowledge Continuity, Personalization, and Transfer

Enterprise support is not only a one-shot retrieval problem. A deployed copilot must decide what interaction state is worth retaining, retrieve it only when relevant, and retire stale or explicitly deleted information. Confidence-gated memory writing and time-aware retrieval provide one formulation of this lifecycle (Sun et al., 2023). Closely related work treats preference learning as a privacy problem, using federated topic models for knowledge-grounded chat and differentially private estimation for cookieless personalization (Kuo et al., 2025; Bai, Wang, et al., 2026). Self-supervised customer representations extend the same principle to preference-sensitive service journeys (Xin, 2026f). For an enterprise help center, these studies imply that personalization should be separable from the authoritative corpus and governed as an auditable data layer.

Knowledge access also varies by domain, language, and query construction. Retrieval-summary integration for code intelligence shows why a system must be both findable and explainable (Li, 2024), while spectral rank aggregation and financial-search reranking illustrate the need to preserve stable ordering and domain terminology (Zhong & Ling, 2024; Meng et al., 2026). Uncertainty-aware late fusion further shows why confidence should inform how heterogeneous rankings are combined (Xin, 2025c). Trust-calibrated multilingual RAG adds the requirement that retrieval and confidence remain comparable across languages (Chen & Xu, 2026). More generally, approximately shared-feature methods formalize how transfer can work when domains overlap only partially (Zhong, Pan, et al., 2025). These findings motivate evaluating WixQA's question sources separately rather than assuming that synthetic and human requests are interchangeable.

2.5 Evidence Grounding, Provenance, and Retrieval Policy

Operational RAG research consistently treats private-document grounding as a first-class design constraint. Evidence-grounded DevOps question answering and therapist-facing multi-turn retrieval connect recommendations to runbooks, dialogue history, and operational records rather than relying on model memory (Zhang et al., 2024; Zhang & Zhang, 2025a). Evidence-chain approaches further decompose support into claim detection, source attribution, and deterministic provenance (Li et al., 2024; Jin, 2025a). Unanswerable-question work adds retrieval coverage and calibrated abstention to this chain (Zhong, Chen, et al., 2025). Together, this literature supports the present separation of retrieval recall, article attribution, answer overlap, and referral.

Complex requests create a policy problem because evidence depth should depend on the question. Budgeted multi-hop retrieval explicitly trades additional searches against compositional answer needs (Su et al., 2025). Scientific-search evidence cards and accounting disclosure cards make provenance legible at the interface (Jin, 2025b; K. Zhang et al., 2025), while numerical guardrails and evidence-constrained financial agents show that retrieved text must still be checked against task-specific rules (Z. Li et al., 2026; Zhou et al., 2026). Narrative-aware claim verification and multi-regulation product counsel broaden the same pattern to scientific and legal decisions (Su, Chen, & Qian, 2026; J. Li & A. Zhou, 2026). The common implication is that more context is useful only when each added item contributes attributable evidence.

Failure narratives offer a useful operational analogue. Retrieved normal prototypes can explain OpenStack anomalies (Xin, 2025a), and retrieval-grounded HDFS analysis can attach deterministic narratives to detected failures (Sun et al., 2026). Cloud troubleshooting copilots extend retrieval to runbook generation and command-safety checks, while incident visualization cards translate distributed logs into reviewable evidence (B. Zhang et al., 2026; Nie et al., 2026). Review-grounded recommendation similarly evaluates whether an explanation is faithful to its cited customer evidence (Chang et al., 2026). Machine-learning prediction of RAG retrieval quality represents a complementary monitoring layer for these pipelines (R. Zhang et al., 2025).

2.6 Calibration, Safety, and Selective Service

Calibration and selective rejection recur in regulated and high-impact decision systems. Credit-default tiering combines probability calibration with uncertainty-driven rejection (Chen et al., 2023), and model-risk workflows add distribution-free uncertainty and local explanations (Jin et al., 2024). Cost-sensitive bankruptcy prediction emphasizes that the consequence of an error depends on the class and decision context (Y. Chen et al., 2024). Selective resume-job matching applies the same logic to screening, while counterfactual credit explanations connect model decisions to actionable recourse (Zhou, Jin, & Zhao, 2025; Xin, 2026c). Risk-calibrated safety cards then show how uncertainty can be communicated to nontechnical users (Zhou, Wang, & Chang, 2025). These studies support reporting coverage together with answered-case quality.

Service copilots also operate under adversarial input. Continual red-teaming treats guardrails as policies that must be updated as new jailbreaks appear (Zheng et al., 2023). Evidence-based self-verification and behavior-level refusal seek to preserve useful responses while controlling unsafe behavior (D. Zheng et al., 2024; Zheng & Li, 2024). Privacy and data-integrity risk cards make prompt injection visible to human overseers, and interpretable CloudTrail sequence models reconstruct attack stages from operational evidence (Su, Rao, & Ma, 2026; Bai, Chen, et al., 2026). These concerns are distinct from retrieval accuracy: a document match does not establish that a generated action is safe.

Security detection research reinforces the value of layered, evidence-retaining controls. TinyLLM-assisted IoT detection and language-guided feature selection show how compact systems can combine semantic guidance with network indicators (Lu & Zhou, 2024; Li & Lu, 2025). Host-based sequence localization retains forensic windows rather than returning only a label (Xin, 2026d), while deep autoencoding and predictive DDoS mitigation address traffic classification and response planning (Xin et al., 2024; Wang et al., 2024). For customer support, the transferable lesson is architectural: confidence gating, content safety, authorization, and incident logging require separate controls and evidence.

2.7 Operational Capacity, Routing, and Lifecycle Control

A support copilot shares infrastructure with other probabilistic services, so retrieval quality cannot be separated from capacity risk. Foundation-model forecasting with calibrated uncertainty addresses heterogeneous GPU clusters (He et al., 2023), while cold-start forecasting studies workload shift for new tenants (He et al., 2025). Uncertainty-aware late fusion provides a complementary view of confidence-guided combination across heterogeneous signals (Xin, 2025c). Conformal demand envelopes constrain oversubscription risk, and semantic multi-horizon forecasts tie demand to operational risk curves (S. Chen et al., 2024; Zhao et al., 2025). GPU-memory prediction provides a complementary task-level estimate (Wang et al., 2025). Collectively, this work motivates measuring latency and context usage while avoiding claims that offline speed alone establishes production readiness.

Incident operations likewise require combining heterogeneous evidence. Multi-source root-cause attribution fuses metrics, logs, and traces around a fault window (Liu et al., 2023), and residual fusion models noisy-neighbor VM degradation without relying on a single signal (Nie & Zheng, 2024). Capacity dashboards make forecast risk visible to operators (Liu et al., 2025). Profit-aware spot-GPU admission and power-aware inventory planning then translate forecasts into resource decisions and policy explanations (Zhao et al., 2026; S. He et al., 2026). The analogous customer-support decision is whether to answer, retrieve more evidence, or refer the request.

Routing policies connect infrastructure cost to service reliability. Hybrid cloud deployment routes requests between local and remote models according to confidence and cost (Xin, 2025b). Cost-aware AIOps routing assigns different log tasks to different analysis paths, and conformal alert control maps anomaly scores to auditable incident tickets (C. Li et al., 2026; Xin, 2026e). Self-supervised LogBERT-style detection supplies a stronger sequence-modeling baseline for that lifecycle (Xin, 2026g). Trajectory reliability prediction adds tool errors, looping, and recovery behavior as monitoring signals for generalist agents (Zhong et al., 2026). These approaches motivate the present controller's explicit stopping and referral states, but also show why a production agent would need broader trajectory-level monitoring.

2.8 Managerial Value and Human Oversight

Economic evaluation must separate prediction from policy value. Uplift modeling targets incremental response rather than average conversion (Mu et al., 2023), and offline counterfactual evaluation estimates how ranking or bidding policies would behave before deployment (Mu et al., 2025; Ye et al., 2026). Constrained budgeting integrates demand forecasts with marketing response, while privacy-safe marketing-mix modeling addresses identifier loss (Lu et al., 2025; Bai & Wu, 2026). E-commerce classification and recommendation studies also show that service value depends on downstream segmentation and choice, not only relevance scores (Xu et al., 2024). This literature supports presenting the WixQA financial analysis as a transparent scenario model rather than observed company ROI.

Human oversight depends on how evidence is displayed. Explainable e-commerce recommendation cards and financial-risk dashboards organize confidence, sources, and consequences for decision makers (B. Zhang et al., 2025; Z. Li et al., 2025). Text-grounded design-rationale interfaces and visual brief cards convert model inputs into reviewable design choices (Wu et al., 2025; Tu et al., 2025; Mu et al., 2026). Human-aligned design critique emphasizes that model feedback should be checked against expert judgment rather than treated as an authority (Y. Li et al., 2025). Evidence-linked counseling cards offer a further human-review pattern in which a recommendation remains subordinate to professional judgment (Zhang & Zhang, 2025b). For a support copilot, article IDs, extracted spans, confidence, and referral reasons should therefore be exposed as decision aids.

Interface governance also includes harms that retrieval metrics cannot detect. Dark-pattern analysis demonstrates how apparently ordinary microcopy can encode manipulative choices (H. Xu et al., 2025). Auditable essay feedback (Xin, 2026a) and student early-warning systems (Xin, 2026b) illustrate the need for rubric-grounded rationales when AI affects learning opportunities. Strategy-controlled emotional support dialogue similarly separates response intent from surface fluency (Zhang & Zhou, 2026). These domains differ from technical customer support, but they share a management requirement: users and reviewers need enough structured evidence to contest, correct, or escalate an automated recommendation.

3. Methodology

3.1 Data and Integrity Protocol

The knowledge base contained 6,221 records with unique article IDs. The evaluated question sets contained 200 ExpertWritten, 200 Simulated, and 6,221 Synthetic records. Every question had a nonempty question, reference answer, and list of relevant article IDs, and all 6,709 question-to-article links resolved to the corpus. ExpertWritten contained 258 links and 26.0% multi-document questions; Simulated contained 230 links and 13.5% multi-document questions. Each Synthetic question referenced one article, and the set covered the entire corpus (Table 1).

