PAY TO REVEAL: ADAPTIVE INFORMATION ACQUISITION UNDER COMMUNICATION BUDGETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) grounds language-model answers in external corpora, and multi-hop questions often make later retrieval depend on evidence found earlier. Existing iterative systems typically query body-derived indexes or rerank retrieved candidate pools. We study a complementary decision: a controller can enumerate document titles, but each body must be acquired under a limited read budget. We formalize this setting as *Pay to Reveal*. A policy may acquire at most K documents; each acquisition returns a question-conditioned excerpt capped at L under a character-based token estimate; and accumulated excerpts guide later choices. We propose BM25-RF, an iterative pseudo-relevance-feedback policy that expands the query with acquired excerpts and reranks unopened titles with BM25. We also propose Residual-CMG, which adds to the static title-BM25 score the difference between a support-document classifier's acquired-state score and its empty-state score at the same step. The classifier is trained only on the training split. Under an explicit expected-progress condition, a finite-K analysis relates per-step selection gaps to final support-document recall. We evaluate six methods under one protocol on 22,246 retrieval questions and 3,000 Qwen3-4B-Instruct-2507 answer questions from 2WikiMultiHopQA, MuSiQue, and HotpotQA FullWiki. Relative to Static-BM25, BM25-RF changes answer F1 by $+.067$, $+.049$, and $-.028$ across the three datasets. Adding the static title-BM25 score to the learned score improves support AP over Direct-CMG on all three datasets.

1 INTRODUCTION

Retrieval-augmented generation (RAG) augments a generator with evidence retrieved from external non-parametric memory (Lewis et al., 2020). Multi-hop QA requires evidence chains across documents, so later retrieval often depends on an entity or relation found earlier (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022). IRCoT interleaves reasoning and retrieval (Trivedi et al., 2023); Self-RAG learns retrieval and critique decisions (Asai et al., 2024); EfficientRAG learns filtering and next-hop queries (Zhuang et al., 2024); and STRIDE selects retrieval, inference, or query reformulation actions (Chen et al., 2026).

These controllers act through body-derived indexes, retrieved candidate pools, or searchable corpus interfaces (Trivedi et al., 2023; Zhuang et al., 2024; Wu et al., 2026; Hui et al., 2026). Title-driven screening studies selection while primary content remains hidden, and budgeted retrieval treats document access as a cost (Yoo & Jeong, 2026; Wu et al., 2026). We study their sequential intersection. Before each action, the policy sees a fixed title catalog and the excerpts acquired so far. It must name one unopened document, spend one acquisition, and receive a bounded observation from that document. The central question is how acquired evidence should change the scores of the remaining titles when no body-derived representation is available for an unopened document.

We formalize this decision as *Pay to Reveal*. The public benchmark catalog exposes titles. Each named-document acquisition consumes one of K steps. A shared deterministic query-conditioned extractive selector scores sentences in the acquired body by IDF-weighted question-term overlap with length normalization, packs them within the L-unit cap, and restores document order. Its returned excerpt becomes state for later choices and the final answer.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

A 2WikiMultiHopQA example makes the update concrete (Ho et al., 2020). Acquiring the article for *Everyday Is Christmas* reveals Jacky Cheung and promotes that still-unopened title to the next acquisition; the static title order selects *Everyday*. Figure 1 gives the full 0.00-to-21.56 ranking transition. The example motivates a state-dependent value update; standard BM25 remains an inherited scoring component.

We propose two policies for this shared decision. BM25-RF adapts classical pseudo-relevance feedback (Rocchio, 1971) by expanding the query with acquired excerpts and reranking unopened titles with BM25. Residual-CMG adds the acquired-state support-document classifier score minus the empty-state score at the same step to the static title-BM25 score, with the classifier trained only on the training split. As controls, Direct-CMG uses the acquired-state classifier score alone, whereas Prior+Full-CMG adds the full acquired-state score to the same static title-BM25 score, distinguishing the value of retaining static relevance from the effect of subtracting the empty-state score. Finally, under an explicit expected-progress condition, a finite-K analysis relates per-step selection gaps to final support-document recall.

We evaluate six acquisition methods under one protocol on 2WikiMultiHopQA, MuSiQue, and HotpotQA FullWiki (Yang et al., 2018). Every method receives the same title catalog, $K = 6$ budget, extractor, and answer model. Evidence selection is evaluated on 22,246 questions; Qwen3-4B-Instruct-2507 (Yang et al., 2025) answers 1,000 questions per dataset. Sparse feedback improves Static-BM25 on 2Wiki and MuSiQue and declines on HotpotQA. Both prior-preserving learned scores recover most of Direct-CMG's AP deficit; Residual-CMG exceeds Prior+Full-CMG by $.0171$, $.0002$, and $.0049$ AP.

Our contributions are:

- a budgeted RAG formulation with title-only pre-acquisition information, named-document actions, acquired-excerpt state, and explicit K/L constraints;
- two acquisition policies: BM25-RF, which adapts iterative pseudo-relevance feedback to sequential document acquisition, and Residual-CMG, which combines static title relevance with the change in a support-document classifier's score after evidence is acquired;
- a conditional finite-horizon bound linking selection gaps to support-document recall under an expected-progress premise; and
- a common three-dataset evaluation of answer quality, support-document acquisition, and measured communication, showing dataset-dependent sparse-feedback effects and consistent support-AP gains from retaining static title relevance in the learned scores.

2 Related Work

Sequential and evidence-conditioned RAG. IRCoT interleaves reasoning with open-corpus retrieval, Self-RAG learns when to retrieve and critique passages, and EfficientRAG replaces repeated language-model calls with learned filtering and next-hop query modules (Trivedi et al., 2023; Asai et al., 2024; Zhuang et al., 2024). KiRAG, CIRAG, S2G-RAG, PAR^2-RAG, CoSearch, and STRIDE use retrieved structure, accumulated trajectories, evidence gaps, planned query refinement, or learned action selection to obtain multi-hop support (Fang et al., 2025; Wei et al., 2026; Li et al., 2026a;b; Zeng et al., 2026; Chen et al., 2026). DeepControl, DynaKRAG, EviReform, and Hi-Q make intermediate evidence explicit when deciding whether or how to continue retrieval (Xiong et al., 2026; Zhou et al., 2026; Xu & Li, 2026; Kim et al., 2026). These systems establish the value of adapting retrieval to observed evidence. Their controllers act through body-derived search results or candidate representations; Pay to Reveal studies the choice of a named document before its body is observable.

Relevance feedback. Classical relevance feedback updates a query from material treated as relevant, with Rocchio as a canonical vector-space construction (Rocchio, 1971). Dense-Rocchio instantiates that family in our title-embedding space. BM25-RF uses direct query concatenation with acquired excerpts. Its contribution is the sequential acquisition policy and information protocol; neither the feedback principle nor concatenation is claimed as a new retrieval primitive.

Budgeted retrieval and corpus interaction. When Knowledge Is Not Free chooses evidence tiers after a full-content retriever forms a candidate pool; Inference-Time Budget Control and Budget-

aware Active-RAG allocate retrieval operations under call or generation limits; PACE and PAGE-RAG allocate reranking or reader-context capacity (Wu et al., 2026; Fang et al., 2026; Qian et al., 2026; Cai & Zafarani, 2026; Deng et al., 2026). Interact-RAG, Sieve, Direct Corpus Interaction, RISE, DR-DCI, and AtlasNav expose search, inspection, navigation, and context-management operations over corpus-derived interfaces (Hui et al., 2026; Wang et al., 2026; Li et al., 2026c; Zhuang et al., 2026; Lu et al., 2026; Guo et al., 2026). Pay to Reveal fixes the operation—open one named document—and assigns unit cost to the body whose content becomes observable, with L capping the returned excerpt.

Content-hidden selection and communication. Title-driven forensic screening selects files from names while primary content is withheld, providing a close one-pass precedent (Yoo & Jeong, 2026). Our setting adds sequential feedback from each acquired body. BANDMAS schedules already generated semantic packets, IterCOMP evaluates and compresses retrieved documents, and AgentPrune and DySCo select links among existing agent states or messages (Dong & Lin, 2026; Yun & Kim, 2026; Zhang et al., 2025; Gou & Liu, 2026). Those decisions allocate observed content; our decision determines which content becomes observed.

Table 1: The action and resource that distinguish Pay to Reveal from adjacent settings.

