

Online Resource 1

Cost-Aware Gated Retrieval: Decoupling Retrieval Initiation from Evidence-Conditioned Escalation

Zida Hao and Yubo Li

School of Software, Xinjiang University, Urumqi, China

Corresponding author: Zida Hao (20242501431@stu.xju.edu.cn)

Journal of Intelligent Information Systems

This resource retains extended studies and implementation details from the main article. Section S1 contains the full development studies; S2–S7 provide operating parameters, full comparisons, and reproducibility details. Table and figure numbers in this resource have an S prefix. The main article’s bibliography is reproduced at the end for consistent reference numbering. Figure S1 summarizes the corresponding measurements.

S1 System Extensions and Ablation Studies

S1.1 Cache continuation and evidence quality

KV-Bridge reduces the cost of generating an additional query by retaining the preliminary answer’s cache. Across 200 paired development questions, mean call time falls from 724.52 to 267.98 ms, a 63.01% reduction and a 2.70-fold speedup. The paired difference is -456.54 ms with a descriptive 95% bootstrap interval of $[-484.43, -428.85]$ ms. These intervals use 10,000 question resamples with seed 42 and are development analyses.

Table S1 summarizes the paired query-generation measurements.

Table S1: KV-Bridge development calls on 200 paired questions; end-to-end call times include the appended suffix, decoding, and host work.

Measure	Query rewriting	KV-Bridge
Mean call time (ms)	724.521	267.980
P95 call time (ms)	1093.41	315.39
Mean new input tokens	669.420	19.005
Mean output tokens	13.235	6.855

The long rewriting prompt processes 669.420 new input tokens on average, compared with 19.005 for KV-Bridge. Output lengths are 13.235 and 6.855 tokens, respectively, so the complete-call comparison combines cache reuse with a shorter generation task. A stricter 20-question control replays the identical complete prefix and obtains identical outputs, with cache continuation saving 190.63 ms (interval $[-218.52, -164.77]$ ms). This control directly isolates the benefit of reusing the processed prefix.

Query-generation tails show a similar component-level opportunity: development P95 falls from 1,093.41 ms for rewriting to 315.39 ms for KV-Bridge, a 71.2% reduction. Lower query-generation cost is particularly relevant to configurations that escalate frequently. The full-pipeline timing results are reported separately from this paired development comparison.

Figure S2 summarizes the corresponding measurements.

The complete action matrix evaluates stop, answer expansion, KV-Bridge, and full rewriting across every development question. Answer expansion constructs the second query from existing question and answer text, avoiding a separate rewriting generation. Previously computed answer-expansion outcomes are reused when their prompt, document order, and tokens match, with timing obtained through component replay.

Table S2 reports the complete development action comparison.

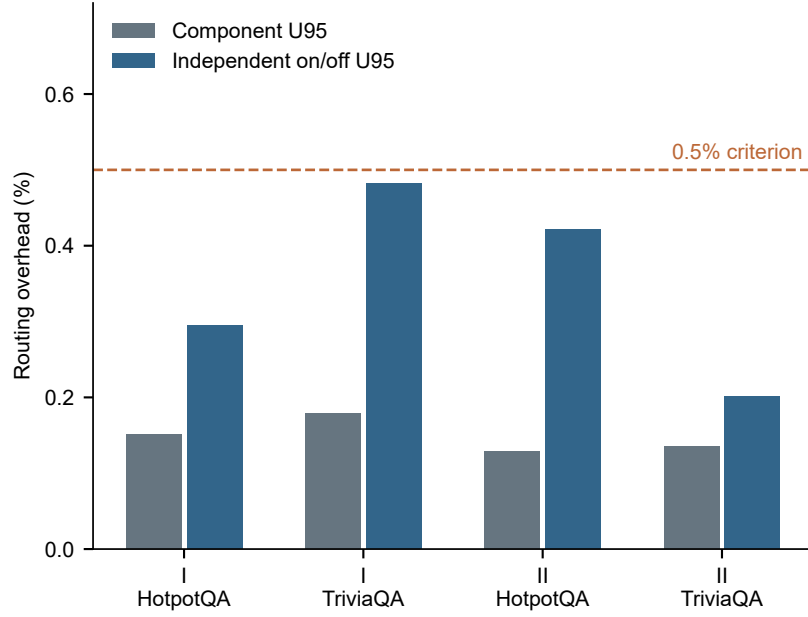


Fig. S1 Routing overhead upper bounds. Bars show one-sided 95% bootstrap upper bounds for component and independent measurements; the dashed line is the specified 0.5% criterion. Each dataset contributes 500 questions per cohort

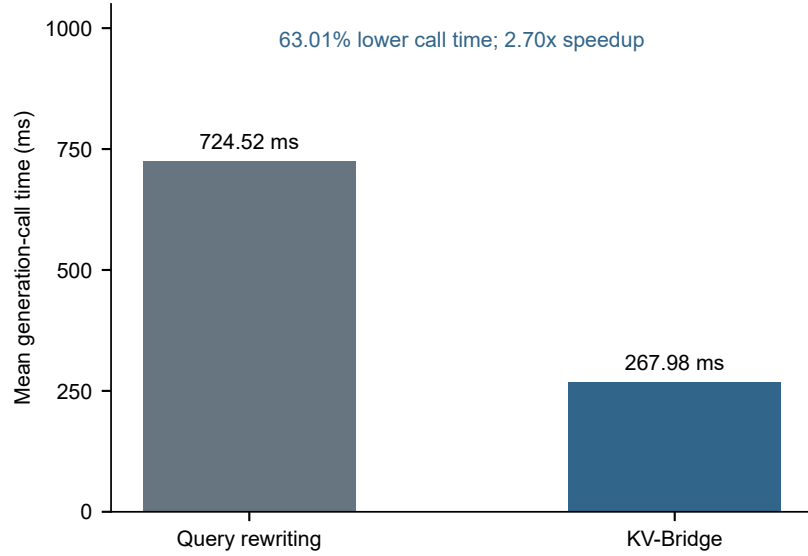


Fig. S2 Development query-generation calls across 200 paired questions. Bars show mean latency for the complete rewriting and bridge-continuation tasks; paired difference intervals are reported in the text

Table S2: Complete four-action development matrix, forced on all 200 questions. Gain/tie/harm is relative to the first retrieved answer; these are potential action outcomes, not a routed method.

