International Journal of Current Sciences and Engineering Intelligence
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
website: *https://ijcsi.com*
Volume 01, Issue 01, September 2026
OPEN ACCESS
E-ISSN :

Cost Aware Retrieval and Abstention for Question Answering in Institutional Repositories

Vivek Menon[1], Prachi Nair[2], Ankit Rao[3]
[1,2,3] Students, MIT Arts, Commerce and Science College, Alandi, Pune

Article Info

Article History:
Published: 14 Nov 2026

Publication Issue:
Volume 01, Issue 01
September 2026

Page Number:
6-12

Corresponding Author:
Vivek Menon

Abstract:
Institutional question answering requires a balance between answering useful questions, controlling computation, and avoiding unsupported claims. This paper develops an analytical framework that selects basic retrieval, expanded retrieval, or abstention according to expected decision utility. A transparent scenario contains 1,000 queries divided into strong, ambiguous, and insufficient evidence groups. Correctness probabilities and processing costs are stipulated inputs rather than measurements of an implemented language model. Four policies are compared using coverage, selective accuracy, expected errors, computation, and a loss function. Under the main assumptions, the proposed cascade answers 80 percent of queries with 93 percent expected selective accuracy, uses 1,900 processing units, and achieves utility of 386. Always using expanded retrieval consumes 4,000 units and achieves utility of 74. Sensitivity analysis shows that the preferred policy changes when error penalties increase or confidence estimates become optimistic. The contribution is a reproducible decision analysis explaining when additional retrieval is worthwhile and when deferral is preferable. The results support conditional design choices and establish requirements for subsequent empirical validation; they do not constitute a benchmark of any deployed question answering system.

Keywords: Retrieval augmented generation; selective prediction; institutional repositories; confidence calibration; abstention; decision utility.

1 Introduction

Universities and research organizations store knowledge in regulations, circulars, course handbooks, technical reports, and administrative notices. Users frequently need a short answer that combines information from several documents. A conventional search engine returns passages, leaving interpretation to the user. A generative interface can summarize those passages, but an apparently fluent answer may combine an outdated regulation with a current exception or attach a citation that does not support its central claim. Reliability therefore depends on the relationship between the question, the retrieved evidence, and the final answer, rather than on linguistic fluency alone.

The transformer architecture established an effective foundation for sequence modeling [1], while bidirectional representation learning improved the extraction of contextual meaning [2]. Encoder and decoder pretraining further supported generation from supplied text [3]. These advances make repository assistants technically feasible, but they do not determine whether an institution should answer every question automatically. A wrong examination deadline can impose a different operational cost from an imperfect description of a research laboratory. A practical system must represent these differences through explicit decision rules and through an accessible route to an authorized human source.

Retrieval is also not a binary resource. A first pass can search a small index and assemble a few candidate passages. An expanded pass can search additional collections, resolve version conflicts, and rerank evidence. The expanded pass may improve an ambiguous answer while contributing little to a question already supported by a clear policy statement. Conversely, more search cannot necessarily answer a question whose relevant document is absent. The design problem

Copyright © 2026 The Author(s) : This is an open access article under the CC BY license (http://creativecommons.org/licenses/by-nc/4.0/)

is therefore to allocate additional computation where its expected benefit exceeds its cost and to recognize when available evidence remains inadequate.

This paper asks whether a simple utility based cascade can improve that allocation under transparent assumptions. It contributes a three action decision model, a fully specified numerical scenario, and sensitivity tests that expose the consequences of calibration error and changing error penalties. The analysis concerns repository information services, not autonomous legal, medical, or disciplinary decisions. Its contribution is explanatory rather than an empirical claim of architectural superiority. Every numerical probability in the scenario is an assumed input, and every reported result follows from arithmetic on those inputs. This separation makes the proposed logic auditable before a costly implementation is attempted.

2 Detailed Literature Review

Dense passage retrieval represents questions and passages in a shared embedding space, making semantic matching possible when their surface vocabulary differs [4]. Retrieval augmented generation combines retrieved documents with a generative model so that external information can inform the output [5]. These studies establish the value of connecting generation to an explicit collection, but retrieval success and answer correctness remain different quantities. A relevant passage can contain an exception that a generator overlooks, and a correct sentence can still be accompanied by an inappropriate citation.

