A Selection-Aware Audit of Learned Stopping in Short-Horizon Iterative Retrieval

Supplementary material

Xuan-Bach Le Phat T. Tran-Truong Son Ha Xuan Duy Nguyen

WSDM 2027, Paper 2126

Correspondence: lexuanbach@hcmut.edu.vn
Assigned paper DOI: https://doi.org/10.1145/3862322.3863667

This supplement documents the per-cell evidence and diagnostics supporting the main paper. The public artifact provides the original saved outputs. The companion `camera-ready-analysis.zip` contains the later learning-curve, deployment, and refit records and is available in the public repository. Paths beginning `revision/` below refer to files inside that archive. RQ labels refer to the main paper. Policy codes are Fk for fixed budget k, Rep for answer repetition, TASR for full TASR with native log probabilities, NLI for FBG with an NLI feature and repeat guard, Clf for the binary future-improvement classifier, Err for the error-head gate, StopRAG for a cell-trained Stop-RAG model, V for FBG-only, and G for guarded FBG. Unless a caption states otherwise, entries report exact-match percentage and exact LLM calls. D denotes domination by a simple policy. F denotes an undominated point that does not Pareto-improve the selected $B^\star$. I denotes an undominated point that Pareto-improves $B^\star$: no lower EM and no higher cost, with at least one strict improvement. Equal EM at strictly lower cost therefore qualifies as I. Dom(G) lists policies dominating G (none if no policy does), lb is a one-sided lower bound, CI is a 95% confidence interval, and NA means that an analysis is unavailable. In the archived call-only panels, FBG-only denotes the first-crossing rule V. Sample letters A to E identify question draws. D/F/I occur only in verdict columns, F2/F3 denote fixed budgets, and C denotes calls.

1 Proof sketch for the certificate risk bound

Fix one target risk α and let $L = \lceil \log_2 m \rceil$. Because the stopping rule and error score are fixed before calibration, each episode contributes one independent stopped-state score and error pair. We group equal scores, which gives a nested family of candidate sets in q and prevents a threshold from splitting a tie group. Binary search queries at most L candidates. For a queried candidate with n_q examples and k_q errors, we use the one-sided Clopper and Pearson upper bound

$$U_q = \mathrm{Beta}^{-1}_{k_q+1,n_q-k_q}\left(1-\frac{\delta}{L}\right),$$

with $U_q = 1$ when $k_q = n_q$. The guarantee is inherited from SGR Theorem 1 of Geifman and El-Yaniv [3], this sketch is not an independent proof for arbitrary adaptive threshold searches. The queried cutoffs depend on calibration order statistics, so a naive fixed-candidate binomial union bound alone does not establish validity. We retain SGR's assumptions and search procedure, grouping tied scores. The capped search may skip the smallest candidate, if no queried candidate passes, it conservatively returns no certificate. Each of the six reported α values has its own $1-\delta$ guarantee. These guarantees are separate rather than joint.

Excluded score-mismatch diagnostic. Nine call-only cells predate retention of deployment-matched stopped-prefix features. Their archived diagnostic calibrates on cross-fitted scores but evaluates full-fit scores, so calibration and test scores do not share the same construction. Thirteen of 54 cell and target-risk pairs appear to fire under this mismatched procedure. We exclude all 54 mismatched values from certificate evidence and mark these cells

NA in the main decision matrix. The main certificate analysis uses the 14 core cells with dedicated splits and deployment-matched scores. External and crossed cells with dedicated splits are reported in Sections 4 and 8.

2 Evidence hierarchy

Table 1 separates the primary endpoint, robustness analyses, and exploratory diagnoses so that each result carries its appropriate inference.

Table 1: Claim-to-evidence hierarchy and the inference licensed by each row.

Claim	Evidence population	Status
Guarded FBG passes the declared full-search comparison	One 400-question HotpotQA corpus cell with EM and exact calls	Primary endpoint
Core FBG comparisons have no resolved matched-cost gain	18-cell calls family, 14-cell token and wall-time family, and answer-once ledger	Robustness
More training resolves a small FBG-only matched-call gain	One four-search cell, 180 to 2,880 training chains, fixed 2,000-question test set	Post-review extension
Question sampling can change the endpoint decision	Same-question reruns and four independent question draws	Prospectively frozen
Signal role and selection explain the frontier gap	Alternative heads, calibration, rerolled futures, nested reselection, and oracle	Exploratory diagnosis
The certificate refuses unsupported subsets	84 cell and target-risk pairs over 14 fully instrumented core cells	Robustness

3 Regime-specific policy guidance

Table 2 connects observed regimes to policy choices. The recommendation follows the matched comparison, including zero certificate coverage when calibration cannot support the requested risk.

Table 2: Regime-specific policy guidance supported by the evaluated cells.

Observed regime	Evidence in this study	Practical decision
Accuracy saturates by the second search	An FBG-family rule is the cheapest policy within 1 EM of full search in 1 of 14 fully instrumented core cells	Select fixed-2 or repetition on development data
Large shared index with cross-document negatives	Fixed-2 reaches 23.0 EM at 5.00 calls and dominates guarded FBG at 21.25 EM and 5.90 calls	Prefer the selected fixed budget
Question populations have heterogeneous saturation	Guarded FBG passes on some independent draws and fails on others	Validate on a new question draw before deployment
The selected threshold lies at the grid boundary	FBG-only can collapse while the latched guard reduces to answer repetition	Treat the score as unstable and require a new equal-cost test
Core cells have 30 calibration episodes and high answer error	Corrected SGR returns no certified subset for targets through 0.6 in the 14 instrumented core cells	Report zero certificate coverage
Evidence sufficiency is predictable	Gold-support AUROC reaches 0.65 to 0.95, yet no cell has a positive cellwise equal-call interval	Evaluate policy utility separately from target discrimination

4 External news-corpus extension

We wrote the MultiHopRAG endpoint protocol before fitting the external cell or evaluating its 400 test questions. The adapter retains inference queries and indexes 53,943 sentence units from all 609 English news articles. It excludes comparison, temporal, and null queries under the declared answer-format rules. The cell uses Qwen2.5-3B, 96 policy-training episodes, 24 certificate episodes, and the same four-search cap as the primary HotpotQA corpus cell.

Table 3: External-cell policies and equal-call differences from the randomized simple-policy hull.

Policy	EM	Calls	ΔEM	95% CI
FBG-only	73.00	7.27	+0.48	$[-0.50, 1.46]$
Guarded FBG	73.25	7.40	+0.71	$[-0.07, 1.57]$
Future-improvement clf.	70.50	4.35	+2.10	$[0.12, 4.10]$
Error-head gate	73.75	7.65	+1.18	$[0.41, 2.06]$

In Table 3, guarded FBG meets its prospectively frozen comparison with full search because its paired difference is +0.50 EM $[0.00, 1.25]$ and calls fall from 9.00 to 7.40. Its equal-call interval remains unresolved. The two positive signal-policy intervals are exploratory results, distinct from the frozen endpoint. Gold-support classifier is ineligible because the adapter lacks aligned support annotations. With 24 dedicated calibration episodes, SGR certifies subsets at $\alpha = 0.3, 0.4, 0.5, 0.6$: test coverage is 77.8%, 87.5%, 99.0%, 99.0%, and test risk is 0.148, 0.183, 0.260, 0.260, respectively. These are empirical test risks, not additional guarantees.

5 Signal calibration and evidence sufficiency

We evaluate target quality separately from policy quality on the 18-cell common family. At each test prefix, the error head has expected calibration error from 0.019 to 0.087 with median 0.043. Its selective-risk ordering improves over a random ordering in all 18 cells. The FBG value head improves over random in 7 of 18. These diagnostics use test labels only for evaluation and do not change any threshold or stopping time.

The gold-support classifier marks a prefix sufficient when the retained evidence covers at least 0.8 of annotated support tokens. Gold support is read during training only. With the nine controller features, cross-fitted AUROC ranges from 0.652 to 0.946. This policy is in category D in 8 cells, F in 8, and I in 2. A twelve-feature lexical variant has AUROC from 0.629 to 0.945 and categories 9/7/2 in D/F/I order. Neither variant has a cellwise positive equal-call interval. The classifier discriminates evidence coverage, while its selected stopping policy leaves gains over the simple-policy frontier unresolved.

6 Positioning against the closest stopping methods

Table 4 compares the signals and evaluation conditions. We use "matched" when a policy replays identical trajectories and shares our call accounting. We use "transplanted" when a released system retains its own retriever or generation schedule on the baseline-implementation cell. "Not run" identifies a declared missing condition. Our matched rows characterize sparse first-stage retrieval. Dense retrieval is a separate evaluation setting. We price transplanted methods in *searches* because their per-question LLM-call counts are not comparable with our exact-call ledger.

The core deterministic family contains fixed budgets and answer repetition. Section 21 additionally tests evidence-growth saturation and query repetition, with thresholds selected on training chains. Retrieval-score saturation, semantic novelty, grounding thresholds, and cheap-controller cascades remain untested as standalone rules. The final row below describes the core family. Expanded-family results are reported separately. No frontier claim covers all possible deterministic controllers.

Table 4: Signals, requirements, matched conditions, and measured outcomes for the closest stopping and adaptive-retrieval methods.

Method	Stop signal	Needs, action space, backend	Comparison here	Outcome in this audit
Stop-RAG [8]	$Q(\lambda)$ stop and continue values	logged trajectories and encoder training	matched on 14 common-family cells with cell-specific retraining	dominated in 9 cells, frontier-only in 4, and improves the selected fixed budget in 1
TASR [6]	answer stability and calibrated logits	native answer-token log probabilities	matched on 11 of 14 fully instrumented core cells	on the pooled all-policy frontier in 2 of 11, beaten by simple policies in 5 of 11 core cells (5 of 12 including Llama)
HALT [10]	generated-claim coverage	claim generator and trained verifier	verifier-inspired substitute on six cells	dominated in all six cells under both public substitute verifiers, by fixed-2 in five (NLI) and four (relevance)
SIM-RAG [12], S2G-RAG [7]	evidence sufficiency	learned sufficiency supervision	related signal only	motivates our separate gold-support classifier
Gold-support classifier	support-token coverage	gold support during training only	matched on 18 cells	cross-fit AUROC 0.65 to 0.95 with no cellwise equal-call win
Error-head gate	current answer error	correctness labels	matched on 18 cells	one cellwise equal-call win and no max-t-resolved win
LearnStop [2]	learned exit versus static token budget	Learn-then-Test risk certification	reasoning-model services	selects the cheaper certified family, our audit tests external retrieval at matched calls, tokens and time
Self-RAG [1]	reflection tokens	fine-tuned model with its own dense retriever	transplanted released 7B system	7.9 EM at 0.56 searches on the baseline-implementation cell
FLARE [5]	token-probability uncertainty	mid-generation retrieval and native logits	transplanted direct system	34.3 EM at 1.66 searches on the baseline-implementation cell
Fixed budgets and repetition	deterministic budget or answer stability	no learned predictor	primary matched comparator in every cell	guarded FBG is dominated in 15 of the 26 core cells

7 Frontiers, targets, and margin sensitivity (RQ1)

Table 5: Fourteen-cell frontier summary with nondominated counts, cheapest policies within 1/2/3 EM, and mean matched-cost differences.