Action	H EM (%)	H F1 (%)	T EM (%)	T F1 (%)	H gain/tie/harm
Stop after first retrieval	18.00	29.32	41.00	45.48	0/100/0
Answer expansion	16.00	26.83	41.00	44.84	4/90/6
KV-Bridge	16.00	27.46	42.00	45.88	5/88/7
Query rewriting	21.00	32.33	45.00	48.09	11/82/7

KV-Bridge improves nine questions, ties 179, and harms 12 relative to the first retrieved answer; full rewriting produces 17 improvements, 172 ties, and 11 harms. In 50 of 200 bridge outputs, all extracted terms already occur in the question. Eighty-eight outputs reach the eight-token cap, including 57 without an end-of-sequence token. These patterns associate short continuation with repeated query terms and limited opportunities to name a missing entity.

The complete development escalation increment is 887.55 ms for KV-Bridge and 1,300.07 ms for full rewriting, including retrieval and final-answer generation. The 267.98 ms bridge call forms one part of this increment; its second retrieval takes 169.25 ms on average, versus 74.42 ms for the rewritten query. Cache continuation therefore reduces repeated context processing while evidence selection remains a separate determinant of final quality and cost.

S1.2 Stage 0: model-state signals versus question structure

We compare three feature families for predicting positive Δ_{01} on the 200-question development split. The scalar Pilot K=5 feature is u . Thirty question-structure features encode bounded length, capitalization spans, punctuation, relative and comparative terms, prepositions, digits, and initial interrogative words. The 61-dimensional fusion concatenates those 30 features, u , and 30 interactions with u . No reference or dataset identifier is supplied to these heads. Each training fold contains 160 questions, with feature standardization restricted to the fold.

Table S3 compares the first-stage feature families.

Table S3: Stage 0 exploratory feature comparison on the Development Calibration Split (N=200, five folds). Each family was selected within a finite development search.

Features	Dimensions	OOF AUC	OOF MSE	OOF F1 (%)	Training F1 (%)
Pilot K=5	1	0.6615	0.1707	38.58	38.78
Question structure	30	0.4985	0.1884	36.10	41.03
Fusion	61	0.6287	0.2397	36.91	43.21

The scalar prefix signal achieves the highest observed OOF AUC, 0.6615, compared with 0.4985 for structure and 0.6287 for fusion. The fused policy’s overall F1 increases from 36.91% OOF to 43.21% on training questions, a 6.29-point gap consistent with fitting and selection optimism. The prefix-only family’s OOF F1 is 38.58%, supporting the compact signal in this development comparison.

Figure S3 summarizes the corresponding measurements.

The families were selected within a 189-configuration development search, and complete routed F1 also reflects family-specific escalation choices. AUC offers the more direct comparison of first-stage ranking and favors retaining model-state information in the tested design. Local fitting and replay used the existing model outputs and completed in 5.50 s.

S1.3 Stage 1: compact signals and evidence text

The `gap_base` escalation head uses the preliminary answer’s mean logit gap, retrieval-score variance, and answer length. The eight-dimensional `gap_text` variant adds five bounded text features: answer-document overlap, question coverage, uncovered capitalized question terms, answer occurrence in the best document, and document-answer agreement. All features are observed after the first retrieved answer and before another search. The comparison uses existing five-fold predictions and the same cost weight and decision thresholds.

Table S4 compares the compact and evidence-text escalation heads.

Table S4: Stage 1 development ablation at common cost weight 0.00003 and utility threshold zero. This diagnostic uses different thresholds from the frozen operating points.

Features	Dimensions	OOF AUC	OOF RMSE	H F1 (%)	T F1 (%)	D/S/E
gap_base	3	0.6467	0.27154	30.49	46.58	62/93/45
gap_text	8	0.6741	0.27229	28.59	46.08	62/88/50

Adding evidence-text features raises positive-gain AUC from 0.6467 to 0.6741, with a paired difference interval of $[-0.0341, 0.0935]$. Gain RMSE increases slightly from 0.27154 to 0.27229. At common thresholds,

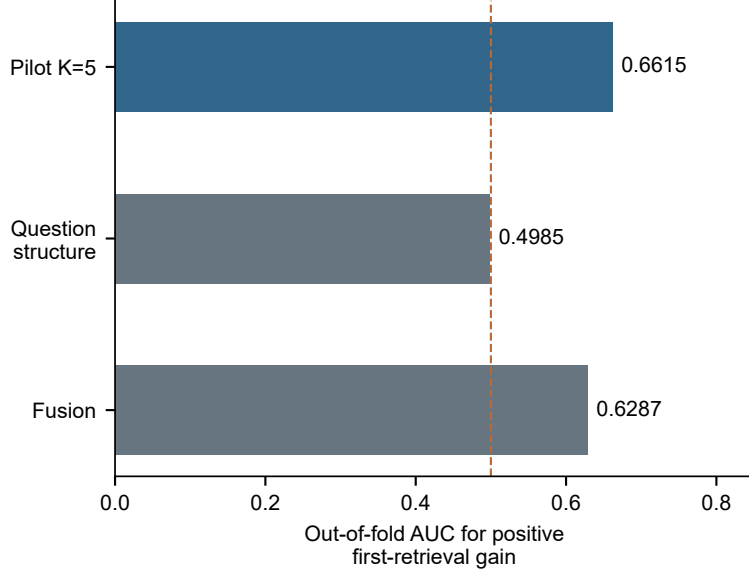


Fig. S3 Stage 0 feature comparison on 200 development questions. Bars show out-of-fold AUC for positive first-retrieval gain; the dashed line marks chance ranking

the larger head selects 50 rather than 45 escalations and lowers F1 on both datasets. This combination of discrimination and routed outcomes favors the compact head under the tested development budget.