Figure 1 summarizes the evaluation architecture from corpus normalization through retrieval, fusion, reranking, confidence control, adaptive stopping, referral, and evidence-card output.

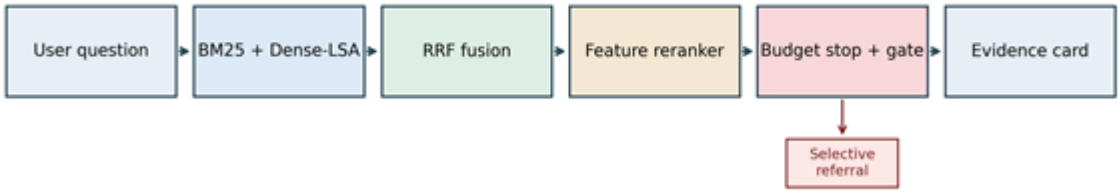

Figure 1. Evaluation architecture

Note. Agentic control denotes deterministic selection of evidence depth and referral. The answer layer is extractive.

Table 1. Question-set integrity and descriptive statistics

Dataset	Questions	Gold links	Unique gold	Multi-doc %	Median Q	Median A
ExpertWritten	200	258	209	26.0	18	139.5
Simulated	200	230	160	13.5	12	44
Synthetic	6,221	6,221	6,221	0.0	14	93

Note. Q and A are lowercase regular-expression word counts. Multi-doc is the percentage of questions with more than one annotated article ID.

The corpus included 4,109 general articles, 2,049 feature requests, and 63 known issues. Cleaned document length had a median of 205 analysis tokens, a mean of 383.07, a 95th percentile of 1,242, and a maximum of 6,722 (Table 2). Figure 2 compares corpus composition and median question and answer lengths. ExpertWritten answers were markedly longer than Simulated answers.

Table 2. Knowledge-base composition and cleaned text length

Measure	Value
Articles	6,221
General articles	4,109
Feature requests	2,049
Known issues	63
Median document tokens	205
Mean document tokens	383.07
P95 document tokens	1,242
Maximum document tokens	6,722

Note. Document length uses the fixed analysis tokenizer after text cleaning; it is not a vendor billing-token count.

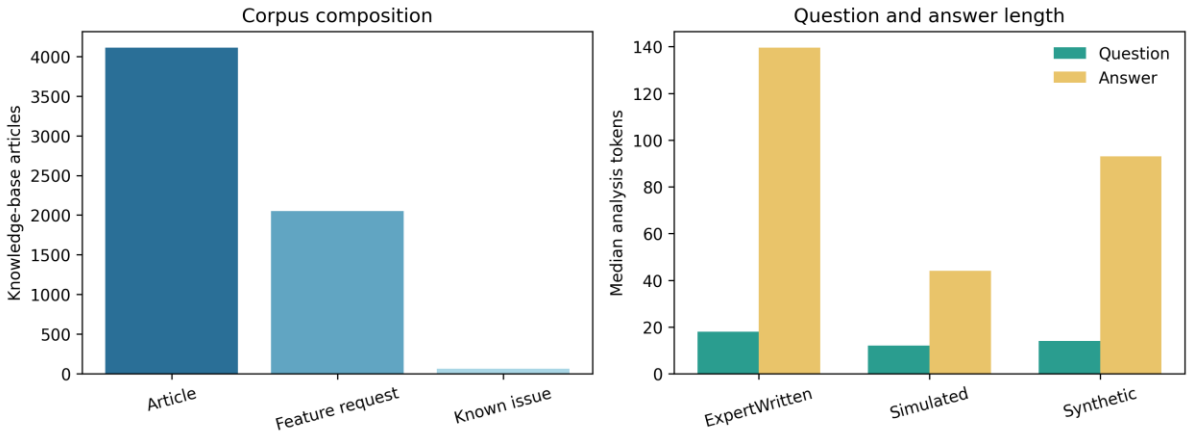

Figure 2. Dataset profile

Note. The length panel uses fixed regular-expression word counts; reference answers retain the visible text of Markdown links.

Text processing applied HTML unescaping, Unicode NFKC normalization, residual-tag and URL-string removal, whitespace normalization, lowercase regular-expression tokenization, and English stop-word removal. The contents field already began with the article title, so the separate title was not concatenated into retrieval text. Titles remained available for overlap features, while article IDs remained the evidence keys.

BM25, Dense-LSA, and RRF were unsupervised and were evaluated on every question. Learned components used five-fold grouped out-of-fold evaluation on Synthetic. StratifiedGroupKFold used article type for stratification and normalized question text for grouping, with seed 42. Fold sizes were 1,244, 1,244, 1,245, 1,244, and 1,244. For each fold, models were fitted on four folds and predicted the held-out fold. Final models were fitted on all eligible Synthetic candidate sets and applied once to ExpertWritten and Simulated; their labels did not influence features, parameters, or thresholds.

3.2 Retrieval, Reranking, and Evidence Assembly

BM25 used k1 = 1.2, b = 0.75, English stop words, and a 60,000-term cap. Query term frequencies were binary, while document weights followed Okapi BM25. Dense-LSA used word unigrams and bigrams, minimum document frequency 2, sublinear term frequency, a 50,000-feature cap, and L2-normalized TF–IDF. Randomized truncated SVD projected documents and questions into 256 dimensions with four iterations and seed 42; cosine similarity ranked normalized vectors.

$$RRF(d) = 1 / [60 + rBM25(d)] + 1 / [60 + rLSA(d)].$$

RRF fused the two top-100 lists. The reranker considered the union of the top-50 candidates from BM25, Dense-LSA, and RRF. Fourteen features captured normalized retrieval scores, reciprocal ranks, query-title Jaccard overlap, query coverage in the

document head, title phrase overlap, document and question length, and article type. Annotated articles were positive instances and up to 12 highest-ranked nonmatches were hard negatives. The standardized, class-balanced logistic model used C = 1, the liblinear solver, a 500-iteration limit, and seed 42. Final fitting used 80,782 pairs from 6,214 Synthetic questions whose annotated article occurred in the candidate pool; all 6,221 questions remained in evaluation.

The evidence-card assembler segmented documents into sentences, split spans longer than 80 tokens into 60-token units, and scored each unit using query coverage, query-conditioned precision, reciprocal document rank, title overlap, and a length preference. It chose no more than three sentences, no more than two from one article, rejected candidates with Jaccard overlap above 0.72 against a selected sentence, and enforced a 120-token limit. Each sentence retained its article ID. Table 3 records the fixed implementation.

Table 3. Components and fixed implementation parameters

Component	Fixed implementation
BM25	Okapi; k1=1.2; b=0.75; 60,000-term cap
Dense-LSA	TF-IDF 1-2 grams; min_df=2; 50,000 features; randomized SVD-256; cosine
Hybrid	RRF; k=60; BM25 and Dense-LSA top-100 union
Reranker	Logistic regression; 14 retrieval/overlap/type features; 12 hard negatives
Evidence card	Query-overlap sentence selection; max 3 sentences and 120 analysis tokens
Gate	Cross-fitted logistic confidence scoring; frozen 5% and 10% top-1-error thresholds

Note. Dense-LSA is a corpus-fitted statistical representation. The evidence card contains extracted corpus spans.

3.3 Confidence Gate and Budget Policy

The confidence scorer predicted whether the reranker's top-ranked article ID belonged to the annotated set. Ten features included the first reranker probability, first-to-second margin, agreement among retrievers, top-10 probability concentration, query-title overlap, first-stage scores, and question length. A standardized, class-balanced logistic model used C = 1, the lbfgs solver, and 500 iterations. Cross-fitted Synthetic predictions produced cumulative risk-coverage thresholds of 0.3472 for at most 5% top-1 article error and 0.1442 for at most 10% error.

The adaptive policy stopped after one document at the stricter threshold and assembled at most two sentences or 80 tokens. Confidence between the thresholds triggered a three-document, three-sentence, 120-token card. Confidence below 0.1442 triggered referral. The 10% target applies to the complete accepted cohort, not to the marginal three-document group. Fixed methods always processed their top three documents.

3.4 Metrics, Statistical Tests, and Scenario Economics

Retrieval metrics were macro Recall@1/3/5/10, Hit@k, Complete@k, MRR@10, and binary nDCG@10. Recall measured the fraction of annotated articles recovered; Hit required at least one; Complete required every annotated article. Answer F1 used lowercase word tokens, removed punctuation and the articles "a," "an," and "the," retained Markdown anchor text while removing destinations, and computed multiset overlap. Annotated citation precision was the fraction of cited article IDs in the annotated set, and citation recall was the fraction of annotated IDs cited. These article-level metrics do not evaluate inline placement, sentence entailment, or factual correctness.

The gate was evaluated with Brier score, ten equal-frequency-bin expected calibration error (ECE), AUROC, service coverage, and selective top-1 article error. Answer and citation means for the adaptive policy were conditional on answered cases and were always reported with coverage. Quality-adjusted yield measured the sum of answer F1 multiplied by annotated citation precision per million retrieved context analysis tokens.

$$QAY = 10^6 \, \Sigma i(F1i \times CPi) / \Sigma i \, Ti.$$

Planned paired comparisons tested Hybrid+Reranker against BM25 for Recall@5 and answer F1. Mean differences and 95% intervals came from 5,000 paired bootstrap resamples. Two-sided Wilcoxon signed-rank tests used the Pratt treatment of zero differences, and Holm adjustment covered all six dataset-endpoint tests. Latency was measured on the first 80 ExpertWritten questions after ten warm-up queries. The reranker row combines retrieval with logistic prediction but excludes candidate-feature

construction, confidence scoring, and evidence-card assembly, so the measurements are component benchmarks rather than end-to-end service latency.