		Band	ΔEM at equal cost		
Policy family	**Nondom.**	**1/2/3**	**calls**	**tokens**	**wall**
Fixed budgets	12 of 14	4/7/9	NA	NA	NA
Answer repetition	5 of 14	1/0/0	NA	NA	NA
FBG-only	11 of 14	0/2/1	+0.01	−0.02	+0.02
FBG + repeat guard	3 of 14	0/0/0	−0.59	−0.64	−0.60
FBG + NLI guard	NA	1/0/0	NA	NA	NA
Future-improvement clf.	13 of 14	5/5/4	−0.06	−0.07	−0.06
Full TASR	2 of 11	2/0/0	−0.32	−0.34	−0.33
Stop-RAG	5 of 14	NA/NA/NA	−1.20	NA	NA
Error-head gate	3 of 14	0/0/0	−0.47	−0.47	−0.48
Verifier-inspired coverage	0 of 6	0/0/0	NA	NA	NA

In Table 5, denominators follow signal eligibility. Nondominated counts use the pooled all-policy frontier, except Stop-RAG, whose expanded count is against the simple family. Stop-RAG has equal-call vectors for fourteen cells, but its complete token, wall-time, and all-policy band ledgers are unavailable. We report those entries as NA.

Table 6: D/F/I counts across the 26 core cells under ϵ-dominance. The zero tolerance matches the point matrix.

ϵ **(EM)**	**FBG-only D/F/I**	**Guarded FBG D/F/I**
0	10/13/3	15/6/5
0.5	15/9/2	18/4/4
1	18/8/0	23/2/1
2	20/5/1	25/1/0

Table 6 varies the EM tolerance around the core matrix, which assigns guarded FBG to category I in five cells. These are Sonnet 4.5, the call-only Qwen and MuSiQue with Llama cells, and independent corpus draws D and E. Category D includes GPT-oss-120B, Command-R, two of the three same-question runs, the Wikipedia 25M cell, and all three 2Wiki draws. The remaining six cells are frontier-only and need not be accuracy ties.

In Tables 7 and 8, "logged" is the deployed run and "refit" reruns the same pipeline with seeded tie-breaking. The net and discounted targets remain nonnegative and use $\max_{t'>t}[u_{t'} - u_t - c(t'-t)]_+$ and $\max_{t'>t} \gamma^{t'-t}[u_{t'} - u_t]_+$, respectively. Clf is a binary future-improvement classifier. Err thresholds the error head. All thresholds use the frozen band rule on cross-fitted training chains. Table 9 then compares each selected policy with the simple hull in three cost coordinates. Intervals spanning zero leave the direction unresolved.

Table 7: Guarded-policy EM/calls and verdicts under alternative targets and learners.

(a) Logged, refit, and net targets

Run	FBG logged	FBG refit	net c=.02	net c=.04	net c=.08
A s7	24.8/6.75 F	24.8/6.75 F	24.8/6.77 F	24.5/6.67 D	24.5/6.65 D
A s11	25.3/6.51 D	25.0/6.90 D	24.8/6.90 D	25.3/6.89 D	25.5/6.84 D
A s19	25.5/6.68 D	25.5/6.68 D	25.8/6.63 F	25.8/6.66 F	26.0/7.07 D
B	29.3/6.71 F	28.0/5.97 D	28.3/5.83 D	28.3/5.82 D	27.8/5.83 D
C	23.3/5.90 F	23.3/5.90 F	23.3/5.90 F	23.3/5.90 F	23.3/5.90 F
D	27.0/6.36 I	26.5/5.90 D	26.5/5.90 D	26.5/5.90 D	26.5/5.90 D
E	34.3/7.76 I	34.3/7.76 I	34.0/8.21 F	34.0/8.20 F	34.0/8.16 F
2Wiki d1	23.5/5.50 D	23.5/5.50 D	23.5/5.50 D	23.5/5.50 D	23.5/5.50 D
2Wiki d2	23.5/5.59 D	24.0/7.32 D	23.8/7.34 D	23.8/7.26 D	23.3/6.66 D
2Wiki d3	25.8/5.66 D	25.8/5.82 D	25.8/5.81 D	25.8/5.66 D	25.8/5.66 D
MuSiQue d1	7.5/6.54 D	7.5/6.54 D	7.5/6.53 I	7.5/6.53 I	7.5/6.53 I
MuSiQue d2	9.3/6.50 D	9.8/6.89 D	9.8/6.99 D	9.8/7.00 D	9.8/6.98 D
MuSiQue d3	10.8/6.43 D	10.8/6.43 D	10.8/6.43 D	10.8/6.43 D	10.8/6.43 D
Wikipedia 25M	21.3/5.90 D	21.3/5.90 D	21.3/5.95 D	21.3/5.90 D	21.3/5.90 D

(b) Discounted targets, one-step learner, and classifiers

Run	disc. γ=.5	disc. γ=.8	one-step	Clf	Err
A s7	25.0/6.64 I	24.5/6.56 D	25.0/6.70 I	24.5/5.11 D	24.5/8.80 D
A s11	25.0/6.36 D	25.5/6.76 D	24.5/6.40 D	25.8/4.96 I	25.0/8.90 D
A s19	25.8/7.05 D	26.0/7.06 D	26.3/6.95 I	24.0/4.94 F	26.0/8.88 D
B	27.8/5.75 D	28.0/5.81 D	28.3/5.89 D	27.8/4.97 F	30.3/7.70 D
C	23.3/5.90 F	23.3/5.90 F	23.3/5.90 F	18.5/3.91 F	26.5/8.84 F
D	26.5/5.90 D	26.5/5.90 D	26.5/5.90 D	25.5/4.97 I	26.0/8.64 D
E	34.3/7.62 I	32.5/5.86 F	34.3/6.86 I	27.3/4.00 F	34.0/8.67 F
2Wiki d1	23.5/5.50 D	23.5/5.50 D	23.5/5.50 D	23.5/4.11 F	23.5/6.79 D
2Wiki d2	23.3/5.51 D	23.8/7.06 D	23.3/5.51 D	19.3/4.15 F	25.3/9.00 D
2Wiki d3	25.3/5.61 D	25.5/5.64 D	25.5/5.66 D	25.3/3.79 F	25.8/8.75 D
MuSiQue d1	7.5/6.53 I	7.5/6.53 I	7.5/6.53 I	5.3/3.41 F	9.3/9.00 F
MuSiQue d2	9.5/6.89 D	9.8/6.98 D	9.5/6.99 D	9.8/4.95 F	10.3/9.00 F
MuSiQue d3	10.8/6.43 D	10.8/6.43 D	10.8/6.43 D	6.5/3.08 F	12.0/9.00 F
Wikipedia 25M	21.3/5.88 D	21.3/5.88 D	21.3/5.90 D	17.8/3.00 F	21.5/8.97 D

Table 8: Same targets under the FBG-only rule (EM/calls and verdict).

Run	FBG (logged)	FBG (refit)	net c=.02	net c=.04	net c=.08	disc. γ=.5	disc. γ=.8	one-step
A s7	25.0/6.35 I	25.0/6.35 I	25.0/6.37 I	24.8/6.26 F	24.8/6.24 F	25.5/6.20 I	25.0/6.12 I	25.3/6.27 I
A s11	25.3/5.74 D	25.8/6.39 D	25.5/6.37 D	25.5/6.27 D	25.8/6.20 D	25.0/5.48 D	25.8/6.10 D	24.5/5.55 D
A s19	25.3/6.38 D	25.3/6.38 D	25.5/6.32 D	25.5/6.35 D	25.8/6.87 F	25.3/6.88 D	25.8/6.86 F	26.0/6.83 I
B	29.5/6.62 F	27.8/4.23 F	28.0/4.04 F	27.8/4.03 F	27.3/3.96 F	7.0/1.01 D	26.8/3.87 F	27.8/3.96 F
C	13.5/2.67 F	13.5/2.67 F	13.5/2.67 F	13.0/2.65 F	11.0/2.16 F	6.5/1.00 F	11.0/2.16 F	6.5/1.00 F
D	26.3/5.42 I	19.5/3.07 D	19.5/3.06 D	19.8/3.08 F	19.5/3.06 D	19.5/3.05 D	19.5/3.03 D	16.8/2.68 F
E	33.3/7.51 F	33.3/7.51 F	33.0/8.03 F	33.0/8.01 F	33.3/7.96 F	33.8/7.25 F	26.5/3.42 F	33.8/6.46 F
2Wiki d1	20.3/3.11 F	20.3/3.11 F	20.5/2.98 F	20.5/2.97 F	20.5/2.97 F	8.8/1.00 F	19.8/2.87 F	10.5/1.40 F
2Wiki d2	19.0/3.84 F	23.3/7.06 D	22.5/7.02 D	23.0/6.98 D	21.8/6.20 D	6.0/1.02 D	22.5/6.70 D	6.0/1.02 D
2Wiki d3	24.3/3.12 F	25.5/4.07 F	25.5/4.03 F	24.0/3.10 F	11.0/1.57 F	6.0/1.00 F	10.0/1.39 F	9.8/1.41 F
MuSiQue d1	3.8/3.03 F	3.8/3.03 F	3.3/2.70 F	0.5/1.00 F	0.5/1.00 F	0.5/1.00 F	0.5/1.00 F	0.5/1.00 F
MuSiQue d2	2.5/1.44 F	9.8/6.21 D	9.8/6.80 D	9.8/6.81 D	9.8/6.79 D	9.3/6.71 D	9.8/6.78 D	10.3/6.60 I
MuSiQue d3	0.3/1.00 F	0.3/1.00 F	0.3/1.00 F	0.3/1.00 F	0.3/1.00 F	0.3/1.00 F	0.3/1.00 F	0.3/1.00 F
Wikipedia 25M	18.0/3.14 F	18.0/3.13 F	18.0/3.19 F	18.0/3.11 F	17.8/3.09 D	10.0/1.51 F	10.0/1.51 F	10.0/1.55 F

Table 9: Paired EM difference from the randomized simple-policy frontier at equal mean calls, input tokens, or wall time (95% bootstrap CI).

