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

Xuan-Bach Le*†
Ho Chi Minh City University of Technology (HCMUT), VNU-HCM
Ho Chi Minh City, Vietnam
lexuanbach@hcmut.edu.vn

Phat T. Tran-Truong*
Ho Chi Minh City University of Technology (HCMUT), VNU-HCM
Ho Chi Minh City, Vietnam
phatttt@hcmut.edu.vn

Son Ha Xuan
The Business School
RMIT University
Ho Chi Minh City, Vietnam
ha.son@rmit.edu.vn

Duy Nguyen
University of Economics Ho Chi Minh City
Ho Chi Minh City, Vietnam
duynguyen.726202321440@st.ueh.edu.vn

Abstract

An extra search can uncover a missing fact or lead an agent away from a correct answer. We test whether learned stopping rules improve accuracy over simple rules at comparable cost. We separate useful stopping opportunities, predictive signal quality, and the value of acting on those signals. We compare four signals on recorded runs of at most four searches. Future-best gain (FBG) measures the largest improvement available later in a run. We use its predicted value alone or with a guard that also requires a repeated answer. We select thresholds and simple baselines on development questions, then freeze them before testing. Across 26 core settings, simple policies outperform guarded FBG in accuracy or cost without worsening the other in 15 settings. Guarded FBG improves on the selected fixed budget by this criterion in five. Neither FBG policy shows a statistically resolved accuracy gain at equal call counts across 18 settings (14 core and four crossed-backbone). In our training-size study, small FBG-only gains emerge from 180 training runs in one setting. Choosing stopping points with hindsight improves full-search exact-match accuracy by 4.84 percentage points on average across 14 fully measured settings. Our equal-cost comparisons test whether learned rules capture these exits. The results identify training size, search budget, and tuned simple alternatives as central to evaluating adaptive stopping.

CCS Concepts

• **Information systems** → **Retrieval models and ranking**; **Web searching and information discovery**; • **Computing methodologies** → **Intelligent agents**.

*Both authors contributed equally to this research.
†Corresponding author.

This work is licensed under a Creative Commons Attribution 4.0 International License.
WSDM '27, Hong Kong, Hong Kong
© 2027 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-3117-4/2027/02
https://doi.org/10.1145/3862322.3863667

Keywords

iterative retrieval, search agents, adaptive stopping, future-best gain, evaluation audit, efficiency

ACM Reference Format:
Xuan-Bach Le, Phat T. Tran-Truong, Son Ha Xuan, and Duy Nguyen. 2027. A Selection-Aware Audit of Learned Stopping in Short-Horizon Iterative Retrieval. In *Proceedings of the Twentieth ACM International Conference on Web Search and Data Mining (WSDM '27), February 15–19, 2027, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3862322.3863667

1 Introduction

A large language model (LLM) agent can search, read the results, and revise its answer [20, 26, 36, 41].[1] Each search can help or harm the answer. A new result may fill a gap, repeat what the agent knows, or distract it from a correct answer [9, 12, 16]. Each search also costs calls, tokens, and time. The agent must therefore decide whether to stop before it knows what the next search will bring.

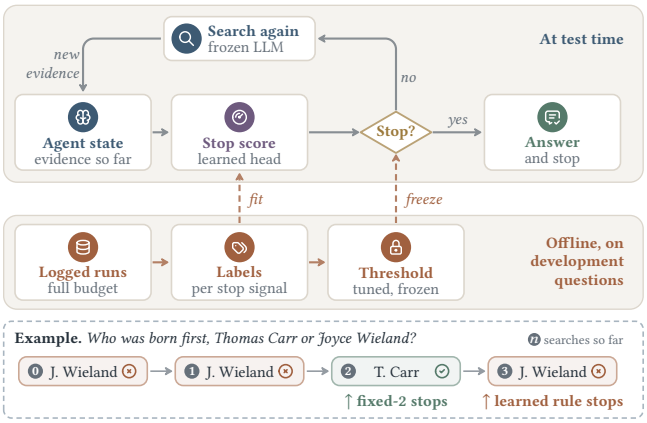

Figure 1: Learned stopping at test time (top), offline fitting and threshold selection (middle), and a logged trajectory where the learned rule misses a useful exit (bottom).