The run used Python 3.12.13, NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0, scikit-learn 1.8.0, and Matplotlib 3.10.8 on Linux x86-64. Peak resident memory was 943.04 MB. Data hashes, dependency versions, fold sizes, thresholds, and query-level results were saved with the run manifest.

The dataset contains no labor cost, resolution time, revenue, satisfaction, or deployment-cost fields. Scenario economics therefore classified an accepted case by top-1 annotated article match, assigned value 6 to a match, cost 12 to a mismatch, cost 2 to a referral, and processing cost 0.50 per million measured analysis tokens. Sensitivity analysis varied match value from 2 to 10 and mismatch cost from 4 to 20. The ratio describes the decision assumptions and is not company financial performance.

$ROI = [6C - (12W + 2R + 0.5T / 10^6)] / [12W + 2R + 0.5T / 10^6]$.

4. Results and Discussion

4.1 Retrieval Performance

The three question sources differed in ways that shaped retrieval. ExpertWritten had the highest multi-document proportion and longest answers, Simulated questions were shorter, and Synthetic questions had a one-to-one relation with articles (Tables 1–2; Figure 2). The latter structure made Synthetic retrieval easier and its confidence distribution unlike the external sets.

Table 4. Retrieval performance on ExpertWritten

Method	R@1	R@3	R@5	R@10	Complete@5	MRR@10	nDCG@10
BM25	0.2075	0.3933	0.5050	0.6250	0.4450	0.3808	0.4218
Dense-LSA	0.1317	0.2842	0.3850	0.5233	0.3200	0.2851	0.3228
Hybrid-RRF	0.1817	0.4108	0.5200	0.6542	0.4650	0.3722	0.4248
Hybrid+Reranker	0.2492	0.4033	0.5058	0.6158	0.4400	0.4149	0.4431

Note. R is macro annotated-article recall. Complete@5 requires all annotated articles. Synthetic learned-component values are grouped five-fold out-of-fold results.

Table 5. Retrieval performance on Simulated

Method	R@1	R@3	R@5	R@10	Complete@5	MRR@10	nDCG@10
BM25	0.1900	0.3542	0.4358	0.5500	0.4050	0.3245	0.3693
Dense-LSA	0.1267	0.2808	0.3558	0.4392	0.3350	0.2367	0.2791
Hybrid-RRF	0.2075	0.3708	0.4692	0.5858	0.4450	0.3369	0.3889
Hybrid+Reranker	0.2275	0.3933	0.4617	0.5642	0.4300	0.3611	0.3990

Note. R is macro annotated-article recall. Complete@5 requires all annotated articles. Synthetic learned-component values are grouped five-fold out-of-fold results.

Table 6. Retrieval performance on Synthetic

Method	R@1	R@3	R@5	R@10	Complete@5	MRR@10	nDCG@10
BM25	0.7992	0.9318	0.9568	0.9780	0.9568	0.8680	0.8952
Dense-LSA	0.4940	0.6677	0.7308	0.8041	0.7308	0.5958	0.6460
Hybrid-RRF	0.6812	0.8447	0.8928	0.9425	0.8928	0.7724	0.8139
Hybrid+Reranker	0.8301	0.9328	0.9534	0.9746	0.9534	0.8846	0.9068

Note. R is macro annotated-article recall. Complete@5 requires all annotated articles. Synthetic learned-component values are grouped five-fold out-of-fold results.

On ExpertWritten, BM25 achieved Recall@5 of 0.5050, compared with 0.3850 for Dense-LSA, 0.5200 for Hybrid-RRF, and 0.5058 for Hybrid+Reranker (Table 4). The reranker nevertheless produced the best Recall@1 (0.2492), MRR@10 (0.4149), and nDCG@10 (0.4431), improving early ordering without improving total five-document recall. On Simulated, Recall@5 was 0.4358, 0.3558, 0.4692, and 0.4617, respectively (Table 5). The reranker again led Recall@1, MRR@10, and nDCG@10. Fusion broadened the evidence pool, while supervision concentrated useful articles nearer the top at a small cost to deeper recall.

Synthetic favored lexical matching (Table 6). BM25 reached Recall@5 of 0.9568; Dense-LSA reached 0.7308; Hybrid-RRF reached 0.8928; and the out-of-fold reranker reached 0.9534. Reranking improved Recall@1 from BM25's 0.7992 to 0.8301 and nDCG@10 from 0.8952 to 0.9068, but did not surpass BM25 at depth five. Figure 3 displays the complete recall curves and makes the difficulty gap visible.

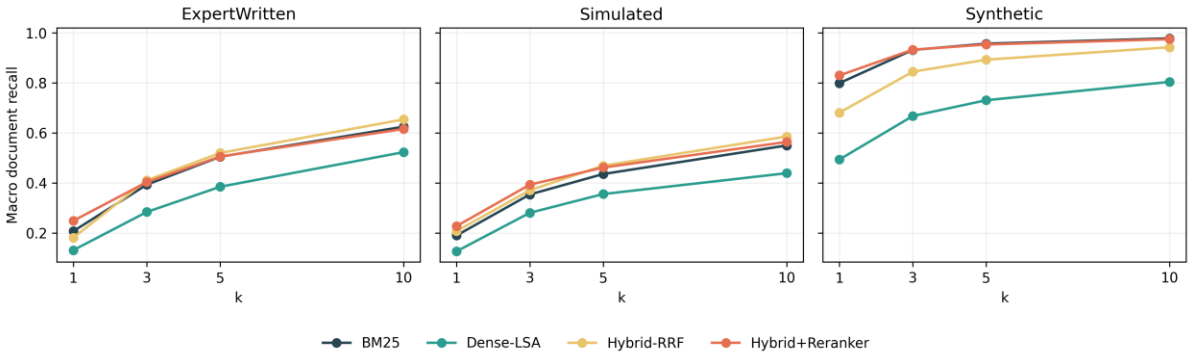

Figure 3. Recall across retrieval depths
Note. Higher is better. Synthetic reranker values are out-of-fold.

4.2 Answer and Article-Attribution Results

Table 7. External answer overlap, article attribution, and context cost

Dataset	Method	Coverage	F1	Cit. P	Cit. R	Median ctx.	P95 ctx.	QAY
ExpertWritten	BM25	1.000	0.3008	0.2233	0.3733	88	120	887.2
ExpertWritten	Dense-LSA	1.000	0.2930	0.1508	0.2442	80	120	684.6
ExpertWritten	Hybrid-RRF	1.000	0.2993	0.2017	0.3392	86	120	835.8
ExpertWritten	Hybrid+Reranker	1.000	0.3071	0.2283	0.3783	86	120	948.9
ExpertWritten	Agentic-RAG	0.565	0.3051	0.3835	0.4322	71	120	1,904.0
Simulated	BM25	1.000	0.2718	0.1675	0.3117	84	120	639.5
Simulated	Dense-LSA	1.000	0.2656	0.1342	0.2483	74.5	120	554.8
Simulated	Hybrid-RRF	1.000	0.2780	0.1708	0.3158	84	120	662.1
Simulated	Hybrid+Reranker	1.000	0.2729	0.1858	0.3383	79.5	120	712.0
Simulated	Agentic-RAG	0.485	0.2927	0.2766	0.2990	63	116.6	1,615.8

Note. Adaptive-policy F1, citation metrics, and context summaries are conditional on answered cases; coverage reports the retained share. QAY is quality-adjusted yield per million retrieved context analysis tokens.

Table 8. Synthetic answer overlap, article attribution, and context cost

Dataset	Method	Coverage	F1	Cit. P	Cit. R	Median ctx.	P95 ctx.	QAY
Synthetic	BM25	1.000	0.3926	0.4348	0.9113	90	120	1,957.2
Synthetic	Dense-LSA	1.000	0.3652	0.3103	0.6623	86	120	1,535.5
Synthetic	Hybrid-RRF	1.000	0.3852	0.3888	0.8301	91	120	1,783.9
Synthetic	Hybrid+Reranker	1.000	0.3949	0.4337	0.9166	90	120	1,969.2
Synthetic	Agentic-RAG	0.888	0.4829	0.8660	0.9338	56	103.8	7,380.8

Note. Synthetic learned-component values use out-of-fold rankings and confidence. Metrics for the adaptive policy are conditional on answered cases.

At full coverage on ExpertWritten, Hybrid+Reranker had the highest answer F1 (0.3071) and annotated citation precision (0.2283), compared with 0.3008 and 0.2233 for BM25 (Table 7). On Simulated, Hybrid-RRF had the highest full-coverage F1 at 0.2780, while reranking raised citation precision to 0.1858. These modest overlap scores combine retrieval errors with a strict comparison between short extractive cards and reference answers that often paraphrase or synthesize their sources.

The adaptive policy answered 56.5% of ExpertWritten cases. On those cases, F1 was 0.3051, annotated citation precision was 0.3835, citation recall was 0.4322, median context was 71 tokens, and QAY was 1,904.0. Simulated coverage was 48.5%; conditional F1 was 0.2927, citation precision 0.2766, citation recall 0.2990, median context 63 tokens, and QAY 1,615.8. Pooled external conditional F1 was 0.2994, but treating referrals as zero gives 0.1572. Selection therefore concentrated article matches among a smaller set of cases; it did not improve full-population response quality.

Synthetic produced much stronger adaptive results: 88.83% coverage, conditional F1 of 0.4829, citation precision of 0.8660, citation recall of 0.9338, median context of 56 tokens, and QAY of 7,380.8 (Table 8). Figure 4 summarizes the quality-context relationship using equal-weight means of the three dataset summaries. Its values consequently differ slightly from the query-pooled external values in Table 10.