	FBG + guard			FBG-only		
Run	**equal calls**	**equal tokens**	**equal wall**	**equal calls**	**equal tokens**	**equal wall**
A s7	-0.19 [-1.22,0.84]	-0.23 [-1.35,0.79]	-0.20 [-1.25,0.83]	+0.16 [-0.85,1.17]	+0.14 [-0.91,1.16]	+0.17 [-0.84,1.17]
A s11	-0.50 [-2.50,1.25]	-0.50 [-2.50,1.25]	-0.50 [-2.50,1.25]	-0.50 [-2.00,1.00]	-0.50 [-2.00,1.00]	-0.50 [-2.00,1.00]
A s19	-0.37 [-1.44,0.69]	-0.42 [-1.55,0.66]	-0.37 [-1.44,0.69]	-0.51 [-1.88,0.84]	-0.52 [-1.89,0.79]	-0.48 [-1.86,0.84]
B	-1.89 [-3.56,-0.39]	-2.01 [-3.72,-0.45]	-1.89 [-3.56,-0.39]	-1.52 [-3.17,-0.01]	-1.61 [-3.30,-0.06]	-1.52 [-3.17,-0.01]
C	-0.12 [-1.54,1.32]	-0.36 [-1.75,1.04]	-0.16 [-1.57,1.28]	+0.32 [-1.18,1.77]	+0.25 [-1.25,1.69]	+0.28 [-2.39,3.04]
D	+0.50 [-0.50,1.50]	+0.50 [-0.50,1.50]	+0.50 [-0.50,1.50]	+0.28 [-0.92,1.53]	+0.35 [-0.74,1.55]	+0.30 [-0.86,1.53]
E	+0.85 [-0.03,1.80]	+0.80 [-0.03,1.64]	+0.80 [-0.03,1.63]	+0.02 [-1.46,1.43]	-0.01 [-1.12,1.08]	-0.01 [-1.12,1.08]
2Wiki d1	-1.50 [-3.25,0.00]	-1.50 [-3.25,0.00]	-1.50 [-3.25,0.00]	+1.86 [-0.19,4.02]	+1.44 [-0.60,3.54]	+1.95 [-0.10,4.12]
2Wiki d2	-0.94 [-2.37,0.38]	-1.05 [-2.48,0.27]	-0.94 [-2.37,0.39]	-1.80 [-3.34,-0.34]	-1.75 [-3.30,-0.29]	-1.78 [-3.32,-0.32]
2Wiki d3	-0.50 [-2.00,1.00]	-0.50 [-2.00,1.00]	-0.50 [-2.00,1.00]	+1.06 [-0.63,2.79]	+1.01 [-0.67,2.72]	+1.07 [-0.62,2.81]
MuSiQue d1	-0.52 [-1.77,0.73]	-0.61 [-1.82,0.61]	-0.60 [-1.82,0.61]	-0.17 [-1.80,1.47]	+0.26 [-0.29,1.04]	-0.08 [-1.69,1.55]
MuSiQue d2	-0.84 [-1.91,0.13]	-0.86 [-1.89,0.11]	-0.85 [-1.90,0.10]	+0.98 [-0.15,2.27]	+0.74 [-0.38,1.99]	+0.96 [-0.18,2.24]
MuSiQue d3	-0.45 [-1.55,0.61]	-0.48 [-1.58,0.53]	-0.48 [-1.58,0.54]	+0.00 [0.00,0.00]	+0.00 [0.00,0.00]	+0.00 [0.00,0.00]
Wikipedia 25M	-1.75 [-3.25,-0.50]	-1.75 [-3.25,-0.50]	-1.75 [-3.25,-0.50]	-0.10 [-0.46,0.46]	-0.08 [-0.43,0.49]	-0.07 [-0.41,0.50]

8 Crossed backbone × question population

We repeat independent corpus-index draws on additional backbones under the same frozen pipeline: three 200-question Haiku draws and one 400-question Llama draw. Because we refit each stopping head within its cell, these crossed cells jointly vary backbone and question population. Table 10 measures recurrence under this per-cell refitting. With 30 dedicated calibration episodes, Haiku data23 certifies at $\alpha = 0.5$ (64.5% coverage, 0.364 test risk) and 0.6 (87.5%, 0.434), data29 certifies at 0.6 (47.0%, 0.351). Haiku data17 and the Llama cell return no certificate at the tested targets. Thus calibration size alone does not determine refusal.

Table 10: Crossed backbone × population draws, including frozen verdicts, margin tests, and equal-call contrasts.

Run	$B^\star$	**V**	**G**	**Verdict** ($\epsilon = 0$)	**Dom(G)**	**lb (1/2/3)**	**Δ@calls**
Llama B	39.0/9.0 (4)	38.8/7.59	38.8/7.61	F	V	-1.25 ×✓✓	+0.27 [-1.17,1.74]
Haiku B	49.5/7.0 (3)	36.0/3.25	49.5/5.59	D	Rep	-2.50 ××✓	-0.01 [-0.04,0.02]
Haiku C	48.5/7.0 (3)	49.5/7.29	49.5/7.32	F	V	-1.50 ×✓✓	+0.76 [-1.00,2.52]
Haiku D	46.0/7.0 (3)	41.0/4.36	45.0/5.94	D	F2, Rep	-2.50 ××✓	-0.47 [-2.41,1.06]

9 The pooled frontier per cell (RQ1)

Table 11 identifies nondominated learned policies (ND-learned), policies dominating the guard, and the cheapest policies within 1/2/3 EM of full search. The lower bounds and tolerance columns separate inferential comparisons from descriptive frontier membership.

Table 11: Pooled accuracy-to-call frontiers and frozen verdicts for fourteen fully instrumented corpus-index cells.

Run	**ND-learned**	**Dom(G)**	**Winner 1/2/3**	**lb (1/2/3 EM)**	**ϵ 0/1**
A s7	V, StopRAG	V, TASR	F2/F2/F2	-0.75 ✓✓✓	F/D
A s11	Clf	F2, V, Clf	Clf/Clf/Clf	-0.75 ✓✓✓	D/D
A s19	Clf	Rep	Rep/Clf/Clf	-1.25 ×✓✓	D/D
B	V, Clf	V	F3/V/F2	-3.25 ×××	F/D
C	V, G, TASR, Clf, Err	none	TASR/F3/F3	-5.00 ×××	F/D
D	V, G, Clf	none	Clf/Clf/Clf	+0.25 ✓✓✓	I/D
E	G, NLI, Clf	none	NLI/F3/F2	-0.50 ✓✓✓	I/I
2Wiki d1	V, Clf	F2, Clf	Clf/Clf/V	-1.50 ×✓✓	D/D
2Wiki d2	V, Clf	F2	F3/F2/F2	-3.25 ×××	D/D
2Wiki d3	V, Clf	F2	Clf/V/F1	-1.25 ×✓✓	D/D
MuSiQue d1	V, NLI, StopRAG, Clf, Err	Rep, NLI	StopRAG/F2/F2	-3.00 ×××	D/D
MuSiQue d2	V, Clf, Err	F2, Rep, Clf	Clf/Clf/Clf	-2.25 ××✓	D/D
MuSiQue d3	V, TASR, Clf	F2	TASR/F2/F2	-2.25 ××✓	D/D
Wikipedia 25M	V, Clf	F2, Rep	F2/F2/F2	-1.75 ×✓✓	D/D

10 The hindsight-oracle ceiling (RQ1)

The hindsight oracle quantifies stopping headroom within the logged trajectories. It reads every prefix grade before selecting a stop, so it provides a diagnostic ceiling outside the deployable simple-policy family and its D/F/I verdicts. It remains a ceiling, not a deployable policy.

For each question, the oracle stops at the first correct prefix. It charges the full budget when no prefix is correct, as does every policy that never stops early. We do not use the stored `oracle_searches` field because its score-maximizing tie break selects the cheapest prefix on never-correct questions and therefore assigns an unattainable zero-search cost. This field averages 0.32 searches on the primary cell. We instead recompute the stopping point under the stated convention and price it with the same `cumulative_cost` routine used for every reported policy.

The main paper reports the full per-cell ceiling. The oracle dominates the full budget in all fourteen cells. Its gain ranges from +1.75 to +8.00 EM, with mean +4.84. The smallest headroom occurs on the three MuSiQue draws (+1.75 to +2.25 EM, where full-budget EM is 9.25 to 12.00). On the primary cell the ceiling is 29.50 EM at 3.14 searches / 7.28 exact calls against the full budget's 24.50 EM at 4.00 / 9.00. The logged trajectories contain five EM points of headroom at lower cost. Guarded FBG reaches 24.75 EM. The oracle turns the 79 harmful and 14 helpful reachable flips into a frontier ceiling. This headroom is measurable, yet the fitted future-gain predictor does not recover it. The ceiling is *not* evidence that stopping works, and the deployable-policy verdicts retain their original simple-policy reference.

11 Question composition and draw variation (RQ2)

Full-budget EM moves from 24.50 to 34.25 across the five HotpotQA corpus draws, and the guarded rule's verdict moves with it. If the draws differed in *composition*, standardization would reduce that disagreement. We test it on the strata the runs already log, including HotpotQA question type, its difficulty level, and gold-support retrieval recall (the share of a question's gold support sentences appearing anywhere in the union of its retrieved evidence, matched after stripping the `Title:` prefix and normalizing to alphanumeric tokens). We compute these strata from the retained logs.

The `level` field is degenerate because all 400 questions in every draw are labeled `hard`. Difficulty therefore cannot be tested as a moderator. We treat correlations over five draws as descriptive and rely on direct standardization for the composition analysis.

In Table 12, Br/Cm are bridge/comparison counts, R is gold-support retrieval recall, and Δ is guarded minus full-budget EM within a stratum.

Table 12: Question-type and gold-support-recall strata across five HotpotQA corpus draws (n=400 each, with point contrasts and no new intervals).