Izacard and Grave investigate the use of multiple retrieved passages within generative question answering [6]. Their work motivates an expanded retrieval option, although it does not imply that increasing the number of passages always improves a local service. Sentence embeddings offer a practical representation for semantic retrieval [7], while BEIR provides evidence that retrieval performance should be examined across heterogeneous tasks [8]. For an institutional repository, disciplinary vocabulary, scanned documents, and policy versioning can change the conditions under which a general retrieval model succeeds.

Confidence calibration is central to turning a predictive score into a decision. Guo and colleagues show that neural network confidence can require postprocessing before it reflects correctness frequency [9]. Geifman and El-Yaniv formalize selective classification, where a system trades coverage for lower risk by rejecting some cases [10]. These contributions support reporting the accuracy of answered queries together with the fraction answered. Selective accuracy alone can look impressive when a system silently excludes most difficult questions.

Abstention can also interact with the quality of supervision. Thulasidasan and colleagues study abstention as a way of dealing with label noise during learning [11]. That mechanism differs from the operational deferral proposed here, but it highlights the difficulty of learning from unreliable correctness labels. Kadavath and colleagues examine what language models know about their own answers [12]. Their findings motivate confidence investigation without establishing that a model's verbal statement of certainty is already a calibrated probability for a particular institutional task.

TruthfulQA evaluates responses to questions that can elicit familiar falsehoods [13]. FActScore instead decomposes generated text into atomic factual claims for finer evaluation [14]. Together, these approaches suggest that a single answer level success label can hide distinct failure modes. For example, a response may state the correct deadline but invent an exemption. A practical evaluation should therefore retain both the binary decision outcome used in the present model and a more detailed annotation of unsupported claims, missing qualifications, and contradictory statements.

Citation generation introduces another evaluation layer. Gao and colleagues assess language models that produce answers with citations [15], while Ragas provides measures directed at retrieval augmented generation [16]. The present framework treats evidence quality as an input to the routing decision, not as a synonym for a citation count. Multiple references to the same irrelevant passage do not increase support. Automated evaluation can assist screening, but a locally adjudicated sample remains necessary to determine whether citations justify the answer under the institution's actual rules.

Self-RAG connects retrieval and generation with reflective assessment [17]. Corrective Retrieval Augmented Generation introduces retrieval quality evaluation and corrective actions [18]. These works make adaptive retrieval a substantive research direction. The analytical model here uses a deliberately smaller action space so that the reason

for each routing choice can be inspected. It does not implement either architecture or reuse their benchmark results. Instead, it asks which policy would be preferred if the relevant correctness and cost estimates were available.

Longer context is not equivalent to reliable use of context. Liu and colleagues show that the placement of relevant information within long inputs can affect performance [19]. Instruction tuning with human feedback addresses another part of the interaction problem by improving alignment with requested behavior [20]. Neither capability removes the need to evaluate evidence sufficiency. The focused literature synthesis therefore supports three complementary requirements: retrieval should be tested in the target collection, confidence should be calibrated on locally defined outcomes, and the answering policy should explicitly value deferral. This is a narrative synthesis of selected foundational and evaluation studies, not an exhaustive systematic review.

3 Detailed Methodology & Framework

3.1 Decision setting and scenario construction

The unit of analysis is one user query. A basic retrieval pass is performed for every query and costs one abstract processing unit. The system then either answers using that pass, performs an expanded pass costing three additional units, or defers the answer. Processing units are normalized computational costs rather than milliseconds, tokens, or currency. Deferral means explaining that the available evidence is insufficient and presenting an authorized contact or source route. The analysis does not assume that deferred queries are automatically solved or that human review is free.

The scenario contains 500 strong evidence queries, 300 ambiguous evidence queries, and 200 insufficient evidence queries. Basic correctness probabilities are respectively 0.96, 0.70, and 0.30. Expanded retrieval probabilities are 0.97, 0.88, and 0.55. These group probabilities are intentionally constructed to represent small, substantial, and inadequate retrieval gains. They are not estimated from a hidden dataset. Group membership is assumed observable to the routing procedure; the difficulty of learning that grouping is an explicit limitation to be investigated during implementation.