Setting	Adaptive action	Information before action	Constrained resource
Sequential RAG	issue/rewrite a query; retrieve or infer	evidence and body-index results	calls, depth, tokens, latency
Budgeted RAG	choose operation, tier, or exposure	retrieved pool or returned passages	access cost, calls, context
Corpus interaction	search, inspect, navigate, shape context	corpus indexes, structure, prior results	rounds and returned content
Title screening	select files from names	case context and exposed names	one-pass review workload
Pay to Reveal	acquire a particular unopened document	fixed titles and earlier excerpts	unit-cost bodies K; excerpt cap L

Adaptive submodular acquisition. Adaptive submodularity supports greedy guarantees for sequential decisions under partial realizations, with extensions to weaker diminishing returns and stochastic or varying costs (Golovin & Krause, 2011; Fujii & Sakaue, 2019; Parthasarathy, 2020; Fern et al., 2017). Bridge questions may exhibit complementarity: observing one document can increase another's value. We therefore use an expected progress condition over policy histories and accumulate per-step selection gaps into a finite-budget bound on support-document recall.

3 Pay-to-Reveal Acquisition

For question q, let E be a fixed document catalog constructed from the public benchmark. Before selection, each $e \in E$ exposes only its dataset-provided title; a document ID supports identity and deterministic tie-breaking. The body $B(e)$ remains unavailable. Selecting e sends an acquisition request to a shared environment-side channel. The deterministic extractor H_q then reads $B(e)$, ranks its sentences by length-adjusted IDF overlap with q, selects within a per-document cap L, and restores document order, yielding $\Phi_q(e) = H_q(B(e))$. The full body remains on the environment side; the policy and final answer model receive only $\Phi_q(e)$. All policies use the same extractor.

We fix $L = 256$ under the character-based token estimate below; an excerpt may be shorter. For any string x, define

$$\widehat{c}(x) = \left\lceil \frac{n_{\mathrm{ASCII}}(x)}{4} + \frac{n_{\mathrm{nonASCII}}(x)}{1.6} \right\rceil. \tag{1}$$

K limits opened documents, whereas L limits the content returned by each opening.

After t acquisitions, $S_t = \{e_1, \ldots, e_t\}$ and the observable state is

$$\psi_t = ((e_1, \Phi_q(e_1)), \ldots, (e_t, \Phi_q(e_t))). \tag{2}$$

A policy may choose e_{t+1} using q, exposed titles, the shared cap, tie-breaking identifiers, and ψ_t. It cannot consult the body, body embedding, or body-derived index of any unselected document. The

common three-dataset evaluation uses $K = 6$ unit-cost acquisitions; protocol-separated appendix analyses report additional budgets.

Let $(X, \Omega) \sim D$, where X contains the question, public catalog and protocol parameters, and Ω contains hidden bodies and the nonempty support set $G(\Omega)$. The observation $O_q(e, \Omega) = H_q(B_\Omega(e)) = \Phi_q(e)$ is deterministic for a fixed task and extractor. D specifies no independence or parametric distribution. Let $\mathcal{H}_t = (X, \psi_t)$ include any independent policy randomization. Define support-document recall, used as the coverage utility, by $f_{\mathrm{cov}}(S, \Omega) = |S \cap G(\Omega)|/|G(\Omega)|$ and

$$\Delta(e \mid \mathcal{H}_t) = \mathbb{E}[f_{\mathrm{cov}}(S_t \cup \{e\}, \Omega) - f_{\mathrm{cov}}(S_t, \Omega) \mid \mathcal{H}_t]\,. \tag{3}$$

For each hidden state, coverage is modular and monotone; the completion utility $f_{\mathrm{all}}(S, \Omega) = \mathbf{1}\{G(\Omega) \subseteq S\}$ is monotone with complementarity. For two support documents these utilities follow $0, 1/2, 1$ and $0, 0, 1$, respectively. Answer F1 of $g(q, \psi_t)$ is measured separately and can decrease. Adaptive submodularity concerns the same candidate's conditional gain under nested histories (Golovin & Krause, 2011). The policy need not know D or these oracle gains.

Communication accounting. Let $C_{\mathrm{ex}}(\pi, q)$ be the character-based token estimate for the acquired excerpts and $C_{\mathrm{serial}}(\pi, q)$ the corresponding estimate after titles, identifiers, and formatting are added. We report both using Equation (1). The title catalog and indexes are preloaded and shared. Catalog size is reported as a document count because a common serialized catalog volume was not measured across all three corpora. Answer-model usage is reported from the tokenizer counts returned for each response. The online resource results cover acquisition traffic, local selection time, and answer-model tokens under the shared protocol.

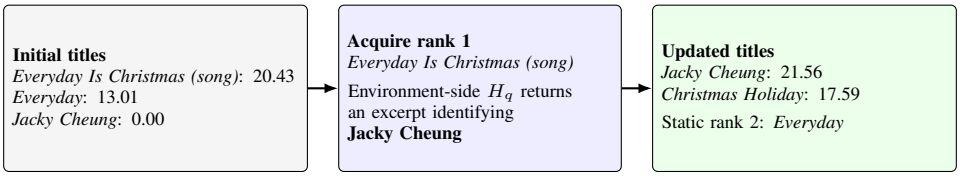

Figure 1: A full-corpus 2Wiki transition. The environment-side extractor returns the acquired excerpt, which promotes the second support document from BM25 score 0.00 to 21.56; the static ranking does not.

Gold answers, support labels, and decompositions are used only for training-set statistics or evaluation; they do not enter test-time selection.

4 Methods and Analysis

We propose BM25-RF and Residual-CMG as two implementations of the same adaptive value-update problem. Both rank unopened candidates from the question, exposed titles, and acquired excerpts; neither receives a selected document's full body. They differ in how the acquired state changes the score of the remaining titles.

In the unified three-dataset evaluation, Static-BM25 and BM25-RF search the complete title index. At each step, Direct-CMG, Residual-CMG, and Prior+Full-CMG score the same candidate set, constructed as the union of the top 256 question–title BM25 matches and the top 256 acquired-excerpt–title BM25 matches after acquired documents are removed. HotpotQA training rows use only documents in each training question's provided context, while evaluation uses the complete fixed title index. Regularization is selected on an official-training validation subset. The common evaluation fixes $\lambda = 0.1$ before evaluating any held-out question; this is the previously developed 2Wiki/MuSiQue setting and is applied unchanged to HotpotQA.

4.1 BM25 relevance feedback

BM25 combines term matching, inverse document frequency, term-frequency saturation, and length normalization (Robertson & Zaragoza, 2009). Let $R(\psi_t) = \Phi_q(e_1) \oplus \cdots \oplus \Phi_q(e_t)$ concatenate

acquired excerpts in order. BM25-RF is an iterative pseudo-relevance-feedback policy: it expands the query with $R(\psi_t)$ and scores each unopened document by

$$s_t^{\text{sparse}}(e) = \text{BM25}(\text{title}_e,\ q \oplus R(\psi_t))\,. \tag{4}$$

The policy acquires the highest-scoring document. At $t = 0$, it equals Static-BM25; later changes can arise only from acquired content. Scoring and state updates are deterministic at inference time, with fixed BM25 parameters.

4.2 Residual state correction

Residual-CMG separates a static relevance prior from the predicted change induced by acquired evidence:

$$s_t^{\text{res}}(e) = s_0(e \mid q) + \lambda[f_\theta(e \mid q, \psi_t) - f_\theta(e \mid q, \varnothing, t)]\,. \tag{5}$$

Here s_0 is normalized title BM25. For each dataset, f_θ is the uncalibrated logit of an L2-regularized logistic classifier trained only on the official training split. A row represents an unselected candidate at a reachable acquisition state; the target is one exactly when that candidate is a support document still unacquired. Features cover static question–title relevance, overlap and mentions induced by acquired excerpts, step, and their interactions. Training states include the empty state, incomplete support prefixes, and Static-BM25 top-one and top-two prefixes.

For raw title-BM25 score $b_q(e)$, $s_0(e \mid q) = b_q(e)/\max_{j\in E} b_q(j)$ when the maximum is positive and zero otherwise. The training analogue normalizes over the titles in that training task.

The bracketed term is the acquired-state classifier score minus the empty-state score at the same step. It cancels query-only terms and is zero before acquisition, so the first choice follows s_0. CMG reflects a conditional-marginal-gain motivation; the implemented quantity is a selection score. Direct-CMG ranks by $f_\theta(e \mid q, \psi_t)$ alone. Prior+Full-CMG uses $s_0(e \mid q) + \lambda f_\theta(e \mid q, \psi_t)$. The three variants share the classifier, target, training rows, features, candidate set, and $\lambda = 0.1$ where applicable. The comparison with Direct-CMG evaluates the combined static score and learned update. The comparison with Prior+Full-CMG changes only the subtraction of the empty-state score.