Draw	Full EM	Br/Cm	EM Br/Cm	$\bar{R}$	$n_{R=1}$	EM R=1	EM R<1	$\Delta_{R=1}$	$\Delta_{R<1}$
A	24.50	324/76	21.3/38.2	0.559	111	46.9	15.9	+0.90	+0.00
B	31.25	329/71	27.7/47.9	0.555	115	52.2	22.8	-5.22	-0.70
C	26.50	331/69	22.7/44.9	0.549	114	47.4	18.2	-10.53	-0.35
D	25.50	336/64	23.2/37.5	0.523	92	46.7	19.2	+1.09	+1.62
E	34.25	323/77	31.0/48.0	0.592	124	50.8	26.8	+0.00	+0.00

The composition null. Question-type mixture barely moves (323 to 336 bridge of 400, an 80.8 to 84.0% share) and its correlation with the draw's full-budget EM is negative and unresolved ($r = -0.45$, p=0.45, n=5). Mean gold-support recall moves a little more (0.523 to 0.592) and correlates positively, also unresolved (r=0.74, p=0.15). To test the composition explanation directly, we standardize every draw to the pooled distribution over the four question-type × recall cells. Standardization leaves the full-budget EM swing at 8.87 EM against a raw 9.75 EM. Composition accounts for 0.88 of 9.75 EM points, about 9%. The remaining 8.87 EM is unexplained by these composition variables. Five draws support this descriptive contrast but leave stable variance components unidentified. **The measured draw-to-draw disagreement is not explained by the tested composition**

variables, and the two failing draws remain two of four independent question draws that fail the guard. Their one-sided lower bounds are −3.25 and −5.00.

Retrieval recall localizes the losses. Retrieval recall predicts answerability within every draw. Full-budget EM ranges from 46.7 to 52.2 inside the $R = 1$ stratum and from 15.9 to 26.8 outside it, while the $R = 1$ subgroup remains between 23.0 and 31.0% of each draw. Its stable prevalence means that it cannot explain the draw-level swing. It does, however, locate the guard's errors: in the two failing draws the loss is almost entirely inside the answerable subgroup (−5.22 and −10.53 EM at R=1, against −0.70 and −0.35 outside it). The losses concentrate on questions whose logged retrieval contains the complete support. This localization identifies where the stopping policy needs to improve. It qualifies where the loss lands and is not evidence that the failures are artifacts.

12 Answer-once cost accounting (RQ1)

The exploratory repeat-conditioned gain policy fits FBG on nonterminal training states with repeated normalized answers. It stops at the first repeated-answer state with predicted gain below a training-selected threshold, or uses the full budget if none qualifies. This policy was defined after a partial look at the primary run, then frozen for its decoding reruns.

The main paper charges every policy the $1+2T$ live LLM calls the runs actually made. These consist of one initial answer, then a query and an answer per search. The FBG critic reads five features that need an answer at *every* prefix (`answer_changed`, `answer_length`, `answer_grounding`, `stable_two`, `distinct_answers`). A conventional agent can instead defer generation until the end and pay $1+T$ calls, comprising one query per search and one final answer. Only fixed budgets can adopt this accounting because their stopping signal needs no per-prefix answer. Answer repetition, the NLI guard, full TASR and every FBG rule compare or score answers at each prefix and therefore stay at $1+2T$.

Repricing changes the currency, not what was executed: EM and per-question search counts are the measured values in the `corpus_confirmation*.json` registries, and nothing was re-run. Under the answer-once accounting of Table 13, the simple-policy advantage grows. Over the 18-cell matched-cost family of fourteen core and four crossed-backbone runs, the guarded rule is Pareto-dominated by the simple family in 17 of 18 cells under the answer-once ledger, against 11 of 18 under the primary one. Answer repetition is dominated by a plain fixed budget in all 18, against 11 of 18. The guard costs more than every fixed budget in all 18 and more than the entire simple family in 15. Under this currency, its cost lies outside the fixed-budget range in every cell. We therefore report point domination: a fixed-budget mixture cannot match the guard's cost under this ledger. We retain the primary accounting because it records the executed answer-at-every-prefix loop. Under answer-once accounting, our recommendation reduces to tuning a fixed budget because repetition also loses its cost advantage.

Table 13: Primary-run accuracy and cost under measured prefix-answer calls and the answer-once counterfactual ledger.

		LLM calls		
Policy	**EM**	**primary**	**answer-once**	**dominated by (answer-once)**
fixed-0	9.00	1.00	1.00	NA
fixed-1	18.25	3.00	2.00	NA
fixed-2	24.50	5.00	3.00	NA
fixed-3	25.00	7.00	4.00	NA
full budget	24.50	9.00	5.00	fixed-2, fixed-3
answer repetition†	24.25	5.72	5.72	fixed-2, fixed-3, full budget
FBG + NLI guard†	24.00	6.03	6.03	answer repetition, fixed-2, fixed-3, full budget
FBG-only†	25.00	6.35	6.35	fixed-3
full TASR†	24.75	6.68	6.68	fixed-3
FBG + repeat guard†	24.75	6.75	6.75	fixed-3
repeat-conditioned gain†	24.50	6.78	6.78	fixed-2, fixed-3, full budget

Tail latency. The retained sequential ledgers cover 18 matched-cost cells on three backbones. Across these cells, full-budget per-question p90 wall time ranges from 1.93 to 17.50 seconds, while p95 ranges from 2.00 to 19.61

seconds with a median of 6.85. Per-call p95 latency ranges from 0.24 to 1.81 seconds. These quantiles describe the logged local and API executions. They do not include queueing, concurrency, retries, or live-service failures, which were not logged consistently enough for inference.

Live deployment benchmark. To complement the logged replays and the answer-once ledger, we ran one narrow live benchmark after acceptance on a single machine (Apple M5 Max laptop, 128 GB, Ollama 0.30.10), which differs from the hardware of the original runs. On the primary HotpotQA corpus cell we used the same quantized Qwen2.5-3B-Instruct weights (matching digest) and the same BM25 corpus. Before any live call, the reconstructed frozen guarded rule reproduced the logged stopping step on all 400 test questions. A protocol frozen before timing fixed four schedules, which we executed sequentially on all 400 test questions with the order randomized within each question: the guarded rule answering after every search, the training-selected budget of three searches answering after every search, the same budget answering once, and the full four-search budget answering once. EM was 25.00, 25.00, 24.75, and 24.75. Mean LLM calls were 6.75, 7.00, 4.00, and 5.00, and mean end-to-end wall time per question was 3.82, 4.00, 3.03, and 3.87 seconds, with no failed calls, retries, or timeouts. Relative to answering once after three searches, the guarded rule used 2.75 more calls and 0.79 more seconds per question (26.2%, 95% paired bootstrap CI 22.8 to 29.5%), while its EM difference of 0.25 points was unresolved (CI −1.00 to 1.50). Relative to the same budget answering after every search, it saved 0.18 seconds (4.5%, CI 2.3 to 6.8%) at identical EM. Against the full budget answering once, the wall-time difference was −0.06 seconds (CI −0.16 to 0.04). The measured accuracy versus wall-time frontier therefore contains only the answer-once fixed budget and the guarded rule. Live stops matched the logged stops on 399 of 400 questions. Unrelated jobs shared the machine during the run, and absolute times drifted by roughly a factor of two. In a post hoc split of the run into halves, the gap against the answer-once fixed budget kept its sign (24.0% and 30.8%). This is one deployment on one machine and one backbone, and it does not measure lifecycle cost, invoices, or energy. Protocol, raw per-call records, and analysis are in `revision/ls9-deployment`.

13 Coverage probes with substitute verifiers (RQ1)

We probe claim coverage with each run's frozen backbone and a fixed public cross-encoder. The backbone generates 1.7 to 2.0 claims per question at a cost of one call. Selected coverage thresholds vary by cell. This substitute does not evaluate HALT's trained verifier, whose released adapter requires a 24GB CUDA GPU. We select thresholds with the 1-EM training-chain band. Token and wall-time totals omit claim-call tokens, so we compare exact calls in Table 14. Both substitute-verifier policies are dominated in all six cells, by fixed-2 in five (NLI) and four (relevance).

Table 14: Accuracy, exact calls, and dominating policies for two verifier-inspired substitute verifiers on six cells.

			HALT (NLI verifier)		HALT (relevance)	
Run	**fixed-2**	**FBG + guard**	**EM/calls**	**dominated by**	**EM/calls**	**dominated by**
A s7	24.5/5.00	24.8/6.75	24.0/9.66	F2, F3, F4, Rep, +7 learned	24.0/8.70	F2, F3, Rep, +5 learned
C	22.3/5.00	23.3/5.90	22.0/7.55	F2, F3, Rep, +3 learned	23.8/8.66	F3, +2 learned
D	25.5/5.00	27.0/6.36	25.3/9.23	F2, F3, F4, Rep, +6 learned	24.0/8.67	F2, F3, Rep, +6 learned
2Wiki d1	25.0/5.00	23.5/5.50	23.0/6.45	F2, Rep, +4 learned	20.5/6.37	F2, Rep, +4 learned
MuSiQue d1	7.3/5.00	7.5/6.54	8.0/9.46	F4, +4 learned	8.5/9.05	F4, +3 learned
Wikipedia 25M	23.0/5.00	21.3/5.90	20.8/6.50	F2, Rep, +2 learned	20.5/6.72	F2, Rep, +3 learned

14 Retrospective moderator analysis (RQ3)

We formulated the 26-cell moderator after the decision matrix was available. It is therefore a retrospective exploratory analysis, which we report here in full with every scored variant. A rule specified within this exploratory moderator analysis (early correctness above the median *and* repetition precision below it) reaches 73.1% leave-one-cell-out accuracy versus a 57.7% majority-class base rate, with balanced accuracy 0.69, precision 0.83, and permutation p=0.029. Substituting the second variable changes little: threshold stability gives the same 73.1% (p=0.032) and repeat-signal AUROC drops to 65.4% (p=0.18). Single variables are weaker. Repetition precision

alone reaches 69.2% (p=0.057), early correctness alone reaches 61.5% (p=0.21), and an eight-feature leave-one-out logistic regression reaches 57.7%, exactly its base rate (p=0.46). No selection among these variants was used to choose the reported rule.