3.2 Utility and routing logic

A correct automatic answer earns one utility unit, an incorrect answer incurs a penalty of five, and deferral incurs a penalty of 0.2. Each processing unit incurs a penalty of 0.02. Equation 1 aggregates these quantities over the query workload. The symbols denote expected correct answers, expected wrong answers, deferred queries, and total processing units. Expected counts may be fractional because they represent the mean implied by the model, not a realized experiment.

$$U = N_c - 5N_w - 0.2N_a - 0.02C$$

Equation 1. Aggregate expected utility for correct answers, wrong answers, deferrals and processing cost.

For an individual query, the basic retrieval cost has already been incurred and is common to all remaining actions. A basic answer has expected value equal to six times its correctness probability minus five. Deferral has value minus 0.2. Consequently, basic answering is preferable to deferral when correctness is at least 0.80. An expanded answer must also recover its additional cost of 0.06 utility units, giving a threshold of 0.81. Expanded retrieval is preferable to basic answering when its correctness gain exceeds 0.01. Ties favor basic retrieval to avoid unnecessary computation.

These rules produce a specific cascade. Strong evidence queries receive basic answers because the gain from expansion exactly matches its cost. Ambiguous queries receive expanded retrieval because their probability improves by 0.18. Insufficient evidence queries are deferred because even their expanded probability remains below the answering threshold. Figure 1 places this rule within a possible operational architecture. Repository version controls and citation checks support the confidence estimate, while an audit sample provides a feedback path for recalibration.

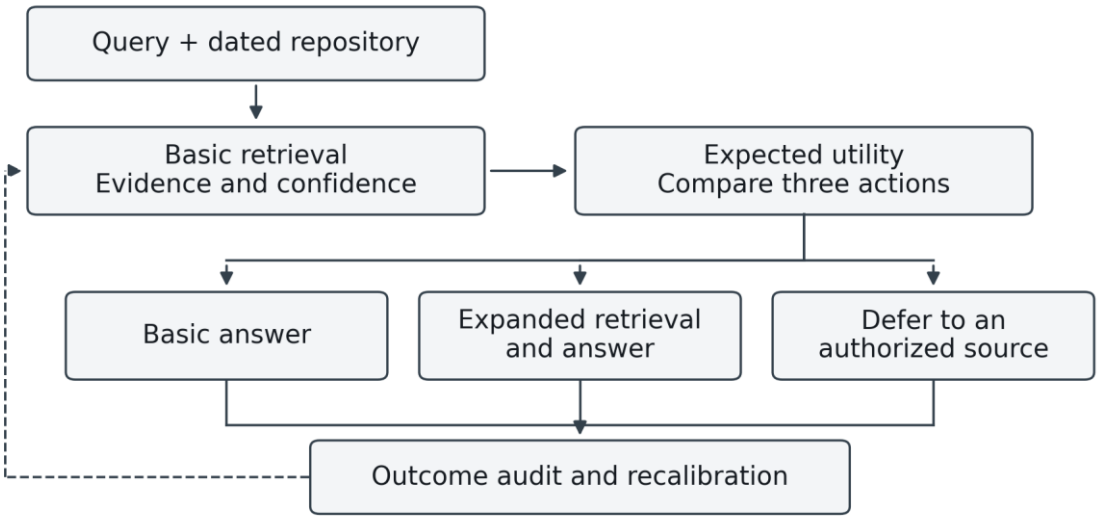


Figure 1. Proposed retrieval and deferral framework. The decision rule is calculated analytically; the audit feedback represents a future validation workflow.

3.3 Comparators and evaluation procedure

The first comparator answers every query after basic retrieval. The second always uses expanded retrieval and answers every query. The third answers only the strong evidence group after basic retrieval, deferring the other groups. All policies incur the initial retrieval cost for all 1,000 queries. The comparison therefore does not give abstention an artificial advantage by removing the cost needed to recognize uncertainty. The cascade differs from the conservative comparator only in its decision to expand and answer the ambiguous group.