4.3 Finite-horizon guarantee

The analysis connects local selection errors to expected support-document recall, used as the coverage utility. Let $F_t = \mathbb{E}[f_{\text{cov}}(S_t, \Omega)]$ and $\text{OPT} = \max_{\pi'\in\mathcal{P}_K} \mathbb{E}[f_{\text{cov}}(S_K^{\pi'}, \Omega)]$. The policies in $\mathcal{P}_K$ share D, the observation map, catalog $E = E(X)$ and at most K acquisitions. Expectations include task and policy randomness. Define the expected one-step *selection regret* by

$$\delta_t = \mathbb{E}\left[\max_{e\in E\setminus S_t} \Delta(e \mid \mathcal{H}_t) - \Delta(e_{t+1} \mid \mathcal{H}_t)\right]. \tag{6}$$

The selected action is the policy's actual action, including any title-based candidate restriction.

Theorem 1 (Finite-horizon coverage bound)**.** *For support-document recall with $F_0 = 0$, suppose the histories induced by (D, π, O) satisfy, for $t = 0, \ldots, K-1$ and a common $0 < \gamma \le 1$,*

$$\mathbb{E}\left[\max_{e\in E\setminus S_t} \Delta(e \mid \mathcal{H}_t)\right] \ge \frac{\gamma}{K}(\text{OPT} - F_t). \tag{7}$$

Then

$$F_K \ge \left[1 - \left(1 - \frac{\gamma}{K}\right)^K\right] \text{OPT} - \sum_{t=0}^{K-1} \left(1 - \frac{\gamma}{K}\right)^{K-1-t} \delta_t. \tag{8}$$

The progress condition averages over histories, allowing already-completed questions to have zero remaining gain. It is an analytical premise; the analysis neither derives a task-level lower bound on γ nor proves that either method controls δ_t. Held-out support labels yield a realized one-step gap r_t whose expectation $R_t = \mathbb{E}[r_t]$ upper-bounds δ_t under the same observation protocol and action reference. Appendix A proves this relation, gives the optional calibration lemma, and shows why clue-first tasks can force γ to be small. Both policies enter the bound through their actual choices. The result concerns support-document recall; answer quality is evaluated separately.

5 EXPERIMENTS

5.1 DATASETS AND PROTOCOL

We use one evaluation protocol on 2WikiMultiHopQA (Ho et al., 2020), MuSiQue-Answerable v1.0 (Trivedi et al., 2022), and HotpotQA FullWiki (Yang et al., 2018). Each policy sees the question, all document titles, and previously acquired excerpts. It may acquire $K = 6$ documents; for each acquisition, the shared query-conditioned extractor returns an excerpt capped at $L = 256$ under the character-based token estimate. Methods, information access, budgets, answer model, metrics, and statistical tests are held fixed across datasets.

Table 2: Unified evaluation matrix. Answer n is the fixed cohort evaluated for every method.

Dataset	Catalog docs	Retrieval n	Answer n
2WikiMultiHopQA	54,943	12,536	1,000
MuSiQue	21,099	2,401	1,000
HotpotQA FullWiki	5,233,329	7,309	1,000
Total	5,309,371	22,246	3,000

The 2Wiki catalog is built from all eligible development contexts; a prespecified 40-question development subset is excluded from evaluation. MuSiQue uses all 2,401 answerable development questions. On HotpotQA, 7,309 of 7,405 validation questions map every supporting title to the public FullWiki catalog. Gold documents are never added per question. The answer cohort is selected by a fixed hash rule before generation.

5.2 METHODS AND BASELINES

Table 3: Methods under the common protocol. Stars mark our methods.

Method	Role	Next-document score
Static-BM25	Static BM25 (Robertson & Zaragoza, 2009)	fixed $\mathrm{BM25}(\mathrm{title}_e, q)$
BM25-RF *	acquired-evidence sparse feedback	Equation 4
Direct-CMG	matched learned control	direct $f_\theta(e \mid q, \psi_t)$
Residual-CMG *	static BM25 plus classifier-score change	Equation 5
Static-BGE	Static BGE title retrieval (Xiao et al., 2023)	fixed question–title cosine
Dense-Rocchio	BGE–Rocchio feedback (Rocchio, 1971)	cosine after $q_t = \mathrm{norm}(q_0 + 0.75\,\mathrm{mean}_{i \le t}\, r_i)$

Static-BM25 and BM25-RF search the complete title index exactly. The dense methods use the complete fixed BGE-base-en-v1.5 title index. Direct-CMG, Residual-CMG, and Prior+Full-CMG score the same candidate set, constructed as the union of the top 256 question–title BM25 matches and the top 256 acquired-excerpt–title BM25 matches after opened documents are removed. Their per-dataset classifiers share training rows, labels, features, regularization, and inference candidates. We use the shared preset $\lambda = 0.1$ on all datasets. Dense-Rocchio fixes $\alpha = 1, \beta = .75$ without development-set tuning and serves as a reference baseline.

5.3 OUTCOMES AND UNCERTAINTY

Qwen3-4B-Instruct-2507 (Yang et al., 2025) answers from the six selected excerpts at temperature 0. Retrieval metrics compare acquired document IDs with public support IDs; the model-generated supporting-message field is not used for these tables. Support AP@6, recall@6, and all-support coverage@6 measure document acquisition. Normalized answer F1 and exact match measure downstream performance after the shared extractor. All contrasts are paired within qid and use 10,000 bootstrap resamples. Matched comparisons are BM25-RF–Static-BM25, Residual-CMG–Direct-CMG, and Dense-Rocchio–Static-BGE; cross-method references are BM25-RF–Direct-CMG

and Residual-CMG–Static-BM25. Holm correction is applied within each endpoint across all pre-specified dataset–pair tests.

We retain actual Qwen prompt and completion tokens, selection and answer runtimes, and peak memory. Local inference has no provider charge. Equal K and L fix the number of reads and their caps, while realized excerpt and serialized-payload sizes can differ. One-time corpus preparation, catalog transfer, and hardware ownership are outside the online-path accounting. Appendix E gives the complete prompt, eligibility rules, tie-breaking, validation, and ablations; Appendix F gives the resource ledger.

6 RESULTS

6.1 ANSWER QUALITY UNDER THE COMMON PROTOCOL

Table 4 reports the complete answer matrix at $K = 6$. BM25-RF gives the highest answer F1 on 2WikiMultiHopQA and MuSiQue, while Static-BGE is strongest on HotpotQA. Residual-CMG is the strongest method by exact match on 2Wiki and nearly matches Static-BM25 on MuSiQue and HotpotQA.

Table 4: Answer quality at $K = 6$ with Qwen3-4B-Instruct-2507. Every cell is F1 / exact match on the same 1,000 questions; bold marks the best value in each metric and dataset.

Method	2WikiMultiHopQA	MuSiQue	HotpotQA FullWiki
Static-BM25	.3931 / .3280	.0646 / .0350	.3007 / .2210
BM25-RF	**.4604** / .3820	**.1134** / **.0810**	.2731 / .2020
Direct-CMG	.3091 / .2510	.0627 / .0360	.2111 / .1570
Residual-CMG	.4484 / **.3830**	.0661 / .0370	.3006 / .2180
Static-BGE	.4011 / .3340	.1048 / .0670	**.3283** / **.2350**
Dense-Rocchio	.3567 / .2890	.1062 / .0650	.2672 / .1950

The matched BM25-RF–Static-BM25 answer F1 differences are $+.0674$ $[.0377, .0961]$ on 2Wiki, $+.0488$ $[.0313, .0661]$ on MuSiQue, and $-.0275$ $[-.0501, -.0062]$ on HotpotQA. The first two remain significant after endpoint-wise Holm correction ($p_{\text{Holm}} = .003$); the HotpotQA comparison has $p_{\text{Holm}} = .061$. Thus the sparse feedback update has a dataset-dependent effect under the same information and answer protocol.

On 2Wiki, the Residual-CMG–Direct-CMG answer F1 difference is $+.1393$ $[.1078, .1705]$. The corresponding differences are $+.0035$ $[-.0117, .0186]$ on MuSiQue and $+.0894$ $[.0696, .1099]$ on HotpotQA. The 2Wiki and HotpotQA effects remain significant after correction ($p_{\text{Holm}} = .003$), whereas the MuSiQue interval crosses zero. On HotpotQA, Residual-CMG raises Direct-CMG's F1 and reaches the static BM25 answer level (.3006 versus .3007). Dense feedback is also update-dependent: Dense-Rocchio lowers answer F1 relative to Static-BGE on 2Wiki and HotpotQA and is statistically indistinguishable on MuSiQue.