[1] Artifact: https://github.com/lexuanbach/wsdm2027-learned-stopping-audit

A useful exit can be easy to miss. The bottom strip of Figure 1 shows a recorded example from HotpotQA: "Who was born first out of Thomas Carr and Joyce Wieland?" The agent answers Joyce Wieland before searching and after one search, correctly switches to Thomas Carr after two, then returns to Joyce Wieland after three. Our learned rule with a repetition guard stops at the third search and gives the wrong answer. A fixed two-search budget is cheaper and correct. Our task is to select the useful exit from the evidence available at that point.

The middle band fits a score and threshold on complete development runs. The top band applies them to new questions. In this example, the guard remembers early repetition but waits for the score to cross its threshold. Later answers supply training labels unavailable at deployment.

What should a learned rule beat? We compare learned rules with development-selected fixed budgets and repetition. We distinguish the cost of answering once at the end from generating the intermediate answers that a stopping rule monitors.

Our contribution. We introduce a *selection-aware audit*: select learned thresholds and simple baselines on development questions, freeze these choices, then test on new questions. This comparison isolates the value of choosing an exit for each question beyond a single tuned budget.

We train four signals on the same recorded runs. Future-best gain (FBG) measures how much the best later answer improves on the current one, treating no improvement as zero. We learn to predict this gain from the run so far. The other signals predict whether any improvement occurs, whether the current answer is wrong, and whether the retrieved text covers the supporting evidence. None requires token probabilities from the answer model.

Our hindsight oracle improves exact-match (EM) accuracy, the percentage of answers matching the reference, over full search in all 14 fully measured core settings, averaging 4.84 percentage points. Simple policies beat guarded FBG in accuracy or cost without worsening the other in 15 of 26 settings. Neither FBG policy has a statistically resolved gain at equal call counts across the 18-setting comparison. Current-error scores better identify risky answers, but their policy gains are below 0.2 EM and none survives max-t simultaneous correction. Testing on new question samples can also change whether a rule meets the allowed accuracy loss, even when rerunning generation on the same questions gives agreement.

Our core study fits small predictors separately in each setting and evaluates their choices on recorded runs of at most four searches. After related work (Section 2), Section 3 explains the signals and stopping rules. Section 4 defines the comparisons needed to measure their value. Section 5 examines tradeoffs between accuracy and cost, replication, signal quality, and guarantees for stopped answers. Section 6 sets out the practical implications and evaluation scope.

2 Related Work

Learned values and stop decisions. Stop-RAG learns stop and continue values from full-budget retrieval trajectories [28]. LearnStop compares learned reasoning-prefix exits with a static token budget, certifies both with Learn-then-Test at a lost-correct risk bound, and selects the cheaper certified candidate [13]. We study external retrieval on logged query trajectories, comparing EM at matched calls, tokens, and time against fixed-budget and repetition frontiers. We test replication on independent question samples and account for multiple comparisons with simultaneous intervals. Concurrent work estimates task-level value of information across retrieval, decomposition, and answer commitment [14]. FBG labels one logged continuation rather than counterfactual search actions. Matched-cost comparisons test whether that logged-future signal yields a useful stopping policy.

Confidence, stability, and evidence sufficiency. Think Straight, Stop Smart (TSSS) uses retriever repetition, whereas TASR combines answer stability with calibrated token probabilities [7, 21]. We run full TASR when native answer-token probabilities are available and use answer repetition as a stability baseline that needs no token probabilities elsewhere. SIM-RAG and S2G-RAG learn whether retrieved context is sufficient for answering or whether evidence gaps remain [23, 40]. HALT instead tests whether accumulated evidence covers generated hop claims [31]. Our current-error predictor estimates whether the answer is wrong. Our gold-support classifier predicts evidence coverage using labels from training annotations. At test time, neither sees those annotations. We also test claim coverage in six settings using substitute verifiers compatible with our hardware. These comparisons measure the stopping value of each signal.