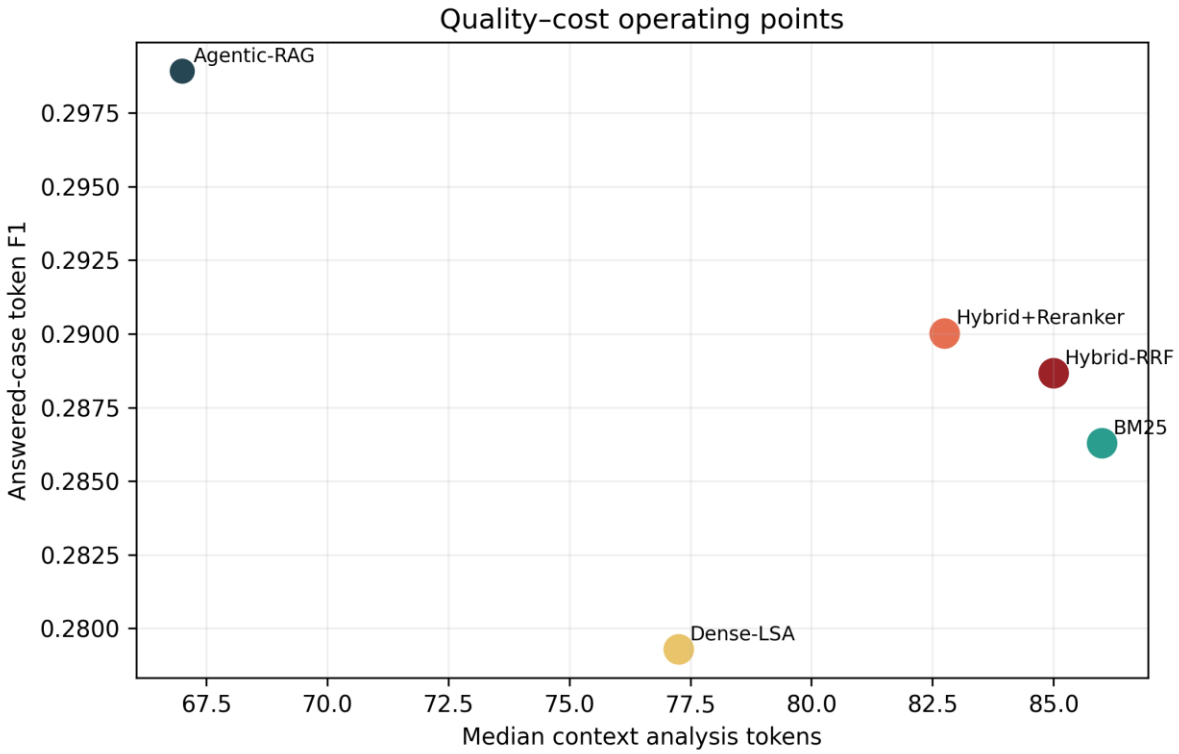

Figure 4. Answered-case quality and retrieved context

Note. Each plotted coordinate equally averages the three dataset-level summaries; Table 10 pools the two external sets by query.

4.3 Confidence Transfer and Selective Referral

Table 9. Confidence discrimination, calibration error, and selective top-1 article error

Dataset	Brier	ECE	AUROC	Coverage	Selective error
ExpertWritten	0.2282	0.1529	0.7064	0.5650	0.5664
Simulated	0.2065	0.1611	0.6743	0.4850	0.6907
Synthetic	0.1273	0.1529	0.9123	0.8883	0.0999

Note. The binary target is whether the reranker's top-ranked article ID belongs to the annotated set. Selective error is one minus top-1 match rate among accepted cases.

Table 9 and Figure 5 provide the central trust result. Synthetic cross-fitted AUROC was 0.9123, Brier score was 0.1273, and ECE was 0.1529. At the frozen 10% threshold, coverage was 0.8883 and selective top-1 article error was 0.0999. The threshold met its target within Synthetic out-of-fold evaluation.

External transfer failed. ExpertWritten AUROC fell to 0.7064, coverage to 0.5650, and selective error rose to 0.5664. Simulated AUROC was 0.6743, coverage was 0.4850, and error was 0.6907. The scorer retained some ordering ability, but its frozen probability scale did not preserve the Synthetic guarantee. Higher conditional citation precision must be read alongside this error: more than half of accepted ExpertWritten cases and more than two thirds of accepted Simulated cases lacked an annotated article at rank one.

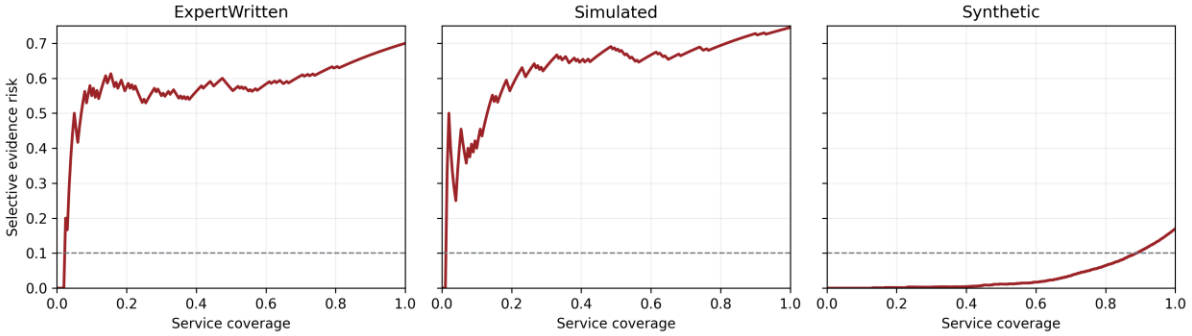

Figure 5. Risk-coverage curves under confidence shift

Note. Risk is selective top-1 annotated-article error, not hallucination, factuality, or response-level error.

For operations, the result is decisive: an acceptance threshold must be set with representative support traffic and monitored after knowledge-base or demand shifts. Article-derived questions were useful for initial discrimination, but their calibration was insufficient for autonomous answering.

The transfer failure is consistent with evidence from other operational domains: thresholds that work for one tenant or benchmark need not retain their meaning after workload or query shift (He et al., 2025). Approximately shared features may preserve some ranking signal while leaving absolute probabilities miscalibrated (Zhong, Pan, et al., 2025). The external WixQA results exhibit precisely that pattern: AUROC remained above chance, but the frozen acceptance threshold did not control selective error.

4.4 Ablation, Subgroups, and Statistical Evidence

Table 10. Pooled external component ablation

Method	Recall@5	Coverage	F1	Cit. P	Median ctx.
BM25	0.4704	1.000	0.2863	0.1954	84
Dense-LSA	0.3704	1.000	0.2793	0.1425	77
Hybrid-RRF	0.4946	1.000	0.2887	0.1862	85
Hybrid+Reranker	0.4838	1.000	0.2900	0.2071	83
Agentic-RAG	0.4838†	0.525	0.2994	0.3341	66.5

Note. External values pool ExpertWritten and Simulated by query. Adaptive-policy quality is conditional on answered cases. †The adaptive controller uses the Hybrid+Reranker base ranking; 0.4838 is the pre-gate Recall@5.

The pooled ablation clarifies endpoint trade-offs (Table 10). Recall@5 was 0.4704 for BM25, 0.3704 for Dense-LSA, 0.4946 for Hybrid-RRF, and 0.4838 after reranking. Dense-LSA was not a substitute for lexical matching in this corpus. RRF delivered the best external depth-five recall; reranking exchanged 0.0108 recall for better early ordering and article attribution.

Table 11. External retrieval by single- and multi-document subgroup

Dataset	Subgroup	Method	N	R@5	Complete@5	nDCG@10
ExpertWritten	Single	BM25	148	0.5270	0.5270	0.4113
ExpertWritten	Single	Dense-LSA	148	0.3986	0.3986	0.3017
ExpertWritten	Single	Hybrid-RRF	148	0.5270	0.5270	0.4178
ExpertWritten	Single	Hybrid+Reranker	148	0.5135	0.5135	0.4391
ExpertWritten	Multi	BM25	52	0.4423	0.2115	0.4519
ExpertWritten	Multi	Dense-LSA	52	0.3462	0.0962	0.3828
ExpertWritten	Multi	Hybrid-RRF	52	0.5000	0.2885	0.4449
ExpertWritten	Multi	Hybrid+Reranker	52	0.4840	0.2308	0.4544
Simulated	Single	BM25	173	0.4509	0.4509	0.3674
Simulated	Single	Dense-LSA	173	0.3642	0.3642	0.2840
Simulated	Single	Hybrid-RRF	173	0.4740	0.4740	0.3874
Simulated	Single	Hybrid+Reranker	173	0.4682	0.4682	0.3892
Simulated	Multi	BM25	27	0.3395	0.1111	0.3816
Simulated	Multi	Dense-LSA	27	0.3025	0.1481	0.2473
Simulated	Multi	Hybrid-RRF	27	0.4383	0.2593	0.3984
Simulated	Multi	Hybrid+Reranker	27	0.4198	0.1852	0.4619

Note. Complete@5 requires retrieval of every annotated article and is stricter than Recall@5 on multi-document questions.

Multi-document cases explain part of the fusion advantage (Table 11). For ExpertWritten multi-document questions, Hybrid-RRF raised Recall@5 from BM25's 0.4423 to 0.5000 and Complete@5 from 0.2115 to 0.2885. For Simulated multi-document questions, it raised Recall@5 from 0.3395 to 0.4383 and Complete@5 from 0.1111 to 0.2593. The reranker delivered the highest multi-document nDCG@10 on both sets but lower Complete@5 than the unfettered hybrid list. Fusion was therefore most useful when an answer depended on several articles.