S1.4 Matched quotas and prefix-length sensitivity

Random Gate assigns questions to saved potential-answer paths while preserving the observed direct/single/escalated counts. One diagnostic holds the global 590/319/91 quotas from Test Cohort I, and another holds the quotas within each dataset. A fixed random stream initialized with seed 42 generates 20,000 assignments, characterizing allocation variability under these constraints.

Table S5 reports the matched-quota assignment diagnostics.

Table S5: Test Cohort I matched-quota Random Gate: 20,000 assignments from seed 42. SD describes the random-policy distribution, not uncertainty from new test sampling.

Assignment	H F1 %, mean (SD)	T F1 %, mean (SD)
Global quotas	26.44 (1.08)	39.80 (1.03)
Within-dataset quotas	28.02 (1.08)	39.45 (0.94)
CAGR-R observed	31.60	42.80

CAGR-R’s observed F1 exceeds the random-assignment means under both quota definitions. Preserving dataset-specific quotas narrows the HotpotQA difference, showing the contribution of task allocation to the global comparison. The remaining gap supports selection value from the recorded signals at matched quotas, with random-policy standard deviations describing assignment variability.

We also truncate saved prefix observations to $K = 1, \dots, 5$ while holding the K=5 head and threshold fixed. For CAGR-R’s first gate, the retrieval fraction changes from 81.2% at K=1 to 41.0% at K=5. Overall first-stage F1 is 37.74% at K=1 and 36.85% at K=5; these are binary first-stage outcomes and exclude escalation. The change illustrates that an uncertainty threshold calibrated for one observation length need not preserve its operating point at another length.

Figure S4 summarizes the corresponding measurements.

The complete prefix table is provided in the appendix, reporting first-stage sensitivity to observation length under the fixed five-token heads and thresholds.

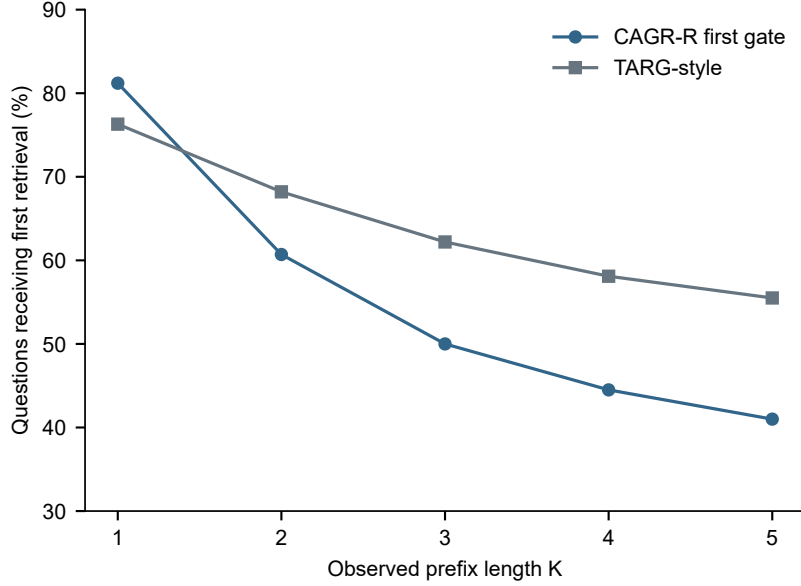


Fig. S4 Prefix truncation sensitivity in Test Cohort I. Curves show first-retrieval fractions from saved token-level signals, with the five-token heads and thresholds held fixed

S1.5 Novelty Stop and the retrieval-cost spectrum

Novelty Stop omits 11 final generations among CAGR-R’s 91 escalations in Test Cohort I and two among CAGR-F’s ten escalations in Test Cohort II. The 200-question forced-action matrix stops 30 full-rewriting paths and 32 KV-Bridge paths. All preceding query-generation and retrieval costs remain included, while deterministic unchanged prompts preserve the saved first answer.

A fixed-path sensitivity analysis isolates the contribution of retrieval cost in Cohort II. For mean latency L_A , observed retrieval time B_A , and retrieval rate r_A , we define $L_A(c) = L_A - B_A + r_A c$. Assigning cost c per retrieval preserves each method’s answers, query text, and routing decisions while varying the retrieval-time component.

Table S6 summarizes fixed-path retrieval-cost sensitivity.

Table S6: Fixed-path cost substitution on Test Cohort II, in milliseconds per question. These are algebraic scenarios with observed answers and branch counts held fixed, not new hardware measurements.

Cost per retrieval (ms)	Always-RAG	TARG-style	CAGR-R	CAGR-F
131	551.30	478.25	533.66	483.85
300	720.30	573.06	620.87	577.31
500	920.30	685.26	724.07	687.91
2500	2920.30	1807.26	1756.07	1793.91

The slope equals the observed retrieval rate: CAGR-F’s 0.553 is substantially below Always-RAG and close to TARG-style’s 0.561. Increasing retrieval cost consequently amplifies savings over indiscriminate retrieval, while similarly selective gates retain comparable sensitivity. Answer quality remains at the measured values throughout these fixed-path scenarios.