Table 15 reports the eight training-chain properties: early correctness, the precision and AUROC of the repetition signal, the saturation gap, the FBG-positive rate, cross-fitted score quality, threshold stability, and training EM. Two unadjusted marginal associations among these eight tests reach $p < 0.05$: the FBG-positive rate ($\rho = +0.44$, p=0.026) and repetition precision ($\rho = -0.42$, p=0.032). Neither survives a correction for eight tests. The rule-level tests are not significant: FBG-only survival p=0.87, "some FBG rule survives" p=0.57, and the deployable-frontier rule reaches 21.4% against an 85.7% base rate (p=0.97). The 1-EM band, won by an FBG rule in one cell, has too few positives to test.

Table 15: Exploratory associations between training-chain properties and learned-rule survival across 26 cells.

	mean property		any FBG		guarded	
Property (training chains only)	**any survives**	**none survives**	ρ	p	ρ	p
early correctness	0.257	0.346	-0.26	0.21	+0.03	0.88
repetition precision	0.955	0.953	+0.07	0.72	-0.42	0.03
repeat-signal AUROC	0.574	0.566	+0.10	0.64	-0.04	0.86
saturation gap (EM)	1.472	1.759	-0.09	0.68	+0.26	0.21
FBG-positive rate	0.136	0.144	-0.07	0.75	+0.44	0.03
cross-fitted ρ	0.362	0.394	-0.22	0.28	+0.28	0.17
threshold stability	0.561	0.570	+0.00	1.00	-0.25	0.21
training EM	32.222	41.481	-0.23	0.27	+0.12	0.56

15 Selection, training size, and feature sensitivity (RQ3)

Table 20 compares joint feature-family ablations: step only, stability only, all retrieval/count proxies, answer features only, no-step, and no-stability. No family is consistently dominant, and simple families often match the full critic on cross-fitted training chains.

The cost-adjusted target variants leave mean EM across the fully instrumented cells at 22.0 to 22.2 and 6.2 to 6.5 calls. In D/F/I order, guarded FBG has tallies 9/3/2 when logged, 10/3/1 at $c = 0.04$, 10/1/3 with discounting at $\gamma = 0.5$, and 9/1/4 with one-step gain. The binary future-improvement classifier's tally is 1/11/2. Table 19 summarizes one-factor sensitivity. The broader 154 selection-protocol configurations vary the threshold grid over 6 or 12 points, the selection band over 0.5/1/2 EM, the folds over 3/5/10 with two cross-fitting seeds, and the certificate-split fraction over 0.20/0.25/0.33.

Logged policies, refits, and signed targets. All tallies here use D/F/I for beaten/frontier/better, matching the main paper's Pareto rule. A better point is undominated and Pareto-improves $B^\star$, so equal EM at strictly lower cost qualifies. Across the 14 instrumented cells, logged guarded FBG is 9/3/2, the full-feature refit with threshold reselection is 11/2/1, and the reduced-feature refit is 11/1/2. Only 8 of 14 full refits recover the logged threshold. These are different fitted policies, not conflicting counts for one policy.

A separate signed-target refit uses $\max_{k>t}[u(a_k, y^\star) - u(a_t, y^\star)] \in [-1, 1]$ and zero at the terminal prefix, with the original features, five-fold chain cross-fitting, and threshold grid. The guarded signed policy gives 11/3/0 versus 11/2/1 for the clipped refit. Mean EM changes from 22.13 to 22.05 and mean calls from 6.44 to 6.26. This isolates clipping on the same logged futures without estimating the value of counterfactual queries. Per-cell results and the refit code are `revision/numerical-audit.json` and `revision/analyze_revision.py`.

Training-size sensitivity. We use two measurements to test whether additional training episodes change the guarded rule's outcome. First, the certificate-split factor doubles as a training-set-size axis, because fractions 0.20/0.25/0.33 leave 96/90/80 policy-training episodes. Across the fourteen fully instrumented cells the guarded rule is Pareto-dominated by the simple family in 9, 9, and 11 of 14 cells at 96, 90, and 80 episodes, with mean guarded EM 22.23, 22.20, and 22.12 at 6.45, 6.34, and 6.35 mean calls. Because that axis is narrow and confounded with calibration-split size, we add a direct subsampling analysis.

`training_size_curve.py` subsamples the policy-training chains to $n \in \{30, 45, 60, 75, 90\}$ episodes, uniformly at random, and re-runs the *identical* frozen pipeline with cross-fitted scores, threshold selection, and replay on untouched questions. We repeat the process 20 times per size and cell, giving 1,400 refits over the fourteen cells and 252 s on CPU with no failed draws. Table 18 reports the size-specific averages. Tripling the training set from 30 to 90 episodes moves the pooled domination rate by 6.1 points, from 79.6% to 73.6%. The trend is not monotone because it rises to 81.1% at n=60, giving a 7.5-point full range. Mean guarded EM increases by 0.14 points (21.98 to 22.12), with a fitted slope of −0.86 points of domination and +0.02 EM per ten extra episodes. The guard stays dominated in roughly three-quarters of draws at every size, and 11 of the 14 cells are dominated in at least half of their draws at *every* size.

Across the 30 to 90 sweep, the guard remains dominated in most refits and mean EM changes by 0.14 points. The full-search power calculation does not measure precision for this learning curve or the matched-cost frontier. Subsampling cannot exclude gains beyond its training range. The fresh same-population curve below tests larger training sets in one cell. These observations establish neither equivalence nor saturation.

Pooled training beyond 90 episodes (post-review, exploratory). All 14 cells share the Qwen2.5-3B backbone, four-search horizon and nine features, so policy-training chains can be pooled across cells without new generation. For each target cell we add the policy-training chains (never certificate or test chains) of donor cells. We use either the same retrieval environment (seven HotpotQA-corpus, three 2Wiki and three MuSiQue cells, with no donor for the Wikipedia 25M cell) or all 14 cells. A donor chain is dropped if its question is in the target's test set (6 to 55 chains per target). Only one trajectory per question is kept, with the target's own first, so cross-fitting folds stay question-disjoint. Cross-fitting, the threshold grid, the 1-EM selection band, the simple family and $B^\star$ are unchanged. Table 16 reports the result. The own-cell rows reproduce the fresh 90-episode refit. Pooling lowers point-level domination, most clearly for FBG-only, while mean EM moves by at most 0.89 points. Mean paired gaps against the matched-cost fixed-budget mixture remain negative. Of the 56 pooled-policy intervals, one is a win (guarded, HotpotQA sample E, same-environment pool, +1.56 [0.17, 3.03] EM), 49 are unresolved and six are losses. These per-cell 95% intervals are unadjusted for multiplicity. The isolated win leaves a family-wise gain unestablished. Pooling also mixes populations, so fresh same-population trajectories are tested separately below. Script and per-cell output: `revision/pooled_training.py` and `revision/pooled-training.json`.

Table 16: Pooled policy training. D/F/I are point dominance verdicts against the target cell's simple family. Δ is the mean paired EM gap to the matched-cost fixed-budget mixture. W/U/L counts per-cell 95% intervals (unadjusted). Fourteen cells per row.

Pool	Episodes	Rule	D/F/I	EM	Calls	Δ	W/U/L
Own cell	90	Guarded	11/2/1	22.13	6.44	−0.67	0/11/3
Own cell	90	FBG-only	10/2/2	21.68	5.79	−0.48	0/11/3
Same environment	90 to 421	Guarded	8/3/3	22.23	6.26	−0.53	1/12/1
Same environment	90 to 421	FBG-only	4/6/4	22.23	6.05	−0.22	0/14/0
All 14 cells	932 to 959	Guarded	9/4/1	22.09	6.01	−0.51	0/11/3
All 14 cells	932 to 959	FBG-only	6/5/3	22.57	6.64	−0.29	0/12/2

Fresh same-population learning curve (post-review). We froze a protocol before generating any trajectory and collected 4,480 new four-search trajectories in the primary HotpotQA corpus cell with the same Qwen2.5-3B digest, prompts, decoding, retriever and corpus. The questions come from the same validation split and are disjoint by ID and text from every archived question. A pilot on ten logged training questions reproduced all prefix answers, queries and evidence. Generation used one sequential local stream for 3.5 hours, with 40,320 calls and no failures. We refit the unchanged pipeline on nested random subsets of 30 to 2,880 training trajectories, 20 subsets per size, with cross-fitting, training-only threshold selection in the 1-EM band, and a training-selected $B^\star$ per subset. Every fit is evaluated on one fixed test set of 2,000 questions, the 400 logged test questions plus 1,600 fresh ones. Table 17 reports the curve. Up to 90 trajectories it agrees with the subsampling analysis. At 30 trajectories FBG-only's gap to the matched-call fixed-budget mixture is −0.78 EM (pointwise 95% CI [−1.12, −0.44]), at 90 it is −0.10 [−0.34, 0.14], also pointwise. From 180 trajectories on, the FBG-only gap is positive and grows from 0.39 to 0.76 EM at 2,880 trajectories. The conditional simultaneous band over 32 contrasts excludes zero at every size from 180, with a lower bound of 0.01 EM at 2,880. Guarded FBG rises from 0.07 EM at 90 trajectories to 0.46 EM at 2,880, but all of its simultaneous bands include zero. Neither rule resolves an accuracy gain over

$B^\star$. At 2,880 trajectories FBG-only reaches 30.75 EM with 6.19 calls against 30.55 EM with 9.0 calls for $B^\star$, a difference of 0.20 EM [−0.35, 0.75]. Fourteen richer state features and the signed target, fitted on the same subsets, do not improve on the nine-feature clipped target. The gain is concentrated on the fresh test questions, whose full-search EM is 32.1 against 24.5 for the logged test sample. The value head is trained on a logged hindsight target and is not a Bellman value. These results come from one cell, one backbone and a four-search horizon, and they are conditional on the fitted policies. They establish a small matched-cost gain for FBG-only with more same-population data. Replication across backbones and search horizons requires separate tests. The protocol, script and output are in `revision/`, under `ls11-learning-curve-protocol.json`, `ls11_learning_curve.py` and `ls11-learning-curve/`.

Table 17: Fresh same-population learning curve in the primary HotpotQA corpus cell. Gap is the mean paired EM difference to the matched-call mixture on the simple family's upper hull, averaged over 20 nested subsets (one distinct fit at 2,880). Sim. is the conditional simultaneous band over 32 primary contrasts. Test set of 2,000 questions.