Table 12. Paired Hybrid+Reranker versus BM25 comparisons

Dataset	Endpoint	Mean Δ	95% low	95% high	Raw p	Holm p
ExpertWritten	Recall@5	0.0008	-0.0484	0.0492	0.5668	1.0000
ExpertWritten	Answer F1	0.0063	-0.0018	0.0148	0.8209	1.0000
Simulated	Recall@5	0.0258	-0.0183	0.0692	0.1267	0.7600
Simulated	Answer F1	0.0012	-0.0066	0.0089	0.8631	1.0000
Synthetic	Recall@5	-0.0034	-0.0082	0.0014	0.1707	0.8536
Synthetic	Answer F1	0.0023	0.0007	0.0038	0.7079	1.0000

Note. Intervals are 5,000-resample paired bootstrap intervals for mean differences. p values are two-sided Pratt-Wilcoxon results; Holm adjustment covers all six tests.

The paired results did not establish a reliable reranking advantage (Table 12). Recall@5 differences were 0.0008 on ExpertWritten, 0.0258 on Simulated, and −0.0034 on Synthetic; all intervals included zero and all Holm-adjusted p values

exceeded 0.75. Answer-F1 differences were 0.0063, 0.0012, and 0.0023. The Synthetic bootstrap interval for the mean excluded zero, but the tie-sensitive Wilcoxon comparison was nonsignificant after correction. Because the procedures test different aspects of the paired distribution, the inferential conclusion remained nonsignificant.

4.5 Efficiency, Stopping, and Scenario Value

Table 13. Measured query-scoring latency and index construction

Measurement	Component	Median / seconds	P95 ms
Query scoring	BM25	9.97	20.77
Query scoring	Dense-LSA	29.17	57.54
Query scoring	Hybrid-RRF	38.33	67.77
Query scoring	Hybrid+Reranker	38.94	69.91
Index build	BM25 index	1.939	—
Index build	Dense-LSA index	17.540	—
Index build	Total indexes	19.479	—

Note. Query rows report milliseconds on 80 ExpertWritten questions after warm-up; build rows report elapsed seconds. The reranker row excludes candidate-feature construction, confidence scoring, and evidence-card assembly.

BM25 median scoring latency was 9.97 ms and its 95th percentile was 20.77 ms. Dense-LSA required 29.17 and 57.54 ms; Hybrid-RRF required 38.33 and 67.77 ms; and Hybrid+Reranker required 38.94 and 69.91 ms. BM25 indexing took 1.939 seconds, Dense-LSA indexing 17.540 seconds, and total construction 19.479 seconds (Table 13). These values characterize components, not networked service latency.

Table 14. Adaptive stopping and referral decisions

Dataset	Decision	Questions	Share
ExpertWritten	Stop@1	64	0.3200
ExpertWritten	Stop@3	49	0.2450
ExpertWritten	Refer	87	0.4350
Simulated	Stop@1	57	0.2850
Simulated	Stop@3	40	0.2000
Simulated	Refer	103	0.5150
Synthetic	Stop@1	4,688	0.7536
Synthetic	Stop@3	838	0.1347
Synthetic	Refer	695	0.1117

Note. Stop@1 uses the stricter threshold; Stop@3 is the intermediate cohort; Refer produces no evidence card.

Figure 6 and Table 14 show the workload created by the controller. Of ExpertWritten requests, 32.0% stopped after one document, 24.5% after three, and 43.5% were referred. Simulated shares were 28.5%, 20.0%, and 51.5%; Synthetic shares were 75.36%, 13.47%, and 11.17%. The high external referral rate reduced context use, but it also moved almost half the workload to another service path.

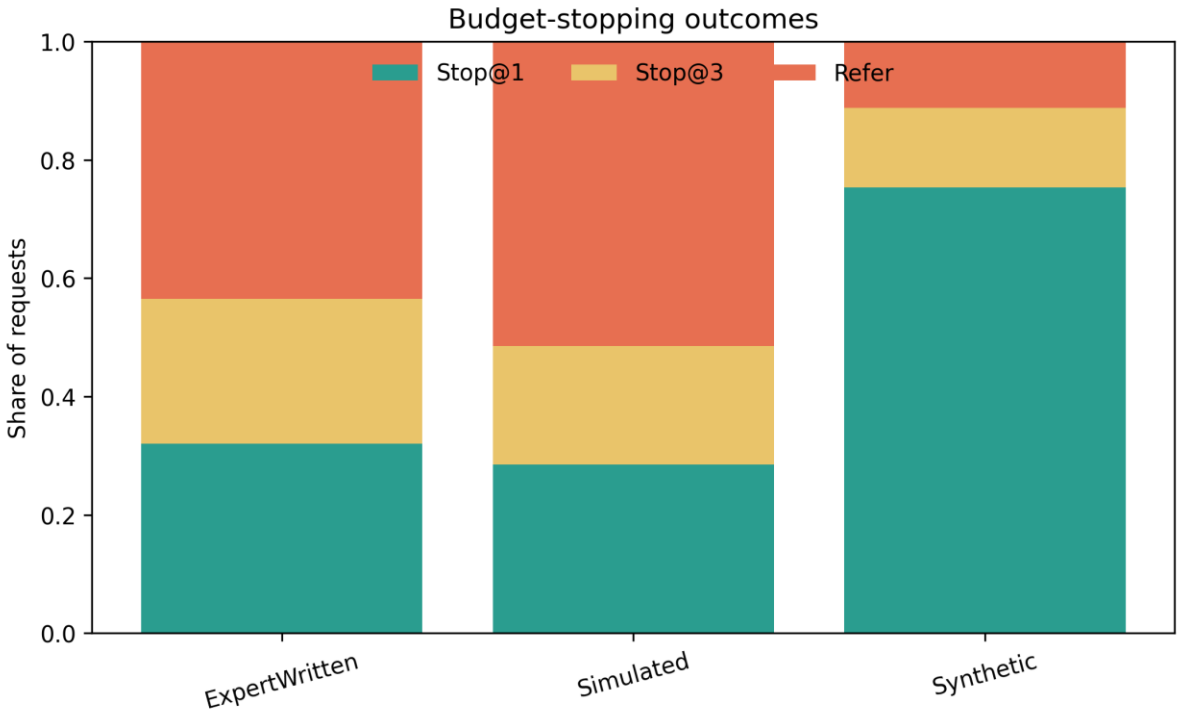

Figure 6. Budget-stopping decisions by question source
Note. The shares are derived from frozen confidence thresholds selected on Synthetic.

Table 15. Base-case top-1 article-match decision economics

Method	Matched	Mismatched	Referred	Processing	Net value	ROI
BM25	93	307	0	0.0370	-3126.04	-0.8485
Dense-LSA	62	338	0	0.0351	-3684.04	-0.9083
Hybrid-RRF	90	310	0	0.0367	-3180.04	-0.8548
Hybrid+Reranker	111	289	0	0.0370	-2802.04	-0.8080
Agentic-RAG	79	131	190	0.0174	-1478.02	-0.7572

Note. Matched and mismatched refer to top-1 article-ID agreement with annotations, not response correctness. Assumptions: match value 6, mismatch cost 12, referral cost 2, and processing cost 0.50 per million analysis tokens.

All base-case ratios were negative (Table 15). BM25 produced net value −3,126.04 and ROI −0.8485; Dense-LSA produced −3,684.04 and −0.9083; Hybrid-RRF produced −3,180.04 and −0.8548; and Hybrid+Reranker produced −2,802.04 and −0.8080. The adaptive policy accepted 79 annotated top-1 matches, accepted 131 mismatches, and referred 190 cases. Its net value was −1,478.02 and ROI −0.7572. Referral reduced the mismatch penalty enough to make the ratio less negative, but it did not approach profitability.

Every point in the complete sensitivity grid was also negative. Even the most favorable tested combination—match value 10 and mismatch cost 4—produced net value −114.02 and ROI −0.1261. Figure 7 visualizes this result. The chart's internal "correct response" label operationally means an accepted top-1 article match; it does not mean that an answer was human-verified as correct.

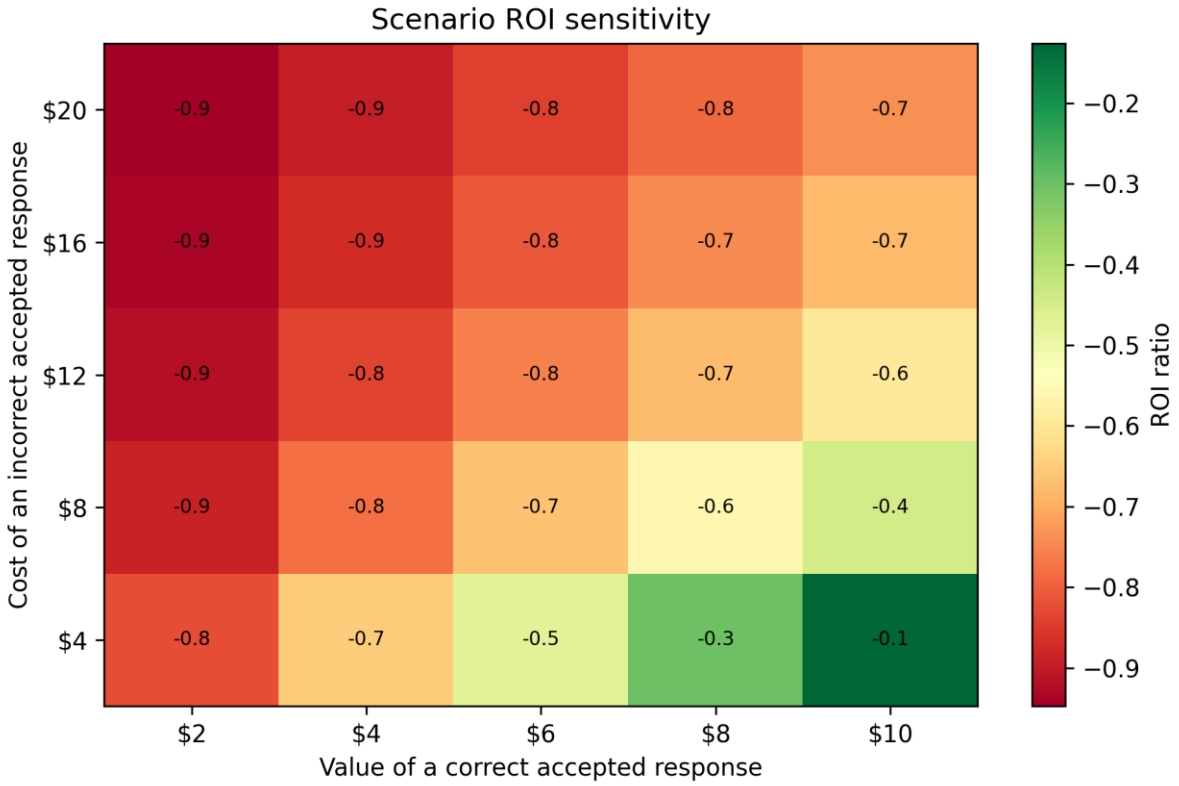

Figure 7. Scenario ROI sensitivity

Note. The figure's response terminology denotes article-ID match status. Referral cost is 2 and processing cost is 0.50 per million analysis tokens.