These answer comparisons establish the performance of the complete Residual-CMG score relative to Direct-CMG. The current answer matrix does not contain Prior+Full-CMG, so it does not attribute those answer gains specifically to empty-state subtraction.

6.2 SUPPORTING-DOCUMENT ACQUISITION

Table 5 evaluates document selection on all questions whose supporting documents are present in the evaluation corpus. The best method varies across datasets: BM25-RF leads all three retrieval measures on 2Wiki and all-support coverage on MuSiQue, while Static-BGE leads AP and recall on MuSiQue and all three measures on HotpotQA.

Table 5: Supporting-document acquisition at $K = 6$. Each cell is AP / recall / all-support coverage. Retrieval n is 12,536 for 2Wiki, 2,401 for MuSiQue, and 7,309 for HotpotQA. Bold marks the best value in each metric and dataset.

Method	2WikiMultiHopQA	MuSiQue	HotpotQA FullWiki
Static-BM25	.6429 / .6708 / .3432	.3147 / .3614 / .0508	.4214 / .4806 / .2233
BM25-RF	**.7719 / .8065 / .6104**	.3528 / .3814 / **.1437**	.3669 / .3824 / .1394
Direct-CMG	.5225 / .5443 / .1713	.2801 / .2832 / .0450	.3022 / .3101 / .0265
Residual-CMG	.6820 / .7115 / .4197	.3187 / .3671 / .0579	.4195 / .4788 / .2285
Static-BGE	.6396 / .6743 / .3577	**.3700 / .4304** / .0820	**.4479 / .5206 / .2634**
Dense-Rocchio	.5735 / .6283 / .3181	.3487 / .4006 / .0804	.3539 / .3973 / .0852

For support AP, BM25-RF changes Static-BM25 by $+.1290$ $[.1240, .1339]$, $+.0381$ $[.0306, .0458]$, and $-.0544$ $[-.0598, -.0492]$ across the three datasets. Residual-CMG improves over Direct-CMG on every dataset: $+.1594$ $[.1544, .1645]$, $+.0386$ $[.0309, .0463]$, and $+.1173$ $[.1122, .1224]$, with $p_{\text{Holm}} = .003$ throughout. Against Static-BM25, Residual-CMG improves AP on 2Wiki and MuSiQue and is statistically indistinguishable on HotpotQA. By contrast, Dense-Rocchio reduces AP relative to Static-BGE on all three datasets.

Both learned scores that retain static title BM25 achieve higher support AP than Direct-CMG on all three datasets. Prior+Full-CMG reaches $.6649$, $.3185$, and $.4147$ AP, compared with $.6820$, $.3187$, and $.4195$ for Residual-CMG. Relative to Direct-CMG, Prior+Full-CMG accounts for about 89%, 99%, and 96% of the rounded Direct-to-Residual AP increase. Residual-CMG exceeds Prior+Full-CMG by $.0171$, $.0002$, and $.0049$ AP. Paired intervals and answer-model outputs were not computed for this comparison, so these differences are reported descriptively without a significance claim. The supported conclusion is that retaining the static title-BM25 score improves support AP over Direct-CMG under the present learned model.

6.3 Communication and acquisition depth

At $K = 6$, mean serialized acquisition traffic ranges from 587 to 887 character-based token estimates per question across method–dataset cells. For BM25-RF, the means are 791.8, 861.9, and 632.5, compared with 694.4, 823.9, and 613.5 for Static-BM25. For Residual-CMG, they are 660.0, 811.5, and 620.7. These totals include the six returned excerpts together with titles, identifiers, and formatting. The comparison fixes the number of acquisitions and the per-document cap; realized traffic can still differ because excerpts vary in length. The title catalog is preloaded, and local search computation is reported separately. Exact excerpt and serialized totals appear in Table 8.

The 300-question prefix analysis also shows that answer quality need not increase monotonically with acquisition count. For example, Static-BGE on 2Wiki reaches .4202 F1 at four acquisitions and .4091 at six, and Dense-Rocchio on HotpotQA reaches .2813 at four and .2779 at six. This empirical behavior is compatible with the theoretical separation between monotone support coverage and a downstream answer model whose output can change in either direction. Complete prefix summaries and the structural Prior+Full-CMG ablation are reported in the appendix.

Across the common evaluation, sparse lexical feedback improves answer F1 and support AP on 2Wiki and MuSiQue, with the opposite direction on HotpotQA. Within the learned methods, adding the static title-BM25 score raises support AP over Direct-CMG on all three datasets. Residual-CMG has the highest support AP among the three learned variants, exceeding Prior+Full-CMG by $.0171$, $.0002$, and $.0049$ in the descriptive ablation.

7 Conclusion

Pay to Reveal captures a budgeted RAG decision in which acquired evidence updates the value of unopened documents. We propose BM25-RF, an iterative pseudo-relevance-feedback policy that expands the query with acquired excerpts and reranks unopened titles with BM25. We also

propose Residual-CMG, which adds the acquired-state support-document classifier score minus the empty-state score at the same step to the static title-BM25 score. BM25-RF improves answer F1 over Static-BM25 on 2Wiki and MuSiQue and decreases it on HotpotQA. Adding the static title-BM25 score to learned classifier scores yields higher support AP than Direct-CMG on all three datasets. Under an explicit expected-progress condition, the finite-horizon analysis relates per-step selection gaps to final support-document recall.

AI Use Statement

Generative AI tools assisted with literature organization and search, research ideation and hypothesis refinement, feedback on methodology and experimental design, code drafting and debugging, dataset cleaning and deterministic reformatting, mathematical formalization and proof drafting, result interpretation, manuscript drafting and editing, translation, and reference formatting. They were not used to generate synthetic datasets; surveys, interviews, recordings, and qualitative human-subject analysis are not applicable to this work. The authors defined the research question and system architecture, selected the experimental protocols, directed implementation and debugging, inspected the cited sources, tested the code, recomputed the reported statistics, and checked the mathematical and textual claims. The answer model evaluated in Section 5 is part of the experiment and is reported separately from research assistance. The authors take responsibility for the final manuscript, claims, code, and artifacts.

Reproducibility Statement

Section 3 specifies the information boundary and evaluation protocol; Section 4 defines both acquisition policies; and Section 5 reports the common three-dataset design, answer model, budget, metrics, and paired statistical comparisons. The appendices provide implementation details, the exact prompt, primary paired comparisons, the structural CMG ablation, acquisition-prefix results, resource accounting, and the proof of the theoretical result. These materials support independent reproduction under the same title-only acquisition setting.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Weibin Cai and Reza Zafarani. Less can be more: Relieving RAG bottlenecks via evidence frontloading and pressure-adaptive budgeting. *arXiv preprint arXiv:2608.25115*, 2026. URL https://arxiv.org/abs/2608.25115.

Wei Chen, Lili Zhao, Zhi Zheng, HuiJun Hou, and Tong Xu. STRIDE: Strategic iterative decision-making for retrieval-augmented multi-hop question answering. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2026. URL https://arxiv.org/abs/2604.17405. Accepted full paper.

Haokun Deng, Xunkai Li, Hongchao Qin, and Rong-Hua Li. PAGE-RAG: Provenance-aware graph evidence promotion for fixed-budget multi-hop retrieval-augmented generation. *arXiv preprint arXiv:2608.29753*, 2026. URL https://arxiv.org/abs/2608.29753.

Jiangwen Dong and Wanyu Lin. BANDMAS: Causality-inspired semantic packet scheduling for bandwidth-efficient multi-agent collaboration. *arXiv preprint arXiv:2608.00458*, 2026. URL https://arxiv.org/abs/2608.00458.

Jinyuan Fang, Zaiqiao Meng, and Craig MacDonald. KiRAG: Knowledge-driven iterative retriever for enhancing retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18969–18985. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.929. URL https://aclanthology.org/2025.acl-long.929/.

Zhengru Fang, Senkang Forest Hu, Zhonghao Chang, Yu Guo, Yihang Tao, Hongyao Liu, Mengzhe Ruan, Jun Huang, and Yuguang Fang. Inference-time budget control for LLM search agents. *arXiv preprint arXiv:2605.05701*, 2026. URL https://arxiv.org/abs/2605.05701.

Alan Fern, Robby Goetschalckx, Mandana Hamidi-Haines, and Prasad Tadepalli. Adaptive submodularity with varying query sets: An application to active multi-label learning. In *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, volume 76 of *Proceedings of Machine Learning Research*, pp. 577–592, 2017. URL https://proceedings.mlr.press/v76/fern17a.html.

Kaito Fujii and Shinsaku Sakaue. Beyond adaptive submodularity: Approximation guarantees of greedy policy with adaptive submodularity ratio. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2042–2051, 2019. URL https://proceedings.mlr.press/v97/fujii19a.html.

Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42:427–486, 2011. doi: 10.1613/jair.3278. URL https://jair.org/index.php/jair/article/view/10602.

Wanshuang Gou and Zihan Liu. Dynamic trust-aware sparse communication topology for LLM-based multi-agent consensus. *arXiv preprint arXiv:2606.01828*, 2026. URL https://arxiv.org/abs/2606.01828.

Hongyu Guo, Zhiyu Zheng, and Zhao Cao. Evidence blindness in direct corpus interaction: Persistent navigation with AtlasNav. *arXiv preprint arXiv:2608.24764*, 2026. URL https://arxiv.org/abs/2608.24764.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Yulong Hui, Chao Chen, Zhihang Fu, Yihao Liu, Jieping Ye, and Huanchen Zhang. Interact-RAG: Reason and interact with the corpus, beyond black-box retrieval. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b61f288da3c106f65d57b0d45b470b6b-Abstract-Conference.html.

Jueun Kim, Sungho Park, and Wook-Shin Han. Hi-Q: Hierarchical evidence-guided query refinement for multi-hop question answering. *arXiv preprint arXiv:2608.30468*, 2026. URL https://arxiv.org/abs/2608.30468.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Minghan Li, Junjie Zou, Xinxuan Lv, Chao Zhang, and Guodong Zhou. S2G-RAG: Structured sufficiency and gap judging for iterative retrieval-augmented QA. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25846–25862. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.acl-long.1185. URL https://aclanthology.org/2026.acl-long.1185/.

Xingyu Li, Rongguang Wang, Yuying Wang, Mengqing Guo, Chenyang Li, Tao Sheng, Sujith Ravi, and Dan Roth. PAR^2-RAG: Planned active retrieval and reasoning for multi-hop question answering. *arXiv preprint arXiv:2603.29085*, 2026b. URL https://arxiv.org/abs/2603.29085.

Zhuofeng Li, Haoxiang Zhang, Cong Wei, Pan Lu, Ping Nie, Yi Lu, Yuyang Bai, Shangbin Feng, Hangxiao Zhu, Ming Zhong, Yuyu Zhang, Jianwen Xie, Yejin Choi, James Zou, Jiawei Han, Wenhu Chen, Jimmy Lin, Dongfu Jiang, and Yu Zhang. Beyond semantic similarity: Rethinking retrieval for agentic search via direct corpus interaction. *arXiv preprint arXiv:2605.05242*, 2026c. URL https://arxiv.org/abs/2605.05242.

Yi Lu, Zhuofeng Li, Ping Nie, Haoxiang Zhang, Yuyu Zhang, Kai Zou, Wenhu Chen, Jimmy Lin, Dongfu Jiang, and Yu Zhang. DR-DCI: Scaling direct corpus interaction via dynamic workspace expansion. *arXiv preprint arXiv:2606.14885*, 2026. URL https://arxiv.org/abs/2606.14885.

Srinivasan Parthasarathy. Adaptive submodular maximization under stochastic item costs. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3133–3151, 2020. URL https://proceedings.mlr.press/v125/parthasarathy20a.html.

Pin Qian, Su Wang, Chong Peng, Junxian You, Lifei Liu, Haoran Yu, Yihang Chen, and Xiaochong Jiang. When should active RAG retrieve? a budget-aware evaluation of utility, calibration, and cost. KDD 2026 Workshop on Evaluation and Trustworthiness of Agentic AI, 2026. URL https://arxiv.org/abs/2607.24010. Accepted workshop paper.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Joseph J. Rocchio. Relevance feedback in information retrieval. In Gerard Salton (ed.), *The SMART Retrieval System: Experiments in Automatic Document Processing*, pp. 313–323. Prentice-Hall, 1971.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 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 (Volume 1: Long Papers)*, pp. 10014–10037. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

Shuai Wang, Haodong Chen, Yu Yin, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. Search, inspect, fetch: Exploiting structure-aware boolean retrieval for deep-search agents. *arXiv preprint arXiv:2608.02751*, 2026. URL https://arxiv.org/abs/2608.02751.

Zili Wei, Yilin Wang, Xiaocui Yang, Shi Feng, Weidong Bao, Daling Wang, Zihan Wang, and Yifei Zhang. CIRAG: Construction–integration retrieval and adaptive generation for multi-hop question answering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26181–26197. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1203. URL https://aclanthology.org/2026.acl-long.1203/.

Mingyan Wu, Han Yang, Omer Ben-Porat, and Yftah Ziser. When knowledge is not free: Cost-aware evidence selection in retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, 2026. URL https://arxiv.org/abs/2606.02245. To appear.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general chinese embeddings. *arXiv preprint arXiv:2309.07597*, 2023. URL https://arxiv.org/abs/2309.07597.

Siheng Xiong, Oguzhan Gungordu, James C. Kerce, and Faramarz Fekri. Adaptive information control for search-augmented LLM reasoning. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026. URL https://arxiv.org/abs/2602.01672. To appear; an earlier version appeared at the First Workshop on Scaling Post-Training for LLMs (SPOT), ICLR 2026.

Xinlong Xu and Yoshua Y. Li. EviReform: Evidence-guided query reformulation for multi-hop graph retrieval. *arXiv preprint arXiv:2608.13006*, 2026. URL https://arxiv.org/abs/2608.13006.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Sanghyun Yoo and Doowon Jeong. LLM-based keyword augmentation for title-driven evidence selection: A practical approach. *Journal of Forensic Sciences*, 71(3):1211–1223, 2026. doi: 10.1111/1556-4029.70311. URL https://doi.org/10.1111/1556-4029.70311.

JungMin Yun and YoungBin Kim. IterCOMP: Reasoning-aware adaptive prompt compression for multi-hop question answering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33827–33840. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1559. URL https://aclanthology.org/2026.acl-long.1559/.

Hansi Zeng, Liam Collins, Bhuvesh Kumar, Neil Shah, and Hamed Zamani. CoSearch: Joint training of reasoning and document ranking via reinforcement learning for agentic search. *arXiv preprint arXiv:2604.17555*, 2026. URL https://arxiv.org/abs/2604.17555.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LkzuPorQ5L.

Chenyu Zhou, Yaqi Wu, Xiaolei Guo, Jiaqi Huang, Xianfa Zhang, Junxu Zhang, Zhuo Yu, Zhubo Shi, Jianghao Lin, and Dongdong Ge. DynaKRAG: A unified framework for learnable evidence control in multi-hop retrieval-augmented generation. *arXiv preprint arXiv:2607.06507*, 2026. URL https://arxiv.org/abs/2607.06507.

Shengyao Zhuang, Yuansheng Ni, Hengxin Fun, Jimmy Lin, and Xueguang Ma. Towards retrieving interaction spaces for agentic search. *arXiv preprint arXiv:2606.06880*, 2026. URL https://arxiv.org/abs/2606.06880.

Ziyuan Zhuang et al. EfficientRAG: Efficient retriever for multi-hop question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3392–3411. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.199. URL https://aclanthology.org/2024.emnlp-main.199/.

A Theory and Proofs

A.1 Finite-horizon coverage

The probability model is the joint task distribution D together with the policy's independent randomization. The visible history $\mathcal{H}_t$ contains the public task information and ordered acquired excerpts. All oracle quantities use the same observation map and document reference $E(X)$ as the stated comparator. We assume a finite catalog with at least K available acquisitions and a nonempty support set. Normalized coverage is zero initially and is pointwise monotone, which implies adaptive monotonicity. The proof instantiates a standard approximate-greedy progress recursion, related to broader guarantees beyond adaptive submodularity (Fujii & Sakaue, 2019), for the support-coverage utility and actual acquisition actions used here.

Proof of Theorem 1. Let $V_t = \mathrm{OPT} - F_t$. By the tower property and the definition of δ_t,

$$\begin{aligned} F_{t+1} - F_t &= \mathbb{E}[\Delta(e_{t+1} \mid \mathcal{H}_t)] \\ &= \mathbb{E}\left[\max_{e \in E \setminus S_t} \Delta(e \mid \mathcal{H}_t)\right] - \delta_t \geq \frac{\gamma}{K} V_t - \delta_t. \end{aligned} \tag{9}$$

Consequently $V_{t+1} \leq (1 - \gamma/K)V_t + \delta_t$. Since $F_0 = 0$, iterating gives

$$V_K \leq \left(1 - \frac{\gamma}{K}\right)^K \mathrm{OPT} + \sum_{t=0}^{K-1} \left(1 - \frac{\gamma}{K}\right)^{K-1-t} \delta_t. \tag{10}$$

Substitute $V_K = \mathrm{OPT} - F_K$ to obtain Equation (8). Under the additional conditions $\gamma = 1$ and $\delta_t = 0$ at every step, the coefficient is at least $1 - 1/e$. □

The expected progress condition permits a mixture of completed and incomplete questions. Its right-hand side is a population difference, so it is compared with a population-averaged conditional gain. Neither γ nor OPT is an input to either implemented policy.