Our core deterministic reference contains fixed budgets and answer repetition. Supplementary checks add training-selected evidence-growth saturation and query repetition. These declared families define each measured frontier. Retrieval-score saturation, semantic novelty, grounding thresholds, and controller cascades provide additional standalone rules for future tests. The supplement compares the closest methods by signal, actions, requirements, matching status, and outcome.

Conditional retrieval with different action spaces. Self-RAG uses reflection tokens, FLARE triggers retrieval from token uncertainty, SKR estimates self-knowledge, and Adaptive-RAG routes questions by complexity [6, 18, 19, 39]. DRAGIN and SeaKR provide further uncertainty-driven retrieval mechanisms [34, 42]. Their retrievers, training objectives, and generation schedules differ from our prefix replay. We evaluate these adapted systems separately to distinguish system-level performance from stopping decisions on shared runs.

Agents, efficiency, and calibrated prediction. ReAct, WebGPT, Toolformer, IRCoT, and Search-R1 illustrate how language models can interleave reasoning with external actions [20, 26, 32, 36, 41]. Efficiency methods such as FrugalGPT, AutoMix, and CALM allocate model or token budgets [1, 8, 33]. We instead control an external retrieval interaction whose output can help or harm the answer. Our optional certificate draws on selective classification and conformal risk control [3, 15, 30, 37]. It certifies a subset after the stopping time and does not select the stopping action. Value of information, sequential testing, and optimal stopping motivate the decision problem [10, 17, 38]. We measure policy utility through observed frontiers and paired inference. Optimality requires additional assumptions about future observations and actions.

3 Learned Stopping Signals

Section 3.1 defines the search budget. Sections 3.2 and 3.3 define prediction targets. Section 3.4 turns scores into policies, and Section 3.5 adds a post-stop certificate.

3.1 Stopping under a search budget

An episode is one attempt to answer a question or information goal g, with at most B searches. A *prefix* is the run so far, after $t = 0, \ldots, B$ searches. It contains retained evidence E_t, with duplicates removed, and remaining budget $b_t = B - t$. These form the state $s_t = (g, E_t, b_t)$. The fixed answer generator produces $a_t = A(g, E_t)$ at each prefix. The next search returns o_{t+1}, giving $E_{t+1} = \text{dedup}(E_t \cup o_{t+1})$. Our stopping rule uses past answers to detect changes and repetition. Its input is therefore $x_t = (s_t, a_{0:t})$, where $a_{0:t}$ contains all answers so far.

Let $y^\star$ be the reference answer and let $u(a, y^\star) \in [0, 1]$ denote answer quality. Token F1 measures overlap with the reference and supplies the training target. Exact match (EM) measures fully matching answers on a 0 to 100 scale. Let $U(x_t) = \mathbb{E}[u(a_t, y^\star) \mid x_t]$ be expected quality given the current input. A stopping policy π should balance this quality against search cost:

$$\max_{\pi} \ \mathbb{E}_\pi[U(x_T) - c_s T], \tag{1}$$

where $c_s > 0$ is a conceptual search price and $T \le B$ is the stopping time. Thus Eq. 1 rewards good answers and penalizes extra searches. In the experiments, we choose stopping thresholds on development runs and measure calls, tokens, and wall time. A threshold determines the observed tradeoff between accuracy and cost without estimating the conceptual price c_s.

3.2 Future-best gain

We collect full-budget trajectories with a fixed-budget ReAct logging policy. A trajectory records the chosen queries, retrieved evidence, and answer at every prefix:

$$\xi = (s_0, a_0, o_1, s_1, a_1, \ldots, o_B, s_B, a_B).$$

Future-best gain asks how much better the best later answer is. Negative differences are clipped to zero by $[\cdot]_+$. For $t < B$, it is

$$\text{FBG}_t = \left[\max_{k \in \{t+1, \ldots, B\}} u(a_k, y^\star) - u(a_t, y^\star) \right]_+, \qquad \text{FBG}_B = 0. \tag{2}$$

For a current token F1 of 0.6 and later scores 0.4, 0.9, and 0.7, FBG is 0.3: the next search hurts, but a later answer improves. If every later score is at most 0.6, FBG is zero. FBG measures one recorded continuation, not the alternative queries and outcomes needed for a Bellman value or calibrated value of information. Clipping also removes a distinction: harmful and neutral continuations both receive zero. The original cost-adjusted, discounted, and one-step variants retain this nonnegative target. We isolate the effect of clipping in a retrospective refit using $\max_{k>t}[u(a_k, y^\star) - u(a_t, y^\star)] \in [-1, 1]$, without clipping (and zero at the terminal prefix). We generate fresh continuations from the same stopping points to test whether another future improves on the recorded path.