Evaluation multiplies each group's size by its relevant probability to obtain expected correct answers. Expected wrong answers are the number answered minus expected correct answers. Coverage is the answered fraction of all queries, while selective accuracy is expected correct answers divided by answered queries. These measures are reported alongside utility and computation so that improvements in one cannot conceal deterioration in another. The calculation is deterministic and has no random seed, sampling distribution, confidence interval, or statistical significance test.

Sensitivity analysis varies the incorrect answer penalty from one to ten and recomputes the preferred action for each group. A second test keeps the main routing policy but reduces the ambiguous group's actual expanded correctness from 0.88 to 0.78, representing an optimistic calibration error. A reproducible script implements the arithmetic directly. An empirical extension would require separate development, calibration, and evaluation samples, document version isolation, and adjudication of answer correctness against dated source passages. Those requirements are part of the proposed validation framework, not completed experiments.

4 Detailed Results and Analysis

4.1 Main policy comparison

Table 1 reports the calculated outcomes. Basic answering produces 750 expected correct answers and 250 expected wrong answers. Although it has complete coverage and the lowest computational burden, its utility is negative under the stated error penalty. Always expanding improves expected correctness to 859 answers, but it still answers the insufficient evidence group and spends additional resources on the strong group. Its positive utility of 74 is therefore substantially below the cascade's value.

Table 1. Calculated comparison for the 1,000-query scenario.

Policy	Coverage (%)	Selective accuracy (%)	Expected errors	Cost (units)	Utility
Always basic	100.0	75.0	250	1,000	-520
Always expanded	100.0	85.9	141	4,000	74
Conservative basic	50.0	96.0	20	1,000	260
Cost aware cascade	80.0	93.0	56	1,900	386

All correctness rates and costs are assumed inputs. Expected outcomes are not measured benchmark results.

The cascade produces 744 expected correct answers, 56 expected wrong answers, and 200 deferrals. Its 93 percent selective accuracy is lower than the conservative policy's 96 percent, but it answers 300 additional queries. These additional answers contribute 264 expected correct outcomes and 36 expected errors. After the cost of expansion and the avoided deferral penalties are considered, this extension increases aggregate utility from 260 to 386. The main benefit is therefore a better balance between service coverage and error cost, not the highest accuracy considered in isolation.

Relative to always expanding, the cascade uses 52.5 percent fewer processing units and generates approximately 60.3 percent fewer expected wrong answers. These percentages follow from the scenario and must not be interpreted as measured savings for a commercial model. The cascade also produces 115 fewer expected correct answers because it declines difficult queries. Whether that exchange is desirable depends on the institution's valuation of delay and error. The utility coefficients make this judgment visible instead of burying it within a headline accuracy score.

4.2 Sensitivity and failure conditions

When the incorrect answer penalty falls to one, the preferred policy uses basic retrieval for the strong group and expanded retrieval for both remaining groups. Coverage rises to 100 percent, expected selective accuracy becomes 85.4 percent, and computation reaches 2,500 units. When the error penalty rises to ten, expanded retrieval becomes worthwhile even for the strong group, but the ambiguous and insufficient groups are deferred. Coverage then falls to 50 percent with 97 percent selective accuracy. These changes show that the action boundary is a decision parameter rather than an intrinsic property of the language model.

The calibration stress test is more consequential than a modest computation error. If the ambiguous group's expanded correctness is actually 0.78, the original cascade's expected errors rise to 86 and its utility falls to 206. The conservative policy retains utility of 260 and becomes preferable. With the main loss coefficients, expanded answering of ambiguous queries must achieve at least 0.81 correctness to outperform deferral. This breakpoint gives developers a concrete local validation target, provided that uncertainty around the estimate is also considered.

The analysis does not establish that the three evidence groups can be detected reliably. It also ignores correlations between failures, retrieval latency variability, changing document availability, and the cost of unresolved deferrals. Those omissions constrain generalization but do not invalidate the arithmetic. The appropriate interpretation is conditional: if the group structure, probabilities, and costs hold, the cascade is preferred among the evaluated actions. A deployment study should challenge each assumption and retain a simple conservative comparator throughout evaluation.