S2 Frozen operating points and algorithm

S2.1 Decision parameters

CAGR-R uses $\lambda_0 = 0$, $t_0 = 0.06643697151231595$, $\lambda_1 = 0.0002$, and $t_1 = -0.20252241348899253$. Its initiation gain and cost heads have ridge penalty 100; the escalation gain and cost penalties are 100 and 10. Development calibration selected the negative escalation threshold, producing the configuration’s more frequent escalation.

CAGR-F initiates when $u > 0.3324547106961593$. Its action gain heads use penalty 100 and accept the highest score above 0.08009160634576197. The final runtime score uses $\lambda_1 = 0$, with the gain threshold, hard token/retrieval limits, and Novelty Stop bounding additional work. Under a common-cost approximation, the threshold can express $t_1 + \lambda_1 \bar{\kappa}_{12}$, whose components are jointly represented by the directly calibrated value 0.0801. Development action costs vary, with increments of 887.55 ms for the bridge and 1,300.07 ms for rewriting, motivating the distinction between gain-threshold and explicit-cost parameterizations.

For TriviaQA, a preliminary mean gap of at least 11.237109375 suppresses escalation. Action ties follow the order of answer expansion, KV-Bridge, and query rewriting. The exported `operating_points.json` contains the original means, scales, and coefficients in the order $b, w_1, \dots, w_d$ for prediction $b + \sum_j w_j(x_j - \mu_j)/\sigma_j$.

The TARG-style deployment threshold is 0.322316956977953 in both test cohorts, fixed from the complete development split. Its realized retrieval fraction changes with the questions rather than being forced to remain constant. SUGAR-L-style uses semantic thresholds 0 and $\log(2)/2$, two answer samples, temperature 1, and feedback after each allowed retrieval. These adaptation details are part of the comparison and distinguish it from the published full system.

S2.2 Bounded inference procedure

```

Input: question q, fixed index D, generator M, remaining budget b
Decode at most five direct-answer tokens; retain the pending token and KV cache.
Compute Stage 0 features from the observed prefix and bounded question counts.
If the initiation rule rejects retrieval:
    Continue the same state to the answer limit and return the direct answer.
Charge the abandoned prefix; release its cache.
Retrieve with q and generate the first answer a1 from the evidence prompt.
Compute Stage 1 features from a1 and the first retrieval scores.
Select stop or an allowed escalation action using the frozen rule.
If stop is selected or the remaining budget is insufficient, return a1.
Construct one second query; charge any query-generation call.
Retrieve once more and merge documents in the fixed order.
If the ordered evidence prompt is unchanged, return a1 (Novelty Stop).
Generate and return the final answer under the remaining budget.

```

The controller permits at most two searches, with every performed bridge generation contributing its input tokens, output tokens, and latency even when the subsequent action stops. Complete path accounting includes all attempted computation within the method boundary.

S2.3 Signal storage and complexity

The native signal window uses $K = 5$ and vocabulary width 151,936, giving 1,519,360 bytes of FP16 storage. Additional buffers hold 320 bytes of partial reductions, 160 bytes of indices, 160 bytes of validity state, 320 bytes of clock scratch space, an eight-byte pending token, and a 40-byte host boundary array. Persistent allocation reduces repeated host allocation and binding overhead. The reduction remains $O(K|V|)$ in scanned logits, followed by constant-size linear prediction for the frozen feature dimensions. KV storage is the generator’s ordinary sequence-dependent cache and is not included in this signal-buffer count.

S3 Complete costs and paired contrasts

The following tables report dataset-specific and equal-weight costs, including latency quantiles, token counts, and branch counts. Direct/single/upgrade (D/S/E) describes the final route, and normal-path means cover the complete online method boundary. Preparation and technical repetitions are summarized separately in Appendix D.

Table S7 reports complete Cohort I costs and branch counts.

Table S8 reports complete Cohort II costs and branch counts.

The paired contrast tables report CAGR-R minus each comparator in Test Cohort I and CAGR-F minus each comparator in Test Cohort II. Marginal intervals retain every question and quantify uncertainty around the paired differences.

Table S9 reports the Cohort I paired contrasts.

Table S10 reports the Cohort II paired contrasts.

Table S7: Test Cohort I: complete method costs and branch counts. Tokens include all generation calls within a method.

Dataset	Method	Mean ms	P50 ms	P95 ms	Input tok.	Output tok.	Retr./q	D/S/E
HotpotQA	Direct	212.17	172.24	387.59	68.466	5.090	0.000	500/0/0
HotpotQA	Always-RAG	628.50	553.60	1060.45	622.268	6.374	1.000	0/500/0
HotpotQA	Always-Multihop	1955.90	1829.01	2935.94	2159.848	24.264	2.000	0/0/500
HotpotQA	Matched Binary	547.51	523.23	1188.77	408.186	8.246	0.530	235/265/0
HotpotQA	TARG-style	610.15	632.02	1197.83	468.316	8.940	0.628	186/314/0
HotpotQA	SUGAR-L-style	2646.32	2696.18	4733.91	2904.474	37.952	1.356	89/144/267
HotpotQA	CAGR-R	718.11	596.69	2341.57	601.878	10.660	0.656	235/202/63
HotpotQA	CAGR-R (answer expansion)	606.05	599.92	1457.37	469.280	8.622	0.656	235/202/63
TriviaQA	Direct	206.68	162.15	547.51	66.298	4.984	0.000	500/0/0
TriviaQA	Always-RAG	608.16	514.44	1294.19	653.558	6.600	1.000	0/500/0
TriviaQA	Always-Multihop	1876.39	1741.96	2961.96	2145.282	23.726	2.000	0/0/500
TriviaQA	Matched Binary	400.13	194.52	1195.12	253.368	7.138	0.290	355/145/0
TriviaQA	TARG-style	488.77	424.88	1225.30	374.768	7.900	0.482	259/241/0
TriviaQA	SUGAR-L-style	2274.12	2297.19	4386.22	2614.124	32.924	1.124	163/112/225
TriviaQA	CAGR-R	466.52	194.60	1776.74	339.388	8.008	0.346	355/117/28
TriviaQA	CAGR-R (answer expansion)	427.32	194.97	1341.42	284.666	7.350	0.346	355/117/28
Overall	Direct	209.43	169.10	427.92	67.382	5.037	0.000	1000/0/0
Overall	Always-RAG	618.33	531.49	1144.52	637.913	6.487	1.000	0/1000/0
Overall	Always-Multihop	1916.15	1782.68	2957.06	2152.565	23.995	2.000	0/0/1000
Overall	Matched Binary	473.82	237.64	1192.65	330.777	7.692	0.410	590/410/0
Overall	TARG-style	549.46	539.74	1223.31	421.542	8.420	0.555	445/555/0
Overall	SUGAR-L-style	2460.22	2508.03	4531.01	2759.299	35.438	1.240	252/256/492
Overall	CAGR-R	592.31	242.43	2094.98	470.633	9.334	0.501	590/319/91
Overall	CAGR-R (answer expansion)	516.69	238.83	1416.41	376.973	7.986	0.501	590/319/91