The negative sensitivity grid also clarifies why policy evaluation cannot stop at retrieval accuracy. Incrementality and off-policy research evaluates the value of a decision relative to an alternative action, with uncertainty about selection and counterfactual outcomes made explicit (Mu et al., 2023, 2025; Ye et al., 2026). WixQA does not contain those business outcomes, so the scenario results are best interpreted as a decision boundary: under every tested set of transparent assumptions, autonomous answering was not economically justified.

4.6 Limitations and Management Implications

The experiment used one English-language company knowledge base at one snapshot. Article-level annotations identify intended evidence but may not exhaust every useful document. The answer assembler measured extractive evidence selection rather than conversational fluency, synthesis, or policy compliance. Confidence and ROI both used top-1 article membership as their binary outcome, so neither supports claims about response factuality. External questions introduced unseen language and task construction, but not unseen documents: Synthetic covered every corpus article.

The principal limitation was distribution shift. Synthetic supported strong out-of-fold discrimination, yet its threshold did not control external top-1 error. The management implication is specific: abundant synthetic questions can initialize retrieval and confidence models, but acceptance thresholds require representative human requests. A defensible service also logs article versions, retrieved rankings, confidence, attributions, stopping decisions, referrals, and downstream outcomes as part of the lifecycle controls emphasized in operational RAG research (J. He et al., 2026; X. Xu et al., 2025).

A production interface should preserve the same separation. Evidence cards and risk dashboards can present source hierarchy, confidence, and escalation status without converting a model score into a guarantee (Jin, 2025b; Z. Li et al., 2025). Security-oriented risk cards add prompt-injection and data-integrity warnings (Su, Rao, & Ma, 2026), while incident visualization work shows how raw operational evidence can remain available for SRE review (Nie et al., 2026). These patterns support human review of both the answer and the controller's decision path.

5. Conclusions

Method value depended on the endpoint. BM25 was the strongest single retriever, Dense-LSA underperformed it, and RRF achieved the highest pooled external Recall@5 at 0.4946. The lightweight reranker improved early ordering and article attribution but did not significantly improve Recall@5 or answer F1. Adaptive stopping reduced retrieved context and concentrated annotated citations among answered cases, with F1 of 0.3051 at 56.5% coverage on ExpertWritten and 0.2927 at 48.5% coverage on Simulated.

The frozen confidence threshold's failure to preserve selective error was the primary result. Top-1 error of 0.0999 on Synthetic became 0.5664 and 0.6907 on the external sets. Confidence learned from article-derived questions therefore did not justify autonomous answering of real support requests. All tested ROI scenarios were also negative, although explicit referral produced the least negative base-case ratio.

A defensible enterprise service copilot combines hybrid retrieval, evidence-linked output, conservative referral, and continuing calibration on representative support traffic. Benchmark accuracy, context efficiency, and synthetic confidence are useful diagnostics; organizational trust and net benefit require external calibration, human escalation, and measured operational outcomes.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers
Artificial Intelligence (AI) Use Disclosure: The authors declare that no artificial intelligence tools were used in the preparation of this manuscript.

References

[1] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=hSyW5go0v8

[2] Bai, J., & Wu, Q. (2026). Privacy-safe marketing mix modeling and budget optimization under identifier loss: A controlled simulation study. *International Journal of Electronics and Communication Systems, 6*(1). https://doi.org/10.24042/ijecs.v6i1.30533

[3] Bai, J., Chen, S., Zheng, D., & Kuo, M.-J. (2026). Interpretable attack-chain stage detection from AWS CloudTrail event sequences via linear models and HMM smoothing. *Information, Electrical and Electronics Engineering, 6*(1), 28-43. https://doi.org/10.33474/infotron.v6i1.24923

[4] Bai, J., Wang, H., Wu, Q., & Zhang, B. (2026). Privacy-robust incrementality estimation in cookieless settings via uplift modeling: Reproducible evidence from the Hillstrom e-mail experiment. *Journal of Technology Informatics and Engineering, 5*(1), 17-38. https://doi.org/10.51903/jtie.v5i1.468

[5] Brynjolfsson, E., Li, D., & Raymond, L. R. (2025). Generative AI at work. *The Quarterly Journal of Economics, 140*(2), 889–942. https://doi.org/10.1093/qje/qjae044

[6] Chang, X., Lu, Y., & Zhong, Z. S. (2026). Review-grounded explainable recommendation with faithfulness evaluation on Amazon Reviews. *JEECS (Journal of Electrical Engineering and Computer Sciences), 11*(1), 9-22. https://doi.org/10.54732/jeecs.v11i1.2

[7] Chen, L., Zaharia, M., & Zou, J. (2024). FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=cSimKw5p6R

[8] Chen, S., He, S., & Sun, E. (2024). Risk-bounded GPU resource oversubscription via conformal demand envelopes in production AI clusters. *Journal of Advanced Computing Systems, 4*(5), 119-134. https://doi.org/10.69987/JACS.2024.40509

[9] Chen, Y., & Xu, H. (2026). Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access. *International Journal of Graphic Design, 4*(1), 141-164. https://doi.org/10.51903/ijgd.v4i1.3552

[10] Chen, Y., Zhang, Y., & Sherman, M. (2024). Going concern and bankruptcy prediction under extreme class imbalance: Cost-sensitive learning, resampling, and focal loss with explainable financial-ratio portraits. *Journal of Advanced Computing Systems, 4*(4), 80-96. https://doi.org/10.69987/JACS.2024.40407

[11] Chen, Y., Zhang, Y., Chau, D., & Sherman, M. (2023). Credit card default risk tiering with probability calibration and uncertainty-driven rejection: A reproducible study on the UCI credit card clients dataset. *Journal of Advanced Computing Systems, 3*(4), 31-47. https://doi.org/10.69987/JACS.2023.30403

[12] Cohen, D., Burg, L., Pykhnivskyi, S., Gur, H., Kovynov, S., Atzmon, O., & Barkan, G. (2025). WixQA: A multi-dataset benchmark for enterprise retrieval-augmented generation. *arXiv*. https://doi.org/10.48550/arXiv.2505.08643

[13] Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 758–759). ACM. https://doi.org/10.1145/1571941.1572114

[14] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9

[15] DeLone, W. H., & McLean, E. R. (2003). The DeLone and McLean model of information systems success: A ten-year update. *Journal of Management Information Systems, 19*(4), 9–30. https://doi.org/10.1080/07421222.2003.11045748

[16] Ding, D., Mallick, A., Wang, C., Sim, R., Mukherjee, S., Rühle, V., Lakshmanan, L. V. S., & Awadallah, A. H. (2024). Hybrid LLM: Cost-efficient and quality-aware query routing. *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=02f3mUtqnM

[17] Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 150–158). ACL. https://doi.org/10.18653/v1/2024.eacl-demo.16

[18] Feng, S., Patel, S. S., Wan, H., & Joshi, S. (2021). MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 6162–6176). ACL. https://doi.org/10.18653/v1/2021.emnlp-main.498

[19] Feng, S., Wan, H., Gunasekara, C., Patel, S. S., Joshi, S., & Lastras, L. (2020). Doc2Dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 8118–8128). ACL. https://doi.org/10.18653/v1/2020.emnlp-main.652

[20] Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 6465–6488). ACL. https://doi.org/10.18653/v1/2023.emnlp-main.398

[21] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. *Advances in Neural Information Processing Systems, 30*. https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html

[22] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

[23] He, J., Chen, P., Wu, C., Liang, S., Li, Y., Tan, G., Wen, X., & Zhang, C. (2026). An end-to-end framework for building large language models for software operations. *arXiv*. https://doi.org/10.48550/arXiv.2605.02906

[24] He, S., Li, C., & Rao, H. (2025). Few-shot cold-start workload forecasting for new AI inference tenants with time-series foundation models. *Journal of Technology Informatics and Engineering, 4*(1), 306-324. https://doi.org/10.51903/jtie.v4i1.546

[25] He, S., Nie, J., & Li, C. (2026). Power-aware inventory planning for AI infrastructure using job-level forecasting and LLM workload explanations. *Journal of Technology Informatics and Engineering, 5*(1), 341-359. https://doi.org/10.51903/jtie.v5i1.548

[26] He, S., Tu, H., & Liu, I. (2023). Safe PD capacity forecasting with time-series foundation models and calibrated uncertainty for heterogeneous GPU clusters. *Journal of Advanced Computing Systems, 3*(4), 48-66. https://doi.org/10.69987/JACS.2023.30404

[27] Jeong, S., Baek, J., Cho, S., Hwang, S. J., & Park, J. (2024). Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 7036–7050). ACL. https://doi.org/10.18653/v1/2024.naacl-long.389

[28] Jin, J. (2025a). Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations. *Journal of Technology Informatics and Engineering, 4*(2), 520-533. https://doi.org/10.51903/jtie.v4i2.535

[29] Jin, J. (2025b). LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy. *International Journal of Graphic Design, 3*(2), 397-414. https://doi.org/10.51903/ijgd.v3i2.3698