5 Future Enhancement

An empirical prototype should construct a dated repository benchmark containing answerable questions, conflicting versions, missing information, and questions that require cross-document reasoning. Human annotators should record the specific passage supporting each material claim and adjudicate disagreements before calibration. Evaluation should include citation validity, unsupported statement frequency, response latency, and the proportion of deferred

requests eventually resolved. These outcomes would show whether the modeled utility corresponds to a useful institutional service.

The framework can also replace fixed groups with calibrated query level estimates and uncertainty bounds. Additional retrieval actions could include optical character recognition, specialist indexes, or temporal filtering. Routing should become more conservative when calibration data are sparse or document distributions change. Deferral costs should be estimated from actual response workflows, and users should receive explanations that identify missing evidence without exposing internal confidence scores as guarantees. These extensions require validation before their effects can be claimed.

6 Conclusion

Cost aware retrieval and abstention provide a transparent way to connect question answering performance with institutional priorities. In the specified analytical scenario, the cascade achieves greater utility than always answering, always expanding, or applying a conservative basic retrieval rule. Its advantage arises from concentrating additional retrieval on queries that can plausibly benefit and declining queries whose evidence remains insufficient. The comparison also shows why selective accuracy, coverage, and computation should be reported together.

The strongest limitation is the dependence on assumed and correctly calibrated probabilities. A ten percentage point overestimate for one group reverses the preference between the cascade and the conservative policy. Consequently, the paper supports an auditable design and a testable validation plan rather than a claim of deployment readiness. Institutional adoption should follow local evaluation of evidence grouping, calibration, error consequences, and the operational handling of deferred questions.

References

[1] Vaswani, A., et al. (2017). Attention Is All You Need. arXiv preprint, arXiv:1706.03762. Source

[2] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint, arXiv:1810.04805. Source

[3] Lewis, M., et al. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv preprint, arXiv:1910.13461. Source

[4] Karpukhin, V., et al. (2020). Dense Passage Retrieval for Open-Domain Question Answering. arXiv preprint, arXiv:2004.04906. Source

[5] Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint, arXiv:2005.11401. Source

[6] Izacard, G., and Grave, E. (2020). Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. arXiv preprint, arXiv:2007.01282. Source

[7] Reimers, N., and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint, arXiv:1908.10084. Source

[8] Thakur, N., et al. (2021). BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. arXiv preprint, arXiv:2104.08663. Source

[9] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. Proceedings of ICML, PMLR 70, 1321-1330. Source

[10] Geifman, Y., and El-Yaniv, R. (2017). Selective Classification for Deep Neural Networks. arXiv preprint, arXiv:1705.08500. Source

[11] Thulasidasan, S., et al. (2019). Combating Label Noise in Deep Learning Using Abstention. arXiv preprint, arXiv:1905.10964. Source

[12] Kadavath, S., et al. (2022). Language Models (Mostly) Know What They Know. arXiv preprint, arXiv:2207.05221. Source

[13] Lin, S., Hilton, J., and Evans, O. (2021). TruthfulQA: Measuring How Models Mimic Human Falsehoods. arXiv preprint, arXiv:2109.07958. Source

[14] Min, S., et al. (2023). FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. arXiv preprint, arXiv:2305.14251. Source

[15] Gao, T., Yen, H., Yu, J., and Chen, D. (2023). Enabling Large Language Models to Generate Text with Citations. arXiv preprint, arXiv:2305.14627. Source

[16] Es, S., James, J., Espinosa Anke, L., and Schockaert, S. (2023). Ragas: Automated Evaluation of Retrieval Augmented Generation. arXiv preprint, arXiv:2309.15217. Source

[17] Asai, A., et al. (2023). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. arXiv preprint, arXiv:2310.11511. Source

[18] Yan, S.-Q., Gu, J.-C., Zhu, Y., and Ling, Z.-H. (2024). Corrective Retrieval Augmented Generation. arXiv preprint, arXiv:2401.15884. Source

[19] Liu, N. F., et al. (2023). Lost in the Middle: How Language Models Use Long Contexts. arXiv preprint, arXiv:2307.03172. Source

[20] Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. arXiv preprint, arXiv:2203.02155. Source