	Guarded FBG		FBG-only		
n	EM	Gap	EM	Gap	Sim.
30	29.92	−0.06	27.88	−0.78	[−1.29, −0.27]
60	30.02	0.07	28.95	−0.06	[−0.43, 0.31]
90	30.12	0.07	29.64	−0.10	[−0.46, 0.27]
180	30.10	0.13	29.90	0.39	[0.04, 0.74]
360	30.16	0.21	30.28	0.40	[0.06, 0.75]
720	30.07	0.19	30.19	0.45	[0.10, 0.80]
1,440	30.15	0.32	30.37	0.54	[0.04, 1.03]
2,880	30.30	0.46	30.75	0.76	[0.01, 1.50]

Table 18: Training-size sensitivity over fourteen cells and twenty subsamples per size.

Episodes	Draws	Dominated	Mean EM	Mean calls
30	280	79.64%	21.98	6.21
45	280	78.57%	22.02	6.23
60	280	81.07%	22.01	6.26
75	280	77.86%	22.03	6.30
90	280	73.57%	22.12	6.31

Table 19: Guarded-rule sensitivity over eleven configurations per run: the base configuration, a refit with seed 3, and nine one-factor changes.

Run	θ	selected θ set	EM range	calls range	dom.	win	pass
A s7	0.04	0.04, 0.05, 0.07, 0.12	24.2 to 25.0	5.88 to 6.75	3/11	2/11	11/11
A s11	0.07	0.04, 0.05, 0.07, 0.12, 0.2	24.8 to 25.2	5.78 to 6.90	11/11	0/11	11/11
A s19	0.04	0.01, 0.02, 0.04	25.5 to 26.0	6.47 to 7.38	10/11	1/11	11/11
B	0.04	0.02, 0.04, 0.2, 0.3	28.0 to 30.0	5.83 to 7.17	10/11	0/11	1/11
C	0.2	0.2, 0.3	23.2 to 23.2	5.90 to 5.90	0/11	0/11	0/11
D	0.04	0.04, 0.07, 0.2	26.5 to 27.0	5.89 to 6.36	7/11	4/11	11/11
E	0.02	0.01, 0.02, 0.16, 0.2	32.5 to 34.2	5.79 to 8.21	0/11	6/11	7/11
2Wiki d1	0.2	0.2, 0.3	23.5 to 23.5	5.50 to 5.50	11/11	0/11	11/11
2Wiki d2	0.2	0.005, 0.01, 0.02, 0.04, 0.07, 0.12, 0.2	23.2 to 24.8	5.59 to 7.93	11/11	0/11	7/11
2Wiki d3	0.2	0.12, 0.2, 0.3	25.2 to 25.8	5.61 to 5.82	11/11	0/11	11/11
MuSiQue d1	0.2	0.2, 0.3	7.5 to 7.5	6.53 to 6.54	9/11	2/11	0/11
MuSiQue d2	0.2	0.01, 0.02, 0.04, 0.12, 0.2	9.2 to 10.2	6.50 to 7.36	9/11	2/11	11/11
MuSiQue d3	0.2	0.2	10.8 to 10.8	6.42 to 6.42	11/11	0/11	11/11
Wikipedia 25M	0.2	0.2, 0.3	21.2 to 21.2	5.88 to 5.91	11/11	0/11	11/11

Table 20: Guarded and FBG-only results under joint feature-family ablations with frozen selection and test replay.

	full		no step/evid./stab.		answer content only		no stability	
Run	**G**	**V**	**G**	**V**	**G**	**V**	**G**	**V**
A s7	24.8/6.75 F	25.0/6.35 I	25.0/6.46 I	25.3/5.74 I	24.3/6.87 D	21.8/5.29 D	25.0/6.45 I	25.3/5.78 I
A s11	25.0/6.90 D	25.8/6.39 D	24.5/6.06 D	22.8/4.63 F	24.8/6.77 D	20.8/5.06 D	25.3/6.55 D	25.3/5.79 D
A s19	25.5/6.68 D	25.3/6.38 D	25.5/6.88 D	25.0/6.32 D	25.5/8.04 D	24.5/7.24 D	26.0/7.02 D	25.8/6.47 F
B	28.0/5.97 D	27.8/4.23 F	28.8/6.17 D	26.5/4.25 F	28.5/6.17 D	25.3/4.00 F	29.0/6.14 D	28.8/4.32 F
C	23.3/5.90 F	13.5/2.67 F	23.5/5.91 F	13.3/2.67 F	23.5/5.91 F	13.3/2.66 F	23.3/5.90 F	13.5/2.68 F
D	26.5/5.90 D	19.5/3.07 D	26.5/5.90 D	19.5/3.06 D	26.5/5.91 D	19.5/3.06 D	26.5/5.90 D	19.5/3.06 D
E	34.3/7.76 I	33.3/7.51 F	34.5/7.75 I	34.0/7.51 F	34.3/7.63 I	32.3/6.64 F	34.0/8.13 F	33.5/8.04 F
2Wiki d1	23.5/5.50 D	20.3/3.11 F	23.5/5.52 D	19.3/2.83 F	23.5/5.53 D	19.0/2.78 F	23.3/5.59 D	20.0/3.05 F
2Wiki d2	24.0/7.32 D	23.3/7.06 D	24.8/8.04 D	24.5/7.91 D	23.3/7.08 D	20.0/5.84 D	23.8/6.19 D	22.3/5.40 D
2Wiki d3	25.8/5.82 D	25.5/4.07 F	26.0/5.71 D	24.8/3.18 F	25.8/5.70 D	24.0/3.07 F	25.8/5.67 D	24.3/3.15 F
MuSiQue d1	7.5/6.54 D	3.8/3.03 F	7.5/6.55 D	3.3/2.72 F	7.5/6.57 D	3.3/2.71 F	7.5/6.55 D	3.8/3.03 F
MuSiQue d2	9.8/6.89 D	9.8/6.21 D	9.8/6.92 D	10.3/5.85 I	9.8/7.99 D	9.3/6.91 D	9.8/6.91 D	10.3/5.85 I
MuSiQue d3	10.8/6.43 D	0.3/1.00 F	10.8/6.43 D	0.3/1.00 F	10.8/6.43 D	0.3/1.00 F	10.8/6.43 D	0.3/1.00 F
Wikipedia 25M	21.3/5.90 D	18.0/3.13 F	21.3/5.96 D	18.0/3.20 F	21.3/6.08 D	18.8/3.44 F	21.3/5.97 D	18.0/3.21 F

16 Nested refitting and reselection (RQ3)

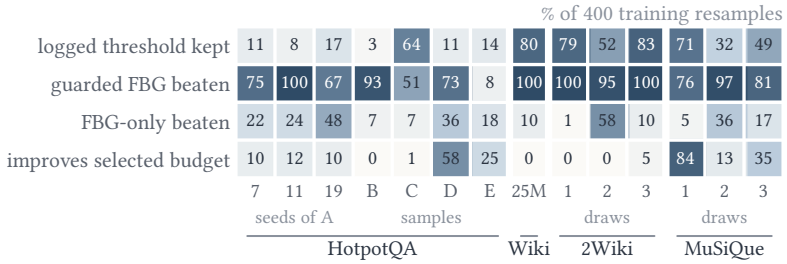

Figure 1: Nested refit/reselection over 400 training-episode resamples per fully instrumented run. Heatmap cells report threshold recovery, simple-policy domination, and guarded-policy wins over reselected $B^\star$.

Figure 1 displays the refit percentages by run. Table 21 gives the numerical contrasts. Here, "dom." is the dominated share and F2/F3/Rep identify the dominating comparator.

Here, θ recovered is the share reselecting the logged threshold. ΔEM is guarded minus full-budget EM with a 95% training-only interval, and C is guarded calls. Across the seven HotpotQA runs, $B^\star$ ranges from 1 to 4 and the interval widths are 0.75 to 2.5 EM. There, FBG-only is dominated less often (7 to 48%) because it is frequently cheapest while losing 1.8 to 11.2 EM. Every resample keeps the questions, logged queries, retrieved evidence, corpus, preprocessing, and decoder realization fixed. It resamples complete training chains, then repeats predictor fitting, threshold selection, and simple-policy selection. The resulting frequencies measure training and selection instability conditional on the logged pipeline.

Table 21: Nested refit/reselection results for the guarded and FBG-only rules.

Run	θ recovered	ΔEM [CI]	C	dom. G	dom. V	G$\succ B^\star$	F2	F3	Rep
Corpus s7, A	11%	-0.07 [-0.50,0.50]	6.24	75%	22%	10%	71%	6%	51%
Corpus s11, A	8%	-0.11 [-0.50,0.25]	6.07	100%	24%	12%	100%	4%	87%
Corpus s19, A	17%	-0.41 [-1.00,0.25]	6.52	67%	48%	10%	0%	18%	55%
Corpus s7, B	3%	-2.98 [-3.75,-1.75]	6.06	93%	7%	0%	91%	2%	24%
Corpus s7, C	64%	-3.10 [-3.50,-1.99]	6.03	51%	7%	1%	0%	1%	49%
Corpus s7, D	11%	+1.01 [0.25,1.50]	6.40	73%	36%	58%	2%	15%	73%
Corpus s7, E	14%	-0.83 [-2.25,0.25]	6.84	8%	18%	25%	0%	0%	7%
Wikipedia 25M	80%	-0.24 [-0.25,0.00]	5.96	100%	10%	0%	100%	0%	92%
2Wiki d1	79%	+0.26 [0.25,0.26]	5.56	100%	1%	0%	100%	0%	90%
2Wiki d2	52%	-1.68 [-2.25,-0.50]	6.08	95%	58%	0%	91%	12%	32%
2Wiki d3	83%	+0.07 [-0.25,0.50]	5.69	100%	10%	5%	100%	0%	29%
MuSiQue d1	71%	-1.83 [-2.00,-1.75]	6.57	76%	5%	84%	32%	0%	76%
MuSiQue d2	32%	-0.73 [-1.00,0.00]	6.78	97%	36%	13%	97%	1%	29%
MuSiQue d3	49%	-1.17 [-1.25,-0.75]	6.55	81%	17%	35%	81%	0%	35%

17 Hierarchical view of the corpus-index runs (RQ2)

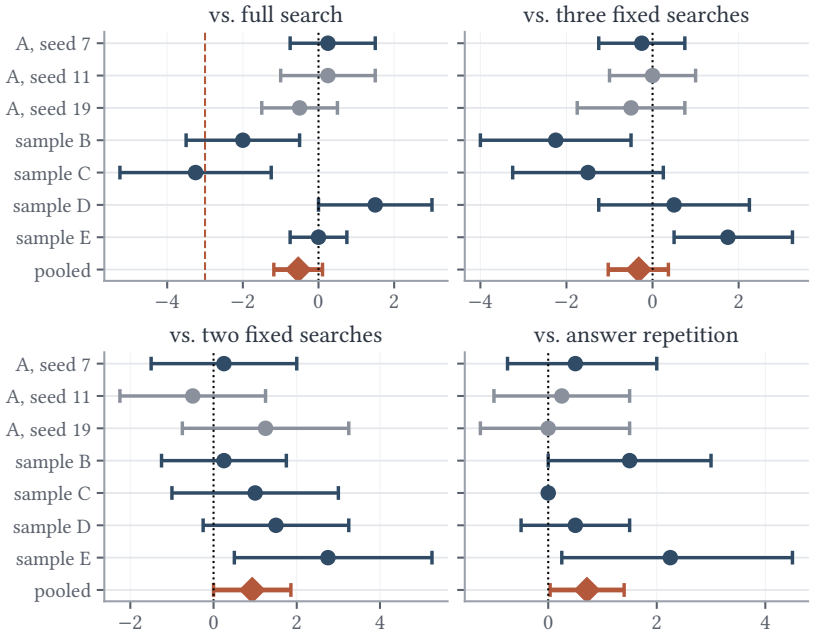

Figure 2: Guarded-FBG contrasts across fixed-question reruns (A) and independent draws (B to E). Bars are 95% CIs, diamonds are pooled estimates, and the dashed line marks the −3 EM margin.