Table S8: Test Cohort II: complete method costs and branch counts. Tokens include all generation calls within a method.

Dataset	Method	Mean ms	P50 ms	P95 ms	Input tok.	Output tok.	Retr./q	D/S/E
HotpotQA	Direct	188.15	158.76	328.02	69.008	4.832	0.000	500/0/0
HotpotQA	Always-RAG	551.61	510.25	826.30	616.386	5.100	1.000	0/500/0
HotpotQA	Always-Multihop	1823.84	1752.52	2615.13	2161.830	22.930	2.000	0/0/500
HotpotQA	Matched Binary	494.06	541.27	985.23	431.018	7.364	0.582	209/291/0
HotpotQA	TARG-style	502.72	549.53	989.14	441.422	7.436	0.598	201/299/0
HotpotQA	SUGAR-L-style	2457.48	2620.06	4105.81	2898.828	35.770	1.346	96/135/269
HotpotQA	CAGR-R	596.46	472.26	1922.59	557.040	9.034	0.612	252/190/58
HotpotQA	CAGR-F	513.10	544.49	1017.77	453.184	7.616	0.596	209/284/7
TriviaQA	Direct	193.27	155.65	414.05	66.636	4.972	0.000	500/0/0
TriviaQA	Always-RAG	574.65	498.12	982.04	650.834	6.292	1.000	0/500/0
TriviaQA	Always-Multihop	1797.55	1727.48	2748.58	2086.634	23.944	2.000	0/0/500
TriviaQA	Matched Binary	458.02	435.19	1033.39	394.938	7.394	0.504	248/252/0
TriviaQA	TARG-style	469.46	470.35	1061.30	406.254	7.578	0.524	238/262/0
TriviaQA	SUGAR-L-style	2252.11	2298.50	4321.18	2713.874	33.560	1.164	155/108/237
TriviaQA	CAGR-R	485.33	193.23	1812.76	401.584	8.256	0.420	329/132/39
TriviaQA	CAGR-F	469.36	442.96	1039.01	402.192	7.670	0.510	248/249/3
Overall	Direct	190.71	157.09	367.17	67.822	4.902	0.000	1000/0/0
Overall	Always-RAG	563.13	505.96	920.64	633.610	5.696	1.000	0/1000/0
Overall	Always-Multihop	1810.69	1743.60	2684.22	2124.232	23.437	2.000	0/0/1000
Overall	Matched Binary	476.04	493.62	992.22	412.978	7.379	0.543	457/543/0
Overall	TARG-style	486.09	511.04	1002.26	423.838	7.507	0.561	439/561/0
Overall	SUGAR-L-style	2354.80	2487.58	4215.52	2806.351	34.665	1.255	251/243/506
Overall	CAGR-R	540.90	231.40	1907.31	479.312	8.645	0.516	581/322/97
Overall	CAGR-F	491.23	493.96	1027.15	427.688	7.643	0.553	457/533/10

Table S9: Test Cohort I: CAGR-R minus each comparator. Paired percentile intervals use 20,000 resamples.

Dataset	Comparator	F1 difference, pp [95% CI]	Latency difference, ms [95% CI]
HotpotQA	Always-RAG	-1.9854 [-5.1267, +1.1712]	+89.6099 [+35.6292, +145.1643]
HotpotQA	Matched Binary	+0.9554 [+0.2156, +1.8335]	+170.5996 [+131.1160, +212.2134]
HotpotQA	TARG-style	-0.4195 [-2.3175, +1.4937]	+107.9535 [+55.7876, +161.9864]
HotpotQA	Always-Multihop	-4.3196 [-7.5485, -1.1202]	-1237.7973 [-1299.3365, -1176.4412]
HotpotQA	SUGAR-L-style	-1.8848 [-4.7308, +0.8806]	-1928.2154 [-2038.9212, -1817.7242]
TriviaQA	Always-RAG	+1.4160 [-1.9004, +4.6453]	-141.6455 [-185.3053, -97.1030]
TriviaQA	Matched Binary	-0.2622 [-0.7511, +0.0933]	+66.3848 [+42.9757, +92.7309]
TriviaQA	TARG-style	-1.6069 [-3.5918, +0.3955]	-22.2478 [-59.9785, +16.5257]
TriviaQA	Always-Multihop	+1.0634 [-2.5138, +4.5580]	-1409.8761 [-1472.2032, -1349.0653]
TriviaQA	SUGAR-L-style	-2.0434 [-4.6378, +0.4845]	-1807.5982 [-1919.5320, -1695.3257]
Overall	Always-RAG	-0.2847 [-2.5806, +1.9846]	-26.0178 [-61.2193, +9.4375]
Overall	Matched Binary	+0.3466 [-0.0900, +0.8167]	+118.4922 [+94.9162, +142.4577]
Overall	TARG-style	-1.0132 [-2.3953, +0.3674]	+42.8528 [+10.3780, +76.2866]
Overall	Always-Multihop	-1.6281 [-4.0214, +0.7532]	-1323.8367 [-1367.8070, -1280.3631]
Overall	SUGAR-L-style	-1.9641 [-3.8635, -0.0757]	-1867.9068 [-1947.1576, -1789.0710]

Table S10: Test Cohort II: CAGR-F minus each comparator. Paired percentile intervals use 20,000 resamples.