A.2 Realized gaps and conditional selection regret

Define the realized coverage gain

$$g_t(e, \Omega) = f_{\mathrm{cov}}(S_t \cup \{e\}, \Omega) - f_{\mathrm{cov}}(S_t, \Omega) = \frac{\mathbf{1}\{e \in G(\Omega) \setminus S_t\}}{|G(\Omega)|}. \tag{11}$$

Held-out labels determine the realized full-information gap

$$r_t = \max_{e \in E \setminus S_t} g_t(e, \Omega) - g_t(e_{t+1}, \Omega), \qquad R_t = \mathbb{E}[r_t]. \tag{12}$$

The maximum is evaluated after selection and does not enter a decision. Since the action is measurable from $\mathcal{H}_t$, conditional expectation yields

$$\begin{aligned} \delta_t &= \mathbb{E}\Big[\max_e \mathbb{E}[g_t(e, \Omega) \mid \mathcal{H}_t] - \mathbb{E}[g_t(e_{t+1}, \Omega) \mid \mathcal{H}_t]\Big] \\ &\leq \mathbb{E}\Big[\mathbb{E}[\max_e g_t(e, \Omega) - g_t(e_{t+1}, \Omega) \mid \mathcal{H}_t]\Big] = R_t. \end{aligned} \tag{13}$$

Both maxima range over $E \setminus S_t$. Thus held-out sample means estimate R_t, an upper bound on conditional regret; they do not directly estimate the conditional oracle or establish the progress coefficient. Substituting R_t for δ_t in Equation (8) gives a weaker bound under the same progress assumption. A bootstrap interval for a sample mean is a sampling summary, not a uniform bound over all histories.

If the policy scores only $A_t \subseteq E \setminus S_t$, its actual action remains in Equation (12). Let $m_t(A) = \max_{e \in A \setminus S_t} g_t(e, \Omega)$, with an empty maximum equal to zero. The exact realized decomposition is

$$r_t = \underbrace{m_t(E) - m_t(A_t)}_{\text{candidate omission}} + \underbrace{m_t(A_t) - g_t(e_{t+1}, \Omega)}_{\text{within-set selection}}. \tag{14}$$

The denominator remains the full $|G(\Omega)|$. A cached union of retrieved documents and a question-specific pool have distinct reference sets. The current prefix appendix reports answer trajectories and does not provide a full-catalog regret estimate; no empirical R_t, γ, or uniform calibration error is substituted into the theorem.

A.3 Optional score-calibration lemma

Lemma 1. *Fix $\mathcal{H}_t$. Suppose a score $\widehat{d}_t$ satisfies $|\widehat{d}_t(e) - \Delta(e \mid \mathcal{H}_t)| \leq \epsilon_t$ for every $e \in E \setminus S_t$, and the policy chooses $\hat{e}$ maximizing that score over this entire set. Then $\max_e \Delta(e \mid \mathcal{H}_t) - \Delta(\hat{e} \mid \mathcal{H}_t) \leq 2\epsilon_t$.*

Proof. For e^* maximizing the true conditional gain,

$$\Delta(\hat{e} \mid \mathcal{H}_t) \geq \widehat{d}_t(\hat{e}) - \epsilon_t \geq \widehat{d}_t(e^*) - \epsilon_t \geq \Delta(e^* \mid \mathcal{H}_t) - 2\epsilon_t.$$

□

If the same ϵ_t bounds every reachable history, then $\delta_t \leq 2\epsilon_t$; a history-dependent error gives $2\mathbb{E}[\epsilon_t(\mathcal{H}_t)]$. For a restricted candidate set, the lemma bounds only within-set regret and candidate omission must also be included. Raw BM25, support-document classifier scores, and s_t^{res} are selection scores; the experiments use actual selected actions to diagnose their realized gaps.

A.4 BM25 RELEVANCE-FEEDBACK COROLLARY

Let $a_{t+1}^{\mathrm{sparse}}$ be the document actually chosen by BM25-RF. Define

$$\delta_t^{\mathrm{sparse}} = \mathbb{E}\left[\max_{e \in E \setminus S_t} \Delta(e \mid \mathcal{H}_t) - \Delta(a_{t+1}^{\mathrm{sparse}} \mid \mathcal{H}_t)\right]. \tag{15}$$

Under the expected progress condition for this policy, Equation (8) holds with F_K^{sparse} and $\delta_t^{\mathrm{sparse}}$. The same substitution applies to the actual Residual-CMG actions. This statement requires no utility-scale calibration of either ranking function; the progress condition remains a separate premise.

A.5 BRIDGE INFORMATION AND COMPLETION

Adaptive submodularity compares the same candidate at nested histories (Golovin & Krause, 2011). Consider two equally likely hidden states with support sets $\{A, B_x\}$ and $\{A, B_y\}$. Acquiring A reveals which state holds. For normalized coverage,

$$\Delta(B_x \mid X) = \tfrac{1}{4} < \tfrac{1}{2} = \Delta(B_x \mid X, A = x). \tag{16}$$

Realized coverage remains modular in each state, while the observation increases the expected value of B_x. This construction illustrates why a fixed-state coverage property does not imply adaptive diminishing returns. It does not assert that the expected progress condition holds for every task.

Clue-first tasks can make the progress coefficient small. Let one supporting document be drawn uniformly from $B_1, \ldots, B_N$, and let a non-support clue document A reveal which B_i is correct. With $K = 2$, a policy that first acquires A and then the revealed B_i attains $\mathrm{OPT} = 1$. At the empty history, however, A has zero one-step coverage gain and every B_i has expected gain $1/N$. Thus Equation (7) at $t = 0$ requires

$$\frac{1}{N} \geq \frac{\gamma}{2}, \qquad \text{hence} \qquad \gamma \leq \frac{2}{N}. \tag{17}$$

The theorem remains valid, while its coefficient can become weak precisely when an action is valuable for the information it reveals about a later acquisition. This example fixes the theorem's role: it accounts for performance once the expected-progress premise and selection gaps are specified; it does not establish those quantities for BM25-RF or Residual-CMG.

Completion is a different utility: if both documents are needed, every singleton has $f_{\mathrm{all}} = 0$, but acquiring the pair gives $f_{\mathrm{all}} = 1$. At the empty history, every one-step completion gain is zero although the two-step optimum is one, so a positive progress coefficient need not exist for completion. The theorem uses f_{cov}. In the empirical analysis, successive selected documents, score changes, and answer jumps are reported separately; they do not alone test the same-candidate conditional expectation required by adaptive submodularity.

A.6 THEORY–IMPLEMENTATION CORRESPONDENCE

Table 6: Probability, decisions and measured outcomes in the coverage analysis.

Object	Definition or experimental counterpart
$(X, \Omega) \sim D$	Public question/catalog and hidden bodies/support relations; task distribution used by the analysis
$O_q(e, \Omega) = \Phi_q(e)$	Deterministic excerpt returned after acquisition; support labels remain hidden
$\mathcal{H}_t$	Public context and ordered observations, with independent policy randomization if used
f_{cov}, F_t	Normalized support recall and its task-and-policy expectation
$s_0, f_\theta, s_t^{\mathrm{res}}$	Title prior, train-only support-document classifier score, and combined selection score
λ	Shared preset residual scale, fixed to 0.1 before held-out evaluation
δ_t, R_t	Conditional oracle gap and expected realized full-information gap, with $\delta_t \le R_t$
γ	Common coefficient in the expected progress assumption; no fitted numerical value
K	Acquisition count; the unified retrieval and answer evaluation uses 6
$g(q, \psi_K)$	Downstream answer model, measured by answer and support metrics separately from the theorem

The analysis uses unit-cost acquisitions. Token-weighted acquisition would require an explicitly different action budget and selection rule.

B ANSWER QUALITY ACROSS ACQUISITION PREFIXES

The prefix analysis uses the same deterministic 300-question subset from each dataset and the same saved acquisition order as the endpoint evaluation. Qwen3-4B-Instruct-2507 answers at $t \in \{0, 2, 4, 6\}$ from exactly the excerpts available after that many acquisitions. At $t = 0$, all methods receive the same empty evidence state, so their within-dataset values coincide.

Table 7: Mean answer F1 across acquisition prefixes. Each cell lists $t = 0/2/4/6$ for the same 300 questions.