[30] Jin, J., Huang, T., & Lu, S. (2024). A model-risk-friendly probability of default workflow: Calibration, distribution-free uncertainty quantification, and SHAP explanations on the UCI credit card default dataset. *Journal of Advanced Computing Systems, 4*(6), 74-85. https://doi.org/10.69987/JACS.2024.40606

[31] Khan, A. F., Khan, A. A., Mohamed, A., Ali, H., Moolinti, S., Haroon, S., Tahir, U., Fazzini, M., Butt, A. R., & Anwar, A. (2025). LADs: Leveraging large language models for AI-driven DevOps. *arXiv*. https://doi.org/10.48550/arXiv.2502.20825

[32] Kuo, M.-J., Zheng, D., & Hires, J. (2025). Federated topic-preference learning for knowledge-grounded chat with differential privacy. *Journal of Technology Informatics and Engineering, 4*(2), 385-401. https://doi.org/10.51903/jtie.v4i2.502

[33] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[34] Li, C., Bai, J., & Wang, S. (2024). Evidence-chain reliable RAG: Word-level hallucination detection, source attribution, and provenance explanation for LLM applications. *Journal of Advanced Computing Systems, 4*(2), 76-92. https://doi.org/10.69987/JACS.2024.40207

[35] Li, C., Liu, G., & Zhao, Z. (2026). Cost-aware LLM-style routing for AIOps log analysis: Log parsing, anomaly detection, fault diagnosis, and incident summarization on LogEval task files. *Journal of Technology Informatics and Engineering, 5*(2), 91-103. https://doi.org/10.51903/jtie.v5i2.538

[36] Li, J., & Zhou, A. (2026). Multi-regulation RAG for AI product counsel: A legal governance framework for cross-border digital commerce. *Rule of Law Studies Journal, 2*(2), 105-123. https://doi.org/10.64780/rolsj.v2i2.225

[37] Li, Y. (2024). Findable then explainable: Retrieval-summary integration for code intelligence on a lightweight CodeSearchNet subset. *Journal of Advanced Computing Systems, 4*(7), 65-82. https://doi.org/10.69987/JACS.2024.40706

[38] Li, Y., & Lu, S. (2025). Language-guided feature selection for DDoS and intrusion detection on CICIDS2017. *Journal of Technology Informatics and Engineering, 4*(1), 284-305. https://doi.org/10.51903/jtie.v4i1.531

[39] Li, Y., Lu, S., & Zhao, L. (2025). LLM-as-design-critic: Aligning AI-generated UI feedback with human graphic design judgment. *International Journal of Graphic Design, 3*(1), 196-215. https://doi.org/10.51903/ijgd.v3i1.3661

[40] Li, Z., Zhang, K., & Wong, A. (2026). Numerical-reasoning guardrails for a quant research assistant: A compact reproducible benchmark using SEC and FRED data. *Journal of Technology Informatics and Engineering, 5*(2), 75-90. https://doi.org/10.51903/jtie.v5i2.541

[41] Li, Z., Zhou, S., & Zhou, Z. (2025). Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench. *International Journal of Graphic Design, 3*(1), 196-210. https://doi.org/10.51903/ijgd.v3i1.3715

[42] Liu, G., He, S., & Liu, I. (2023). LLM-augmented multi-source root cause attribution for CPU and network faults in microservices. *Journal of Advanced Computing Systems, 3*(6), 39-57. https://doi.org/10.69987/JACS.2023.30604

[43] Liu, G., He, S., & Wong, H. (2025). LLM-compatible visual brief cards for AI infrastructure capacity dashboards: A UI/UX framework for turning forecast risk into graphic design decisions. *International Journal of Graphic Design, 3*(1), 196-213. https://doi.org/10.51903/ijgd.v3i1.3723

[44] Lu, S., & Zhou, D. (2024). TinyLLM-assisted intrusion detection for real-time IoT networks. *Journal of Advanced Computing Systems, 4*(8), 72-87. https://doi.org/10.69987/JACS.2024.40809

[45] Lu, Y., Zhou, H., & Zhang, Y. (2025). A constrained, data-driven budgeting framework integrating macro demand forecasting and marketing response modeling. *Journal of Technology Informatics and Engineering, 4*(3). https://doi.org/10.51903/jtie.v4i3.466

[46] Meng, S., Chen, J., & Zheng, I. (2026). LLM-inspired offline reranking for financial search: Query rewriting, hybrid retrieval, and listwise relevance ranking on FiQA. *Journal of Technology Informatics and Engineering, 5*(1), 361-378. https://doi.org/10.51903/jtie.v5i1.537

[47] Mu, J., Lu, Y., & Hwang, E. (2026). Structured visual brief interfaces for advertising design: A UI/UX framework for turning creative intentions into designer-editable graphic design cards. *International Journal of Graphic Design, 4*(1), 192-208. https://doi.org/10.51903/ijgd.v4i1.3702

[48] Mu, J., Lu, Y., & Smith, M. (2023). LLM-assisted incrementality (uplift) modeling for email advertising: From feature interactions to interpretable audience-creative-channel policies. *Journal of Advanced Computing Systems, 3*(1), 31-48. https://doi.org/10.69987/JACS.2023.30103

[49] Mu, J., Ye, T., & Patel, P. (2025). Offline counterfactual evaluation for advertising and recommendation slot policies: A reproducible study on the Open Bandit Dataset (Small). *Journal of Technology Informatics and Engineering, 4*(3), 521-543. https://doi.org/10.51903/jtie.v4i3.500

[50] Nie, J., & Zheng, D. (2024). Noisy-neighbor-aware VM degradation risk modeling with unsupervised residual fusion. *Journal of Advanced Computing Systems, 4*(4), 112-123. https://doi.org/10.69987/JACS.2024.40409

[51] Nie, J., Liu, G., Li, C., & Zou, T. (2026). Evidence-constrained incident visualization cards for distributed cloud logs: A UI/UX framework for turning Hadoop, OpenStack, and ZooKeeper logs into actionable SRE design interfaces. *International Journal of Graphic Design, 4*(1), 179-185. https://doi.org/10.51903/ijgd.v4i1.3703

[52] Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. *arXiv*. https://doi.org/10.48550/arXiv.1901.04085

[53] Ong, I., Almahairi, A., Wu, V., Chiang, W.-L., Wu, T., Gonzalez, J. E., Kadous, M. W., & Stoica, I. (2025). RouteLLM: Learning to route LLMs from preference data. *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=8sSqNntaMr

[54] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2383–2392). ACL. https://doi.org/10.18653/v1/D16-1264

[55] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval, 3*(4), 333–389. https://doi.org/10.1561/1500000019

[56] Ru, D., Qiu, L., Hu, X., Zhang, T., Shi, P., Chang, S., Jiayang, C., Wang, C., Sun, S., Li, H., Zhang, Z., Wang, B., Jiang, J., He, T., Wang, Z., Liu, P., Zhang, Y., & Zhang, Z. (2024). RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems, 37*, 21999–22027. https://doi.org/10.52202/079017-0692

[57] Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024). ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 338–354). ACL. https://doi.org/10.18653/v1/2024.naacl-long.20

[58] Su, W., Chen, S., & Qian, E. (2026). Narrative-aware scientific claim verification agent with evidence ranking for ClimateCheck. *Journal of Technology Informatics and Engineering, 5*(1), 327-340. https://doi.org/10.51903/jtie.v5i1.549

[59] Su, W., Chen, S., & Zhao, C. (2025). Budgeted multi-hop retrieval agent for compositional question answering: A retrieval-policy evaluation on the official MultiHop-RAG benchmark. *Journal of Technology Informatics and Engineering, 4*(3), 649-662. https://doi.org/10.51903/jtie.v4i3.543

[60] Su, W., Rao, H., & Ma, E. (2026). Privacy and data-integrity risk cards for LLM agents: A UI/UX design framework for secure human oversight under prompt-injection attacks. *International Journal of Graphic Design, 4*(1), 186-191. https://doi.org/10.51903/ijgd.v4i1.3699

[61] Sun, X., Lu, Y., & Chen, J. (2023). Controllable long-term user memory for multi-session dialogue: Confidence-gated writing, time-aware retrieval-augmented generation, and update/forgetting. *Journal of Advanced Computing Systems, 3*(8), 9-24. https://doi.org/10.69987/JACS.2023.30802

[62] Sun, X., Zhong, Z. S., & Wu, Q. (2026). Retrieval-grounded HDFS log anomaly detection and deterministic failure narrative generation. *Journal of Computer Systems and Applications, 3*(1), 15-30. https://doi.org/10.64229/j6d7fr94

[63] Tang, Y., & Yang, Y. (2024). MultiHop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv*. https://doi.org/10.48550/arXiv.2401.15391

[64] Trivedi, H., Balasubramanian, N., Khot, T., & Sabharwal, A. (2023). Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (pp. 10014–10037). ACL. https://doi.org/10.18653/v1/2023.acl-long.557

[65] Tu, H., Zhao, S., & Zhou, A. (2025). Visual brief cards for advertising design: A structured UI/UX framework for turning creative intentions into graphic design decisions. *International Journal of Graphic Design, 3*(1), 210-226. https://doi.org/10.51903/ijgd.v3i1.3714

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

[67] Wang, C., Wen, Z., Zhang, R., Xu, P., & Jiang, Y. (2025). GPU memory requirement prediction for deep learning tasks based on bidirectional gated recurrent unit optimization transformer. In *2025 5th International Conference on Artificial Intelligence, Virtual Reality and Visualization (AIVRV)*. IEEE. https://doi.org/10.1109/AIVRV67401.2025.11350369

[68] Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., & Wei, F. (2022). Text embeddings by weakly-supervised contrastive pre-training. *arXiv*. https://doi.org/10.48550/arXiv.2212.03533

[69] Wix.com. (2025). *WixQA* [Data set]. Hugging Face. https://huggingface.co/datasets/Wix/WixQA