Dataset	Comparator	F1 difference, pp [95% CI]	Latency difference, ms [95% CI]
HotpotQA	Always-RAG	-2.7952 [-5.5246, -0.0682]	-38.5164 [-67.9644, -8.7380]
HotpotQA	Matched Binary	+0.3100 [-0.1000, +0.8800]	+19.0371 [+6.8789, +33.9314]
HotpotQA	TARG-style	+0.3767 [-0.0500, +0.9767]	+10.3801 [-3.5260, +26.3058]
HotpotQA	Always-Multihop	-4.6384 [-7.9583, -1.3057]	-1310.7400 [-1353.5674, -1268.7251]
HotpotQA	SUGAR-L-style	-2.8743 [-5.5465, -0.2276]	-1944.3846 [-2037.6402, -1848.8714]
TriviaQA	Always-RAG	+3.8200 [+0.7002, +6.9371]	-105.2828 [-142.7383, -67.6756]
TriviaQA	Matched Binary	-0.1733 [-0.5200, +0.0000]	+11.3428 [-1.5541, +29.9724]
TriviaQA	TARG-style	-0.3150 [-0.9900, +0.2417]	-0.0937 [-17.0228, +20.3579]
TriviaQA	Always-Multihop	+3.8632 [+0.5338, +7.1929]	-1328.1877 [-1379.6652, -1278.2727]
TriviaQA	SUGAR-L-style	-2.3817 [-5.0640, +0.2992]	-1782.7505 [-1887.1486, -1680.7089]
Overall	Always-RAG	+0.5124 [-1.5411, +2.5842]	-71.8996 [-95.5306, -47.9076]
Overall	Matched Binary	+0.0683 [-0.2233, +0.3900]	+15.1900 [+5.7215, +26.7394]
Overall	TARG-style	+0.0308 [-0.3775, +0.4317]	+5.1432 [-6.0650, +17.9462]
Overall	Always-Multihop	-0.3876 [-2.7125, +1.9309]	-1319.4638 [-1352.9872, -1286.6152]
Overall	SUGAR-L-style	-2.6280 [-4.5310, -0.7417]	-1863.5676 [-1934.1499, -1792.2188]

S4 Prefix perturbations and development interpretation

Table S11 reports first-stage prefix-truncation sensitivity.

Table S11: Test Cohort I prefix truncation diagnostic: first-stage quality only. The K=5 heads and thresholds remain fixed; no K-specific retraining or new timed generation was performed.

K	Fixed rule	H retrievals	T retrievals	H F1 (%)	T F1 (%)	All F1 (%)
1	CAGR-R first gate	418	394	32.03	43.45	37.74
1	TARG-style	393	370	31.97	43.02	37.49
2	CAGR-R first gate	330	277	31.33	44.82	38.08
2	TARG-style	353	329	32.15	44.90	38.53
3	CAGR-R first gate	302	198	31.55	44.43	37.99
3	TARG-style	340	282	32.19	45.28	38.74
4	CAGR-R first gate	279	166	31.51	44.57	38.04
4	TARG-style	326	255	32.29	44.74	38.52
5	CAGR-R first gate	265	145	30.65	43.06	36.85
5	TARG-style	314	241	32.02	44.40	38.21

The prefix table reconstructs the first decision using saved Direct and Always-RAG answers as paired potential outcomes. Its K=5 row recovers the corresponding matched binary result, providing the reference for fixed-head prefix perturbations.

The five-fold structure-feature comparison uses 30 bounded lexical proxies. Word and character counts, noninitial capitalized spans, punctuation clauses, relative and comparative indicators, preposition counts, digits, and question marks form the general statistics. Fourteen lexical counts and seven initial interrogative indicators complete the vector. Fusion adds the scalar prefix uncertainty and all 30 scalar-feature products. The finite development search tests ridge penalties and initiation rates together with two second-stage model families and three acceptance quantiles. Its OOF ranking is a development result after selection; training-set F1 is explicitly labeled in the table.

S4.1 First-stage quota alignment in Test Cohort I

This post hoc diagnostic ranks questions by the frozen first-stage utility and uses saved Direct and Always-RAG answers for scoring. Stable identifiers break utility ties. One condition selects 555 questions globally, while another matches TARG-style’s 314 HotpotQA and 241 TriviaQA searches. The exact TARG-style subset provides a separate control for the question identities selected at the same quotas. All values describe binary first-stage quality before escalation.

Table S12 reports the first-stage quota-alignment diagnostic.

Table S12: Post hoc first-stage quota decomposition in Test Cohort I, using saved Direct and single-retrieval answers only. H and T each contain 500 questions. Deltas use the original binary gate as reference. Score ordering uses the frozen initiation utility; dataset quotas match 314 HotpotQA and 241 TriviaQA searches. All values describe the binary first-stage outcome before escalation.