Method	2WikiMultiHopQA	MuSiQue	HotpotQA FullWiki
Static-BM25	.0133/.3838/.4170/.4136	.0020/.0619/.0566/.0698	.0606/.2849/.2932/.2950
BM25-RF	.0133/.4488/.4757/.4850	.0020/.1111/.1149/.1175	.0606/.2696/.2780/.2836
Direct-CMG	.0133/.2507/.2948/.3158	.0020/.0682/.0769/.0805	.0606/.2011/.2032/.2227
Residual-CMG	.0133/.4193/.4541/.4728	.0020/.0642/.0685/.0814	.0606/.2933/.3026/.3222
Static-BGE	.0133/.3758/.4202/.4091	.0020/.0878/.1234/.1258	.0606/.3005/.3297/.3440
Dense-Rocchio	.0133/.2853/.3528/.3592	.0020/.0825/.1097/.1217	.0606/.2551/.2813/.2779

The curves distinguish acquisition utility from answer behavior. Support coverage can only stay constant or increase when another document is acquired, yet answer F1 can fall because the fixed answer model receives a longer and partly distracting evidence set. Two of the six methods on 2Wiki and one on HotpotQA attain their highest listed prefix F1 before $t = 6$; none does so on MuSiQue. These diagnostics motivate reporting support acquisition and downstream answers as separate endpoints and do not alter the predeclared $K = 6$ comparison.

C INFORMATION-ACCESS PROTOCOL

C.1 DECISION-TIME PERMISSIONS

Immediately before acquisition $t + 1$, every policy may read the question, every dataset-provided candidate title, the shared excerpt cap $L = 256$ under the character-based token estimate, document IDs for identity and tie-breaking, and the exact excerpts acquired during the first t steps. Document

IDs contribute no semantic features. Unopened bodies, candidate-specific body embeddings or sparse indexes, answer aliases, support labels, and evaluation-derived parameters are inaccessible. Title embeddings are permitted because they are computed from exposed titles. Corpus-level body IDF weights are used only inside the environment-side extraction channel after a document has been selected; they never score or represent an unopened candidate.

After a policy selects document e, the environment-side channel applies the same deterministic extractive sentence selector H_q to the original body $B(e)$. Implemented by the fixed scoring-and-packing rule below, it operates directly on original body sentences. Only the returned excerpt $\Phi_q(e)$ crosses to the policy. Selection-side length uses

$$\widehat{c}(x) = \left\lceil \frac{n_{\mathrm{ASCII}}(x)}{4} + \frac{n_{\mathrm{nonASCII}}(x)}{1.6} \right\rceil, \tag{18}$$

and $c(s) = \widehat{c}(s)$. For sentence s, the extraction score is

$$a(s,q) = \frac{\sum_{w \in T(q) \cap T(s)} \mathrm{IDF}(w)}{\max\{1, c(s)^{0.2}\}}. \tag{19}$$

Here $\mathrm{IDF}(w) = \log\left(1 + \frac{M - \mathrm{df}(w) + 0.5}{\mathrm{df}(w) + 0.5}\right)$. For 2Wiki and MuSiQue, M and df are computed over the fixed public catalog bodies. For HotpotQA, they are computed only from context documents attached to 1,000 model-fitting tasks and 500 disjoint regularization-selection tasks, all drawn from the official training split; the FullWiki evaluation bodies are not scanned to build this table. This shared aggregate table is used only after a document has been selected and does not provide candidate-specific body features. Tokenization lowercases ASCII alphanumeric spans, removes a fixed stopword list and one-character tokens, and uses unique question terms. Sentences are considered in decreasing score order, with original position breaking ties, and retained when they fit within a cap of 256 character-based token estimates. Selection stops when fewer than eight character-based token estimates remain. Retained sentences are returned in their original document order; if no complete sentence fits, the highest-ranked sentence is truncated to the cap. The resulting $\Phi_q(e)$ is the acquired excerpt supplied to later selection steps and the answer model.

Catalog information is excluded from the K acquisition steps and has nonzero input volume. Acquired-excerpt and serialized-input quantities use $\widehat{c}$: the latter additionally include identifiers and formatting overhead. Answer-model prompt, completion, and cache quantities use the tokenizer usage returned with each response where available; the units remain model-specific and are kept separate from the character-based token estimate. The answer prompt includes the question, selected message identifiers, titles, acquired excerpts, and output instruction, while the full catalog is not transmitted to the answer model. This ledger measures selection-side communication and answer-model token usage, rather than complete RAG-system ownership cost.

The fixed title catalog and its title-only indexes are loaded once and reused across questions. We therefore measure incremental online communication: excerpt content returned by six acquisitions and the serialized payload after adding titles, identifiers, and formatting. Catalog document counts are reported in Table 2; a uniform serialized catalog volume, transfer time, and index-construction cost were not measured across all three corpora and are excluded from the per-question totals below.

Table 8: Mean online selection communication at $K = 6$. Each cell is the character-based token estimate for acquired excerpts / serialized payload per question; lower is cheaper for an otherwise equal outcome.

Method	2WikiMultiHopQA	MuSiQue	HotpotQA FullWiki
Static-BM25	580.0 / 694.4	712.5 / 823.9	511.1 / 613.5
BM25-RF	673.8 / 791.8	742.3 / 861.9	519.5 / 632.5
Direct-CMG	492.0 / 597.4	765.3 / 871.5	536.7 / 629.1
Residual-CMG	548.1 / 660.0	699.1 / 811.5	522.8 / 620.7
Static-BGE	586.2 / 702.2	770.6 / 884.7	494.3 / 598.5
Dense-Rocchio	586.8 / 702.7	772.7 / 887.2	483.6 / 587.3

These units quantify the online information delivered to the selection policy. They do not include hardware ownership, the one-time catalog preload, or local search computation, and therefore are not a cold-start total-system cost.

C.2 Exact answer-model instruction

The system message is:

> You are an evidence-only multi-hop question answering aggregator. Use only the supplied evidence messages. If they cannot support an answer, return UNKNOWN. Return one JSON object with answer as a string and supporting_message_ids as an array of supplied MESSAGE_ID strings.

The user message has the following deterministic form:

> Question: {question}
> Evidence messages:
> MESSAGE_ID: {id}
> TITLE: {title}
> 0: {sentence 0}
> ...
> Return JSON now.

Messages retain acquisition order. Gold answers and support labels are absent. For answer F1, the scorer lowercases, removes punctuation and English articles, normalizes whitespace, computes multiset token-overlap precision and recall, takes their harmonic mean, maximizes over public reference answers or aliases, and macro-averages over questions. Retrieval AP, recall, and all-support coverage compare the document IDs actually acquired by the policy with public support-document IDs. The answer model's predicted message IDs are retained in its JSON output but are not used for those retrieval tables.

C.3 Dataset construction

We construct all three catalogs deterministically from public benchmark fields. For 2Wiki, normalized titles from all eligible development contexts define shared documents; duplicate sentences are removed while retaining their first occurrence. This yields 54,943 title documents searched by every query. A prespecified 40-question development subset is excluded from evaluation, leaving 12,536 questions. For MuSiQue, deterministic eligibility rules retain 2,401 answerable development questions and 21,099 distinct normalized title–paragraph documents. HotpotQA uses the public FullWiki corpus with 5,233,329 catalog documents; 7,309 of 7,405 validation questions map their complete supporting-title set to this catalog. Repeated titles retain distinct document identities for acquisition and deterministic tie-breaking.

Every method receives the complete title catalog and the extractor in Section C. Unopened bodies and body-derived representations remain inaccessible. Retrieval uses the full mapped evaluation set at $K = 6$. Each dataset's answer evaluation uses a 1,000-question subset selected by the same fixed hash rule before any answer is generated. A 300-question subset of that cohort supplies the $t = 0, \ldots, 6$ prefix analysis.

C.4 Policy details

Static-BM25 and BM25-RF rank the complete title index exactly. Dense-Static and Dense-Rocchio use the complete fixed title-vector index. On every dataset, Direct-CMG, Residual-CMG, and Prior+Full-CMG maximize their scores over the same candidate set A_t constructed from title retrieval: the union of the top 256 original-question–title BM25 matches and top 256 acquired-excerpt–title BM25 matches, after removing opened documents. The analysis compares the selected action with the complete catalog, so candidate-set omission remains part of the selection gap.

The sparse title index uses BM25 with $k_1 = 1.5$, $b = .75$, and $\mathrm{IDF}(w) = \log(1 + (M - \mathrm{df}(w) + .5)/(\mathrm{df}(w) + .5))$. Queries are sets of the tokens defined above, so a repeated query token contributes

once. The normalized prior divides each raw question–title score by the maximum raw score in its scoring range and returns zero when that maximum is zero. At evaluation the range is the complete title index; training features use each task's local training candidate set.