[70] Wu, Q., Meng, S., & Zhao, J. (2025). Text-grounded LLM-assisted design rationale interfaces: Turning advertising layout metadata into explainable UI/UX decision cards. *International Journal of Graphic Design, 3*(1), 216-240. https://doi.org/10.51903/ijgd.v3i1.3713

[71] Xin, Q. (2025a). Explaining OpenStack failure-injection log anomalies with retrieved normal prototypes. *Emerging Information Science and Technology, 6*(2), 125-146. https://doi.org/10.18196/eist.v6i2.31232

[72] Xin, Q. (2025b). Hybrid cloud architecture for efficient and cost-effective large language model deployment. *Journal of Information Systems and Informatics, 7*(3), 2182-2195. https://doi.org/10.51519/journalisi.v7i3.1170

[73] Xin, Q. (2025c). Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning). *Journal of Technology Informatics and Engineering, 4*(1), 215-238. https://doi.org/10.51903/jtie.v4i1.485

[74] Xin, Q. (2026a). Auditable automated essay scoring and formative feedback: A rubric-grounded pipeline for secondary and higher education. *Journal of Applied Artificial Intelligence in Education, 2*(1), 1-19. https://doi.org/10.66053/jaaie.v2i1.348

[75] Xin, Q. (2026b). Early-warning analytics with LLM intervention rationales for student retention decisions: Classroom interaction modeling with xAPI-Edu-Data and dropout/success prediction. *Interdisciplinary Journal of Pedagogical Research and Media Technology, 2*(1). https://doi.org/10.64268/inspire.v2i1.117

[76] Xin, Q. (2026c). Explainable and fair credit risk scoring with counterfactual explanations: A reproducible evaluation on the German Credit dataset (HELOC-motivated). *Journal of Information and Technology, 14*(2), 215-231. https://doi.org/10.32664/j-intech.v14i02.2228

[77] Xin, Q. (2026d). Host-based intrusion detection with system call sequences: Window localization and forensic narratives. *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls, 8*(2). https://doi.org/10.28989/avitec.v8i2.3973

[78] Xin, Q. (2026e). Log anomaly detection with conformal alert control and evidence-grounded incident ticket generation. *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls, 8*(2), 247-264. https://doi.org/10.28989/avitec.v8i2.3974

[79] Xin, Q. (2026f). Self-supervised customer representation learning for segmentation and next-purchase prediction on UCI Online Retail. *Journal of Information and Technology, 14*(1), 20-37. https://doi.org/10.32664/j-intech.v14i01.2229

[80] Xin, Q. (2026g). Self-supervised log anomaly detection with LogBERT-style transformers: Full empirical evaluation on a reproducible SynHDFS benchmark. *JEECS (Journal of Electrical Engineering and Computer Sciences), 11*(1), 23-35. https://doi.org/10.54732/jeecs.v11i1.3

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

[82] Xu, H., Chen, Y., & Med, A. (2025). Automatic detection and explanation of dark patterns from interface microcopy: Empirical comparison of BERT-style encoders, RoBERTa-style encoders, and LLM-style decoders on the ec-darkpattern dataset. *Journal of Technology Informatics and Engineering, 4*(3), 590-612. https://doi.org/10.51903/jtie.v4i3.491

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

[84] Xu, X., Weytjens, H., Zhang, D., Lu, Q., Weber, I., & Zhu, L. (2025). RAGOps: An enterprise framework for LLM lifecycle management. *arXiv*. https://doi.org/10.48550/arXiv.2506.03401

[85] Yao, S., Shinn, N., Razavi, P., & Narasimhan, K. (2024). τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv*. https://doi.org/10.48550/arXiv.2406.12045

[86] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=WE_vluYUL-X

[87] Ye, T., Mu, J., & Hunter, J. (2026). Off-policy evaluation and conservative policy selection for slot-level dynamic bidding and ranking on the Open Bandit Dataset (Small). *Journal of Technology Informatics and Engineering, 5*(1), 178-199. https://doi.org/10.51903/jtie.v5i1.503

[88] Zhang, B., Rao, H., & Zhao, D. (2024). Evidence-grounded RAG for cloud-native DevOps: Hallucination-resistant AIOps question answering over private operations documents. *Journal of Advanced Computing Systems, 4*(3), 109-125. https://doi.org/10.69987/JACS.2024.40308

[89] Zhang, B., Ren, Y., & Zou, J. (2025). LLM-style explainable e-commerce recommendation cards: A UI/UX design framework for trust-calibrated product recommendation. *International Journal of Graphic Design, 3*(2), 381-396. https://doi.org/10.51903/ijgd.v3i2.3697

[90] Zhang, B., Sun, X., Liu, G., & Zhou, B. (2026). LLM-style DevOps copilot for cloud-native troubleshooting: Retrieval-augmented runbook generation and command-safety evaluation. *Journal of Technology Informatics and Engineering, 5*(2), 104-118. https://doi.org/10.51903/jtie.v5i2.534

[91] Zhang, K., Chen, Y., & Qian, A. (2025). Evidence-grounded accounting disclosure review cards: A visual communication framework for LLM-style explanations over SEC financial statements and notes. *International Journal of Graphic Design, 3*(2), 395. https://doi.org/10.51903/ijgd.v3i2.3710

[92] Zhang, R., Wen, Z., Wang, C., Tang, C., Xu, P., & Jiang, Y. (2025). Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms. *arXiv*. https://doi.org/10.48550/arXiv.2511.19481

[93] Zhang, Y., & Zhang, H. (2025a). A therapist-facing session copilot for live counseling support: Reasoning-guided retrieval and ranking from multi-turn counseling dialogues. *Journal of Technology Informatics and Engineering, 4*(2), 464-486. https://doi.org/10.51903/jtie.v4i2.547

[94] Zhang, Y., & Zhang, H. (2025b). Visualizing the right counseling support: Evidence-linked recommendation cards for explainable mental health intake interfaces. *International Journal of Graphic Design, 3*(1), 214-229. https://doi.org/10.51903/ijgd.v3i1.3722

[95] Zhang, Y., & Zhou, Z. (2026). Strategy-aware therapist imitation for emotional support dialogues: A reproducible ESConv study for LLM response control. *Advances in Educational Technology and Psychology, 10*(2), 92-97. https://doi.org/10.23977/aetp.2026.100213

[96] Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 295–305). ACM. https://doi.org/10.1145/3351095.3372852

[97] Zhao, S., Bai, J., & Roberson, D. (2025). Multi-horizon GPU demand forecasting with workload semantics and operational risk curves: An empirical study on Alibaba Clusterdata GPU trace. *Journal of Technology Informatics and Engineering, 4*(3), 544-571. https://doi.org/10.51903/jtie.v4i3.498

[98] Zhao, S., Ren, Y., & Chang, X. (2026). Profit-aware spot GPU admission control with cost-sensitive loss and evidence-grounded policy memos for AI workload supply-demand matching. *Journal of Technology Informatics and Engineering, 5*(2), 45-59. https://doi.org/10.51903/jtie.v5i2.545

[99] Zheng, D., & Li, C. (2024). Behavior-level jailbreak resistance via multi-stage refusal and utility preservation. *Journal of Advanced Computing Systems, 4*(1), 83-99. https://doi.org/10.69987/JACS.2024.40107

[100] Zheng, D., Li, C., & Davidson, H. (2023). Continual red-teaming for in-the-wild jailbreaks via online guardrail updates and guardrail distillation. *Journal of Advanced Computing Systems, 3*(2), 35-49. https://doi.org/10.69987/JACS.2023.30203

[101] Zheng, D., Zhang, B., & Geibel, J. (2024). VerifySafe: Toxicity-safe agent responses under adversarial prompts with evidence-based self-verification. *Journal of Advanced Computing Systems, 4*(1), 67-82. https://doi.org/10.69987/JACS.2024.40106

[102] Zhong, Z. S., & Ling, S. (2024). Improved theoretical guarantee for rank aggregation via spectral method. *Information and Inference: A Journal of the IMA, 13*(3), iaae020. https://doi.org/10.1093/imaiai/iaae020

[103] Zhong, Z. S., Chen, J., Zhong, E., & Sun, X. (2025). Evidence-calibrated RAG for unanswerable question answering: Retrieval coverage, abstention calibration, and hallucination-proxy analysis on SQuAD 2.0. *Journal of Technology Informatics and Engineering, 4*(2), 502-520. https://doi.org/10.51903/jtie.v4i2.536

[104] Zhong, Z. S., Li, C., & Rao, H. (2026). Trajectory reliability prediction for generalist AI agents: Tool-use failure analysis and success forecasting on ZClawBench. *Journal of Technology Informatics and Engineering, 5*(1), 341-360. https://doi.org/10.51903/jtie.v5i1.539

[105] Zhong, Z. S., Pan, X., & Lei, Q. (2025). Bridging domains with approximately shared features. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics* (Vol. 258, pp. 559-567). PMLR. https://proceedings.mlr.press/v258/zhong25a.html

[106] Zhou, B., Jin, J., & Zhao, D. (2025). Calibrated resume-job matching for trustworthy LLM-assisted recruiter screening: Pairwise matching, probability calibration, and selective refusal on two public recruitment datasets. *Journal of Technology Informatics and Engineering, 4*(3), 625-648. https://doi.org/10.51903/jtie.v4i3.529

[107] Zhou, B., Wang, H., & Chang, X. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX design framework for rubric-based medical risk communication. *International Journal of Graphic Design, 3*(2), 365-380. https://doi.org/10.51903/ijgd.v3i2.3696

[108] Zhou, S., Chen, Y., & Lee, K. (2026). Accounting-aware evidence-constrained agents for disclosure, settlement, and secondary-market risk monitoring in tokenized RWA infrastructure. *Journal of Technology Informatics and Engineering, 5*(2), 60-74. https://doi.org/10.51903/jtie.v5i2.544