First-stage rule	H/T searches	H F1 (%)	H delta (pp)	T F1 (%)	All F1 (%)	All delta (pp)
Original gate	265/145	30.6456	+0.0000	43.0591	36.8523	+0.0000
Score: 555 total	320/235	31.2000	+0.5544	44.5913	37.8957	+1.0433
Score: dataset quotas	314/241	31.5467	+0.9011	44.7247	38.1357	+1.2833
TARG-style selections	314/241	32.0205	+1.3750	44.4038	38.2122	+1.3598

Global alignment adds 145 first searches and raises overall first-stage F1 from 36.8523% to 37.8957%. Dataset-specific alignment reaches 38.1357%, compared with TARG-style’s 38.2122%. On HotpotQA, these conditions recover 0.5544 and 0.9011 points of the initial 1.3750-point gap, leaving 0.4738 points under dataset-specific alignment. The original subsets overlap on 373 questions, with 182 retrieved only by TARG-style and 37 only by CAGR-R. This decomposition identifies contributions from both retrieval allocation and ranking.

This binary first-stage diagnostic holds the utility ranking fixed to separate resource allocation from question selection. Additional first-stage recipients lack some saved escalation signals, restricting reconstruction to the first decision.

For the evidence-text escalation variant, the original AUC comparison interval is calculated on paired OOF predictions. Better AUC need not yield better F1 at a fixed cost threshold, because AUC measures ordering of positive gains rather than their magnitude, harm severity, or execution time. The shared-threshold table therefore reports the actual selected branches as well as discrimination and regression error.

S5 Cost accounting and measurement scope

Generation costs count each physical call once, including direct answers, observed and discarded prefixes, preliminary answers, rewritten queries, bridge continuations, final answers, and the semantic baseline’s sampled and entailment generations. CAGR routes through lightweight heads using the existing model state. Retrieval costs include searches that return duplicate documents, while cache-continuation token totals count newly processed positions and record the reused context separately.

The completed Test Cohort I sessions contain 27,169 generation calls across normal methods, technical executions, and associated warm-up. These calls process 9,100,484 input and 141,783 output tokens, with 9,211 retrieval calls.

Test Cohort II contains 21,144 normal-method generation calls, 6,244 technical generation calls, and 14 cache-check or warm-up calls, totaling 27,402. Normal and technical input totals are 7,375,831 and 1,710,752 tokens, with corresponding output totals of 99,874 and 30,572. Retrieval counts are 6,428 and 2,212, plus 1,000 shared retrieval warm-ups across the two sessions. The separate cache-extension development study uses 2,426 generation calls and 1,318 retrieval calls, including technical checks.

Model and index verification, index staging, dependency restoration, model loading, native compilation, and warm-up are recorded as preparation events separately from steady-state method latency. One-time asset preparation is counted once, while per-question cost summaries cover the complete online path.

The independent on/off experiment holds the scientific path fixed while enabling or disabling feature extraction and decision computation. Two balanced pairs per question yield 2,000 pairs and 4,000 technical executions in each cohort. Statistical resampling uses the 1,000 independent question units per cohort, with 500 per dataset.

S6 Reproducibility resources

The accompanying source-data tables expose the published aggregate values; standard LaTeX tables and vector PDF/SVG figures accompany the manuscript. The local assembly entry reads immutable result summaries, exports tables, and redraws figures without loading model weights or fitting a routing head. Frozen parameter exports preserve both operating points’ numerical values. The model revision, corpus size, tokenizer, prompt text, budgets, and estimator definitions here specify the scientific configuration independently of an execution interface.

The common answer instruction is: “Answer the question using the provided evidence when available. Output only the short answer, without explanation. If evidence is insufficient, use your knowledge; do not invent evidence.” The direct user prompt is “Question:” followed by the question. Retrieved prompts add the fixed ordered evidence blocks. Rewriting requests one short query using the preliminary answer and evidence as clues while explicitly allowing the answer to be wrong. The exact frozen prompt strings and bridge options are exported in `generation_settings.json`; prompt changes constitute a different experiment.

The backbone’s model card provides Qwen2.5 attribution [20], and the original sources specify the terms for HotpotQA and TriviaQA [11, 26]. Retrieved passages retain attribution to their underlying source contexts. The reproducibility repository is <https://github.com/human-com-h/CAGR>.

S7 Reading the evidence across operating points

The matched binary comparison in Test Cohort I estimates CAGR-R’s escalation contribution under a shared first decision and preliminary answer. Within Test Cohort II, CAGR-F versus Always-RAG quantifies the fast-selective operating point. The 200-question development study and 20-question identical-prefix control characterize cache continuation, while five-fold feature comparisons assess initiation signals. Each comparison thus addresses a defined component of the framework.

Together, the results connect efficient signal extraction, evidence-conditioned quality recovery, and selective initiation within a bounded retrieval pipeline. The initiation diagnostic locates the principal multi-hop

quality loss, while cache continuation reduces repeated processing in an escalation component. This combination identifies both the practical savings of two-stage routing and the next modeling challenge in recognizing questions that require relational evidence.