Table 9: Compared acquisition methods and their scoring rules.

Policy	Scoring rule
Residual-CMG	title-BM25 prior plus 0.1 times the acquired-state classifier score minus its empty-state score
Static-BM25	BM25 between the original question and candidate title, ranked once
Direct-CMG	support-document classifier score from the same train-only logistic model used by Residual-CMG
BM25-RF	BM25 relevance feedback using the original question concatenated with acquired excerpts
Static-BGE	cosine similarity of BGE question and title embeddings with fixed model weights
Dense-Rocchio	cosine similarity after $q_t = \mathrm{norm}(q_0 + 0.75\,\mathrm{mean}(r_1, \ldots, r_t))$ in the same BGE space

C.5 Residual-CMG training construction

For 2Wiki and MuSiQue, model fitting uses 12,000 tasks from the official training split and a disjoint 3,000-task training-split validation subset. The HotpotQA model uses 1,000 official training tasks for fitting and 500 disjoint training tasks for regularization selection. HotpotQA training features are constructed only over the documents provided with each training task; evaluation still uses the complete FullWiki title catalog. At each constructed state, an unselected training candidate is positive exactly when it belongs to the remaining support set. States include the empty observation, incomplete ordered support prefixes, and prefixes formed by static-BM25 choices; completed support states are omitted.

The model is standardized, class-balanced, L2-regularized logistic regression. Its 11 inputs are normalized static BM25; query–title phrase match and IDF overlap; acquired-excerpt–title phrase match and IDF overlap; maximum single-excerpt title overlap; log one plus title mentions; static-score interactions with acquired overlap, acquired phrase match, and step fraction; and acquired-overlap interaction with step fraction. Step fraction is $t/10$. Feature means and scales are fitted on the model-fitting rows. Regularization $C \in \{.01, .1, 1, 10\}$ is selected by mean state-level AP on the disjoint training-split validation subset; the selected values are .1, 10, and .1 for 2Wiki, MuSiQue, and HotpotQA. The residual scale is the shared preset $\lambda = 0.1$ fixed before held-out evaluation. Evaluation answers, support labels, and metrics are absent from fitting and model selection.

BGE is `BAAI/bge-base-en-v1.5` at revision `a5beb1e3e68b9ab74eb54cfd186867f64f240e1a`; the archived safetensors SHA-256 is `c7c1988aae201f80cf91a5dbbd5866409503b89dcaba877ca6dba7dd0a5167d7`. We use its official query prefix, CLS pooling, L2 normalization, and maximum sequence length 512. Rocchio parameters are $\alpha = 1$, $\beta = 0.75$, with no negative-feedback term and no development tuning.

D Primary Paired Comparisons and Structural Ablation

D.1 Matched comparisons

Table 10: Primary matched effects at $K = 6$. Each row reports Δ support AP and Δ answer F1 with paired 95% bootstrap intervals. Differences are the method in the row minus its control.

Comparison	Δ support AP [95% CI]	Δ answer F1 [95% CI]
2WikiMultiHopQA		
BM25-RF − Static-BM25	$+.1290[.1240, .1339]$	$+.0674[.0377, .0961]$
Residual-CMG − Direct-CMG	$+.1594[.1544, .1645]$	$+.1393[.1078, .1705]$
Dense-Rocchio − Static-BGE	$-.0661[-.0685, -.0636]$	$-.0444[-.0649, -.0236]$
MuSiQue		
BM25-RF − Static-BM25	$+.0381[.0306, .0458]$	$+.0488[.0313, .0661]$
Residual-CMG − Direct-CMG	$+.0386[.0309, .0463]$	$+.0035[-.0117, .0186]$
Dense-Rocchio − Static-BGE	$-.0213[-.0259, -.0167]$	$+.0014[-.0088, .0116]$
HotpotQA FullWiki		
BM25-RF − Static-BM25	$-.0544[-.0598, -.0492]$	$-.0275[-.0501, -.0062]$
Residual-CMG − Direct-CMG	$+.1173[.1122, .1224]$	$+.0894[.0696, .1099]$
Dense-Rocchio − Static-BGE	$-.0940[-.0980, -.0899]$	$-.0611[-.0791, -.0439]$

All support-AP effects in Table 10 remain significant after Holm correction ($p_{\text{Holm}} = .003$). For answer F1, the positive BM25-RF effects on 2Wiki and MuSiQue, the positive Residual-CMG effects on 2Wiki and HotpotQA, and the negative dense-feedback effects on 2Wiki and HotpotQA remain significant at the same corrected level. The negative HotpotQA BM25-RF effect has an unadjusted interval below zero and $p_{\text{Holm}} = .061$. Complete recall, all-support coverage, exact-match, cross-method reference, and corrected-p records accompany the released result table.

D.2 Static prior, full prediction, and state change

Prior+Full-CMG adds the acquired-state support-document classifier score to the static title-BM25 score, while Residual-CMG adds the acquired-state minus empty-state score to the same static score. Prior+Full-CMG, Residual-CMG, and Direct-CMG use the same support-document classifier and candidate set.

Table 11: Structural CMG ablation at $K = 6$. Each row reports support AP, recall, and all-support coverage on the full retrieval cohort.

Method	Support AP	Recall	All-support coverage
2WikiMultiHopQA			
Direct-CMG	.5225	.5443	.1713
Prior+Full-CMG	.6649	.6973	.3772
Residual-CMG	.6820	.7115	.4197
MuSiQue			
Direct-CMG	.2801	.2832	.0450
Prior+Full-CMG	.3185	.3675	.0583
Residual-CMG	.3187	.3671	.0579
HotpotQA FullWiki			
Direct-CMG	.3022	.3101	.0265
Prior+Full-CMG	.4147	.4567	.2070
Residual-CMG	.4195	.4788	.2285

Both prior-preserving variants substantially exceed Direct-CMG. Prior+Full-CMG accounts for about 89%, 99%, and 96% of the rounded Direct-to-Residual AP increment. Residual-CMG has higher AP than Prior+Full-CMG by .0171, .0002, and .0049. The subtraction ablation is reported without a significance claim because paired intervals and answer-model outputs were not computed

for this comparison. The ablation supports the importance of the explicit static title-BM25 score under the present learned model.

E Reproducibility Details

E.1 Deterministic selection and response validation

Candidate tokens are sorted lexicographically before floating-point accumulation, and document identifiers plus titles provide deterministic tie-breaking. Independent executions produce identical selected-document sequences and answer prompts.

For a byte-identical serialized request, the corresponding previously validated response may be reused. Malformed or incomplete responses are excluded. The released material specifies the selection rules, eligibility criteria, answer prompt, output validation, metric definitions, and resource accounting.

F Resource and Cost Accounting

Table 12: Qwen inference workload. Token counts use the model service's tokenizer usage and include every executed attempt. "Unique" counts distinct answer requests; "Attempts" also counts retries.

Dataset	Retrieval n	Answer rows	Unique	Attempts	Prompt tok.	Completion tok.
2WikiMultiHopQA	12,536	16,800	12,724	12,724	8,467,339	380,677
MuSiQue	2,401	16,800	12,522	12,785	12,381,147	577,235
HotpotQA FullWiki	7,309	16,800	13,396	13,396	7,796,139	401,845
Total	22,246	50,400	38,642	38,905	28,644,625	1,359,757

The workload uses 30,004,382 actual Qwen tokens, and every required request has a valid answer. MuSiQue's totals include 263 retries and one failed-attempt row; all 12,522 unique requests succeeded. Qwen3-4B-Instruct-2507 runs locally through vLLM on one GPU, with zero external-provider billing. Hardware ownership and electricity are not monetized.

Retrieval wall times are 4,069.3 seconds for 2Wiki, 554.4 seconds for MuSiQue, and 3,340.7 seconds for HotpotQA. HotpotQA sparse retrieval uses 32 CPU workers with 229.0 GiB summed peak RSS; dense retrieval peaks at 53.4 GiB. Recorded answer-generation time is 1,334.2 seconds on 2Wiki and 1,362.9 seconds on HotpotQA. MuSiQue's 1,830.6 seconds covers its initial generation phase and excludes the 263 retries, so it is a lower bound. Offline validation time is excluded.

Table 8 reports selection communication separately because its estimator and the answer-model tokenizer use different units. The title catalog and title-only indexes are preloaded. One-time construction, storage, serialization, and transfer fall outside the online per-query totals and lack uniform three-corpus measurements.