Figure 2 separates fixed-question reruns from independent draws, and Table 22 reports their paired contrasts. Averaged over the three stochastic runs on question sample A the guarded rule reaches an equal-weight 25.17 EM at 6.64 versus 9.00 calls. Over all seven corpus-index runs the equal-weight means are 27.04 EM at 6.67 calls for guarded FBG, 27.57 at 9.00 for the full budget, 27.36 at 7.00 for fixed-3, 26.32 at 5.77 for repetition, and 25.43 at 5.81 for FBG-only. The guarded rule is −0.54 EM against the full budget at 6.67 versus 9.00 calls, −0.32 against fixed-3 at 0.34 fewer calls, and −2.14 for FBG-only against the full budget.

Table 22: Paired EM contrasts across fixed-question reruns, independent question draws, and the question-clustered pooled analysis. SD obs/exp is the observed standard deviation of draw means divided by the expected standard deviation under question sampling, both in EM points.

Contrast	A	B	C	D	E	SD obs/exp	Pooled [CI]
G − full	+0.00	-2.00	-3.25	+1.50	+0.00	1.87/0.73	-0.54 [-1.18,0.11]
G − fixed-3	-0.25	-2.25	-1.50	+0.50	+1.75	1.59/0.80	-0.32 [-1.03,0.37]
G − fixed-2	+0.33	+0.25	+1.00	+1.50	+2.75	1.02/0.95	+0.93 [0.00,1.86]
G − repetition	+0.25	+1.50	+0.00	+0.50	+2.25	0.94/0.69	+0.71 [0.04,1.40]
V − full	+0.00	-1.75	-13.00	+0.75	-1.00	5.67/1.15	-2.14 [-3.06,-1.24]
V − fixed-3	-0.25	-2.00	-11.25	-0.25	+0.75	4.94/1.16	-1.93 [-2.87,-1.00]
V − fixed-2	+0.33	+0.50	-8.75	+0.75	+1.75	4.32/1.05	-0.68 [-1.57,0.21]

18 Per-cell accuracy-cost frontiers

Figure 3 plots every fully instrumented core cell to show whether the headline frontier verdict is localized to one dataset or model. The panels cover five HotpotQA samples, the Wikipedia 25M cell, and the 2WikiMultihopQA and MuSiQue draws.

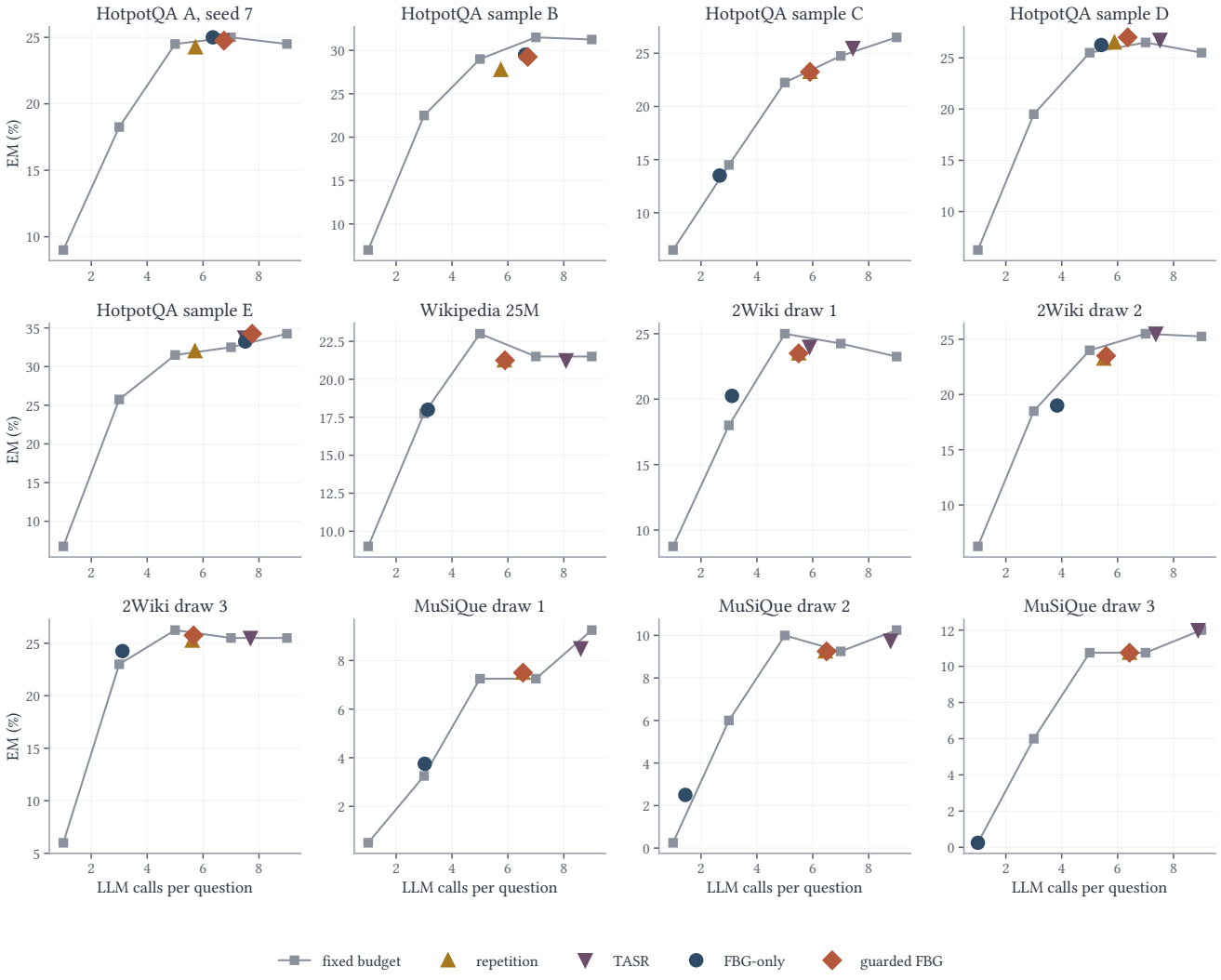

Figure 3: Accuracy-to-call frontiers for fixed budgets, answer repetition, full TASR, and learned policies in fourteen fully instrumented cells.

19 Full corpus-confirmation Pareto tables

In Table 23, S and C denote searches and exact LLM calls. In is input tokens per question. Wall is LLM-plus-retrieval seconds per question. Pareto membership uses EM and calls.

Table 23: Primary corpus-wide confirmation (n=400, question sample A, seed 7).

Method	EM	S	C	In	Wall	Pareto
Full budget	24.5	4.00	9.00	1821	5.5	no
Fixed-2	24.5	2.00	5.00	775	3.1	yes
B^*=3	25.0	3.00	7.00	1282	4.3	no
Answer repetition	24.2	2.36	5.72	1011	3.6	no
TASR (full logit rule)	24.8	2.84	6.68	1233	4.0	no
FBG + NLI + repeat guard	24.0	2.52	6.03	1082	3.8	no
Repeat-conditioned gain	24.5	2.89	6.78	1288	4.2	no
FBG value only	25.0	2.67	6.35	1143	3.9	yes
FBG + repeat guard	24.8	2.88	6.75	1261	4.2	no

20 Call-only full-budget panel

Tables 24 and 25 report the call-only backbone and dataset panels. CB is closed-book EM. ΔEM is the paired FBG-only-minus-ReAct difference with a bootstrap 95% CI. S is searches per question, and ΔS is the saving from ReAct's four searches. Table 26 places adaptive-retrieval systems alongside the frozen controller. These system-level results use searches as the common cost coordinate, retaining differences in retrieval and generation design.

Table 24: Frozen-protocol HotpotQA results across nine backbones versus the full budget in per-question-context cells.

Backbone	n	CB	ReAct EM [CI]	FBG-only EM [CI]	ΔEM [CI]	S	ΔS
Claude Haiku 4.5	140	19	52.1 [43.6,60.7]	50.7 [42.8,59.3]	-1.4 [-5.0,2.1]	1.88	53%
Claude Sonnet 4.5	140	8	52.1 [43.6,60.7]	53.6 [45.0,62.1]	+1.4 [0.0,3.6]	2.85	29%
GPT-oss-120B	140	28	57.1 [48.6,65.7]	55.7 [47.1,64.3]	-1.4 [-4.3,1.4]	2.49	38%
Amazon Nova Lite	140	16	54.3 [45.7,62.9]	52.9 [45.0,61.4]	-1.4 [-3.6,0.0]	2.94	27%
Qwen2.5-3B	140	10	38.6 [30.7,46.4]	39.3 [31.4,47.9]	+0.7 [-3.6,5.0]	2.56	36%
Llama-3.2-3B	140	16	39.3 [31.4,47.9]	37.9 [30.0,45.7]	-1.4 [-6.4,3.6]	2.21	45%
Command-R-35B	100	25	51.0 [41.0,61.0]	53.0 [43.0,63.0]	+2.0 [0.0,5.0]	2.88	28%
Kimi K2.5	140	28	59.3 [51.4,67.1]	57.9 [49.3,66.4]	-1.4 [-5.0,2.1]	1.97	51%
MiniMax M2.5	140	21	55.7 [47.9,64.3]	53.6 [45.7,62.1]	-2.1 [-7.1,2.9]	1.71	57%

Table 25: Frozen-protocol replication across datasets in per-question-context cells. Ref = full-budget ReAct.