References

- [1] Asai A, Wu Z, Wang Y, et al (2023) Self-RAG: Learning to retrieve, generate, and critique through self-reflection. URL <https://arxiv.org/abs/2310.11511>, arXiv:2310.11511
- [2] Borgeaud S, Mensch A, Hoffmann J, et al (2022) Improving language models by retrieving from trillions of tokens. In: Proceedings of the 39th International Conference on Machine Learning, pp 2206–2240, URL <https://proceedings.mlr.press/v162/borgeaud22a.html>
- [3] Boyd S, Vandenberghe L (2004) Convex Optimization. Cambridge University Press, <https://doi.org/10.1017/CB09780511804441>, URL <https://www.cambridge.org/highereducation/books/convex-optimization/17D2FAA54F641A2F62C7CCD01DFA97C4>
- [4] Dao T, Fu DY, Ermon S, et al (2022) FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In: Advances in Neural Information Processing Systems, <https://doi.org/10.52202/068431-1189>, URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/67d57c32e20fd0a7a302cb81d36e40d5-Abstract-Conference.html
- [5] Dhole KD (2025) To retrieve or not to retrieve? uncertainty detection for dynamic retrieval augmented generation. URL <https://arxiv.org/abs/2501.09292>, arXiv:2501.09292
- [6] Dong Z, Qin F, Shah M, et al (2026) Know before you fetch: Calibrated retrieval-budget allocation for retrieval-augmented generation. URL <https://arxiv.org/abs/2606.29959>, arXiv:2606.29959
- [7] Izacard G, Grave E (2021) Leveraging passage retrieval with generative models for open domain question answering. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, pp 874–880, <https://doi.org/10.18653/v1/2021.eacl-main.74>, URL <https://aclanthology.org/2021.eacl-main.74/>
- [8] Izacard G, Caron M, Hosseini L, et al (2021) Unsupervised dense information retrieval with contrastive learning. URL <https://arxiv.org/abs/2112.09118>, arXiv:2112.09118
- [9] Jeong S, Baek J, Cho S, et al (2024) Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. URL <https://arxiv.org/abs/2403.14403>, arXiv:2403.14403
- [10] Jiang Z, Xu FF, Gao L, et al (2023) Active retrieval augmented generation. URL <https://arxiv.org/abs/2305.06983>, arXiv:2305.06983
- [11] Joshi M, Choi E, Weld D, et al (2017) TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, URL <https://aclanthology.org/P17-1147>
- [12] Kamthan A (2026) AB-RAG: Adaptive budgeted retrieval-augmented generation for reliable question answering. URL <https://arxiv.org/abs/2606.29090>, arXiv:2606.29090
- [13] Karpukhin V, Oguz B, Min S, et al (2020) Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, pp 6769–6781, <https://doi.org/10.18653/v1/2020.emnlp-main.550>, URL <https://aclanthology.org/2020.emnlp-main.550/>
- [14] Kwon W, Li Z, Zhuang S, et al (2023) Efficient memory management for large language model serving with PagedAttention. URL <https://arxiv.org/abs/2309.06180>, arXiv:2309.06180
- [15] Lewis P, Perez E, Piktus A, et al (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems, URL <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- [16] Li B, Tian T, Xu Z, et al (2025) Modeling uncertainty trends for timely retrieval in dynamic RAG. URL <https://arxiv.org/abs/2511.09980>, arXiv:2511.09980
- [17] Moskvoretskii V, Lysyuk M, Salnikov M, et al (2025) Adaptive retrieval without self-knowledge? bringing uncertainty back home. URL <https://arxiv.org/abs/2501.12835>, arXiv:2501.12835
- [18] Ni S, Bi K, Guo J, et al (2024) When do LLMs need retrieval augmentation? mitigating LLMs’ overconfidence helps retrieval augmentation. URL <https://arxiv.org/abs/2402.11457>, arXiv:2402.11457

- [19] Qian P, Wang S, Peng C, et al (2026) When should active RAG retrieve? a budget-aware evaluation of utility, calibration, and cost. URL <https://arxiv.org/abs/2607.24010>, arXiv:2607.24010
- [20] Qwen Team (2024) Qwen2.5: A party of foundation models. URL <https://qwenlm.github.io/blog/qwen2.5/>
- [21] Sah CK, Zhang L, Lian X (2026) Beyond self-knowledge: Propagating uncertainty across reasoning and retrieval in LLMs. URL <https://arxiv.org/abs/2607.25600>, arXiv:2607.25600
- [22] Soudani H, Kanoulas E, Hasibi F (2025) Why uncertainty estimation methods fall short in RAG: An axiomatic analysis. URL <https://arxiv.org/abs/2505.07459>, arXiv:2505.07459
- [23] Su W, Tang Y, Ai Q, et al (2024) DRAGIN: Dynamic retrieval augmented generation based on the information needs of large language models. URL <https://arxiv.org/abs/2403.10081>, arXiv:2403.10081
- [24] Trivedi H, Balasubramanian N, Khot T, et al (2022) Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. URL <https://arxiv.org/abs/2212.10509>, arXiv:2212.10509
- [25] Wang Y, Wei L, Ling H (2026) Retrieval as a decision: Training-free adaptive gating for efficient RAG. Transactions on Machine Learning Research URL <https://openreview.net/forum?id=L8gYtUZfVU>
- [26] Yang Z, Qi P, Zhang S, et al (2018) HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp 2369–2380, URL <https://hotpotqa.github.io/>
- [27] Yao Z, Qi W, Pan L, et al (2024) SeaKR: Self-aware knowledge retrieval for adaptive retrieval augmented generation. URL <https://arxiv.org/abs/2406.19215>, arXiv:2406.19215
- [28] Zhang Z, Fang M, Chen L (2024) RetrievalQA: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. URL <https://arxiv.org/abs/2402.16457>, arXiv:2402.16457
- [29] Zheng L, Yin L, Xie Z, et al (2024) SGLang: Efficient execution of structured language model programs. In: Advances in Neural Information Processing Systems, <https://doi.org/10.52202/079017-2000>, URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/724be4472168f31ba1c9ac630f15dec8-Abstract-Conference.html
- [30] Zubkova H, Park JH, Lee SW (2025) SUGAR: Leveraging contextual confidence for smarter retrieval. URL <https://arxiv.org/abs/2501.04899>, arXiv:2501.04899
- [31] Zubkova H, Park JH, Lee SW (2026) Leveraging contextual confidence for smarter retrieval in large language models. Neural Networks 200:108862. <https://doi.org/10.1016/j.neunet.2026.108862>, URL <https://doi.org/10.1016/j.neunet.2026.108862>