			ReAct		FBG-only		
Dataset	Backbone	n	EM	S	EM	S	ΔS
2Wiki	Qwen2.5-3B	140	28.6	4.0	28.6	2.60	35%
MuSiQue	Qwen2.5-3B	140	13.6	4.0	12.1	2.04	49%
MuSiQue	Llama-3.2-3B	100	18.0	4.0	16.0	2.43	39%

Table 26: Adaptive-retrieval baselines versus frozen FBG-only on the HotpotQA Qwen2.5-3B baseline-implementation cell (n=140).

Method	EM	S	ΔS
Self-RAG (released 7B)	7.9	0.56	86%
FLARE-direct	34.3	1.66	58%
SKR_{knn}	27.1	1.00	75%
Adaptive-RAG (trained)	37.9	3.81	5%
ReAct (full budget)	38.6	4.00	0%
FBG-only	**39.3**	**2.56**	**36%**

21 Expanded deterministic reference family

This retrospective check adds evidence-growth saturation to all fixed budgets and answer repetition in 14 fully instrumented cells. Thresholds in $\{0, 1, 2, 3\}$ are selected on training chains only, using the same one-EM admissibility band and lower-call preference as the earlier diagnostic. The control stops after at least one search when newly retained units do not exceed the selected threshold. It uses already logged features and makes no additional LLM calls. Training and test question identifiers are disjoint.

We recompute the observed simple-policy hull for calls, input tokens, and a wall-time sensitivity coordinate. The latter adds the measured local CPU cost of applying the control (0.26 to 0.34 microseconds per question, 100 repetitions) to historical model and retrieval timings. This sensitivity coordinate combines local CPU overhead with historical execution timings. It is distinct from an end-to-end latency measurement.

Table 27 reports mean paired EM gaps across cells. A single 84-contrast family covers 14 cells, two FBG policies, and three cost coordinates. We use 10,000 Gaussian multiplier draws, sharing each question's weight across cells and coordinates, and the 95th percentile of the maximum absolute standardized contrast (3.140). These exploratory bands condition on fitted policies and observed hull vertices and mixture weights. They do not include training or frontier-reselection uncertainty. All 84 bands include zero, which does not establish equivalence. The original frozen 26-cell analysis remains unchanged.

A second retrospective control targets query repetition. After search $t \geq 2$, it stops when the token Jaccard similarity between query t and any earlier query in the episode reaches τ. We fixed the grid $\tau \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, the same one-EM admissibility band, and a lower-call preference before evaluating test questions, and selected τ on training chains only. The control reads logged queries and adds no LLM calls. Across the 14 cells, it matches the EM of the selected fixed budget on average (mean difference 0.00 EM, no cellwise 95% interval excluding zero) at 0.23 fewer calls. Simple policies dominate it in 7 cells, and it improves on the selected budget in 2. With this control added to the reference family, all 84 conditional matched-cost bands for the two FBG policies still include zero, with mean guarded gaps of −0.69, −0.72, and −0.69 EM for calls, input tokens, and the wall-time proxy.

Table 27: Expanded-reference matched-cost sensitivity.

Policy	Cost coordinate	Mean EM gap	Win/unresolved/loss
Guarded	Calls	−0.68	0/14/0
Guarded	Input tokens	−0.72	0/14/0
Guarded	Wall-time proxy	−0.69	0/14/0
FBG-only	Calls	−0.07	0/14/0
FBG-only	Input tokens	−0.09	0/14/0
FBG-only	Wall-time proxy	−0.05	0/14/0

22 Fresh eight-search extension

We prospectively froze a bounded extension on November 18, 2026 before collecting model outputs. From HotpotQA distractor validation, we exclude all question identifiers recoverable from the archived training and test

records, then shuffle with seed 212613 and select 120 training and 200 test questions. The two sets are disjoint. The artifact records their identifiers, the dataset checksum, and the frozen protocol. Policy fitting and five-fold chain cross-fitting use 90 training chains, with 30 reserved and unused for this stopping analysis, preserving the original calibration split.

Claude Haiku 4.5 (`claude-haiku-4-5-20251001`, temperature zero) produces an initial answer and eight search-query/answer pairs per question. Queries adapt to accumulated evidence using the submitted prompts. Retrieval uses SQLite FTS5/BM25 over 269,558 deduplicated sentence units from the public validation contexts, with no gold-support filtering. We replay stopping at horizons four and eight on these same fresh trajectories. This extension tests a longer replay horizon on fresh Haiku trajectories. Joint query-and-stop learning and replication across the original nine backbones remain separate experiments.

At each horizon, we refit the original clipped-FBG and current-error heads, select the guarded threshold by training-chain cross-fitting, and use that threshold for both guarded and FBG-only stopping as in the original protocol. The reference family includes every fixed budget, answer repetition, and training-selected evidence growth. Both training-selected fixed budgets use the full horizon. Table 28 reports every horizon/policy combination. D and I are descriptive categories, not inferential conclusions.

Table 28: Fresh-question longer-horizon check ($n = 200$).

Horizon	Policy	$B^\star$	EM	Calls	Class	Gap	Simultaneous CI
4	Guarded	4	53.50	5.71	D	-0.85	$[-4.24, +2.53]$
4	FBG-only	4	55.00	4.91	I	$+1.69$	$[-2.10, +5.47]$
8	Guarded	8	55.00	8.24	D	$+0.00$	$[-2.43, +2.43]$
8	FBG-only	8	55.50	8.19	I	$+0.50$	$[-2.22, +3.22]$

At horizon eight, guarded FBG reaches 55.0 EM at 8.24 calls and FBG-only reaches 55.5 EM at 8.19 calls. Full-budget EM is 53.5 at horizon four and 55.0 at horizon eight. Their paired difference is 1.5 EM points, with an unadjusted percentile-bootstrap 95% interval $[0.0, 3.5]$. The prefix oracle is 61.5 EM at both horizons. The extra searches thus improve the final answer on average in this sample without raising the observed oracle ceiling.

The 12-contrast exploratory family covers two horizons, two policies, and calls, provider-reported input tokens, and measured request-plus-retrieval latency. We use 10,000 shared-question Gaussian multiplier draws and a maximum absolute standardized-contrast critical value of 2.428. The bands condition on fitted policies and observed hull vertices and mixture weights, without frontier reselection or training uncertainty. All 12 contain zero. In particular, the eight-search equal-call FBG-only gap is 0.50 EM points with a conditional simultaneous interval $[-2.22, 3.22]$. The interval leaves both a gain and a loss plausible. Resolving the longer-horizon effect requires greater precision.

Collection completed all 320 trajectories and 5,440 model requests with zero recorded failures. A separate connectivity probe adds one request. Provider usage totals 1,601,333 input and 86,545 output tokens including the probe, giving an estimated list-price cost of $2.034058 at $1 and $5 per million tokens, respectively. This estimate is not an invoice. Eight concurrent workers affect latency, so these timings are specific to this collection. The record includes prefix costs, queries, answers, retained evidence, features, selected thresholds, and all outcomes.

23 Retrieval and predictor details

HotpotQA uses BM25 over sentence units from the distractor context or one shared validation index [9, 13]. The FEVER condition retrieves 50 candidates with SQLite FTS5 and reranks them to three title-prefixed sentences with a frozen MS MARCO MiniLM cross-encoder. The 2Wiki and MuSiQue cells use their shared validation corpora and credit answer aliases [4, 11]. The external cell indexes the complete MultiHopRAG news corpus with the same FTS5 and BM25 code path. It retains inference queries only. We excluded comparison and temporal yes-or-no items because the answer prompt and alias-aware EM were designed for short entities. We also excluded null items because they require a distinct abstention target. These exclusions were declared before test evaluation.

Predictor settings. Every prefix from a full-budget logged trajectory supplies graded FBG, binary future-improvement, and answer-error labels. Where annotated support exists, it also supplies an evidence-coverage

label. Our FBG head is a gradient-boosted regressor with 80 trees of maximum depth 2. Our error head is logistic regression with `max_iter=1000`. We fit the binary future-improvement and gold-support classifiers with the same chain-level cross-fitting discipline. Token F1 provides the two future-gain targets, while EM remains our primary policy metric. The base LLM stays frozen and each small predictor trains on CPU.

References

[1] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *International Conference on Learning Representations (ICLR)*.

[2] Zhe Dong, Fang Qin, and Manish Shah. 2026. When Does Learning to Stop Help? Configuring Risk-Controlled Early Exit for Reasoning-Model Services. arXiv:2606.30852 doi:10.48550/arXiv.2606.30852

[3] Yonatan Geifman and Ran El-Yaniv. 2017. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[4] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*.

[5] Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[6] Adrian Kieback, Uyiosa Philip Amadasun, Aman Chadha, and Aaron Elkins. 2026. TASR: Training-Free Adaptive Stopping for Iterative Retrieval. arXiv:2606.13814 Agent4IR Workshop @ KDD 2026.

[7] Minghan Li, Junjie Zou, Xinxuan Lv, Chao Zhang, and Guodong Zhou. 2026. 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)*. 25846–25862.

[8] Jaewan Park, Solbee Cho, and Jay-Yoon Lee. 2025. Stop-RAG: Value-Based Retrieval Control for Iterative RAG. *arXiv preprint arXiv:2510.14337* (2025). NeurIPS 2025 MTI-LLM Workshop.

[9] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* 3, 4 (2009), 333–389.

[10] Daeyoung Roh and Donghee Han. 2026. HALT: Verification-Aware Stopping for Retrieval-Augmented Search Agents. arXiv:2608.02009 doi:10.48550/arXiv.2608.02009 Accepted to Findings of EMNLP 2026.

[11] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics* 10 (2022), 539–554.

[12] Diji Yang, Linda Zeng, Jinmeng Rao, and Yi Zhang. 2025. Knowing You Don't Know: Learning When to Continue Search in Multi-round RAG through Self-Practicing. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*. ACM. doi:10.1145/3726302.3730018

[13] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.