Our FBG predictor estimates $\widehat{\text{FBG}}_t \approx \mathbb{E}[\text{FBG}_t \mid x_t]$. Its score can stay positive even when the next retrieval alone does not help.

We also define the binary future-improvement label

$$c_t = \mathbf{1}[\text{FBG}_t > 0]. \tag{3}$$

The classifier estimates $\Pr(c_t = 1 \mid x_t)$: the chance that a later logged answer improves. The label records improvement rather than the correct stopping action.

3.3 Current correctness and sufficiency

We compare FBG and binary future improvement with two current-state targets. The current-error predictor, or *error head*, $\hat{r}(x_t)$ estimates the probability that a_t is incorrect. This target asks whether the current answer is usable, whereas the two future-gain targets ask whether the logged continuation contains a better answer.

The gold-support classifier instead asks whether the retrieved text contains enough of the annotated supporting evidence. Let $z_t = 1$ when the lowercased alphanumeric tokens in E_t cover at least 0.8 of the annotated support-sentence tokens, and let $z_t = 0$ otherwise. We train $\hat{p}_{\text{suff}}(x_t)$ on z_t. The annotated (gold) support supplies training labels but is unavailable at deployment. We test the nine controller features and a declared twelve-feature lexical variant. Covering support, answering correctly, and benefiting from another search are distinct prediction targets.

3.4 Features, predictors, and policies

Operational features. At prefix t, all predictors receive nine features available without model token probabilities. The first two count searches used and distinct retained retrieval units. Question-term coverage is the fraction of lowercased alphanumeric question tokens that occur in E_t. New-evidence mass counts units added by the latest retrieval after exact-string deduplication. Answer change compares normalized answers after lowercasing and removal of punctuation and articles. Answer length is the number of whitespace tokens. Grounding is the fraction of normalized answer tokens found in E_t, with value zero for an empty answer or empty evidence. A two-prefix stability indicator marks consecutive normalized-answer agreement. The final feature counts distinct normalized answers in $a_{0:t}$. These definitions make empty-prefix behavior deterministic and keep the interface identical across backbones.

Predictors. We train small prediction models on complete runs, using the same cross-fitting procedure (Section 4) for every target. Future-gain targets use token F1, whereas policy evaluation uses EM. The LLM remains frozen. The supplement gives the learner settings.

FBG policies. The FBG-only policy continues while $\widehat{\text{FBG}}_t \ge \theta$ and stops at the first crossing $\widehat{\text{FBG}}_t < \theta$ or at B. Guarded FBG stops once both a score crossing and an answer repetition have occurred. The rule remembers each event, so they need not occur together. If the answer repeats at $t = 1$, changes at $t = 2$, and the FBG score crosses at $t = 3$, the policy stops at $t = 3$ because it retains the earlier repetition event. Prefix answers are already generated for the features, so the guard adds no call under our answer-at-every-prefix schedule.

Algorithm 1 stores whether each event has occurred and returns the stopped answer with an optional reliability certificate.

Algorithm 1: FBG-only and repeat-guarded retrieval control with an optional post-stop certificate.

input : goal g, frozen LLM, scores $(\widehat{\mathrm{FBG}}, \hat{r})$, threshold θ, rule $\rho \in \{\mathrm{FBG}, \mathrm{guarded}\}$, certificate cutoff q_α, budget B

output : stopped answer a_t and certificate flag

$E_0 \leftarrow \varnothing$, $\ell_{\mathrm{score}}, \ell_{\mathrm{rep}} \leftarrow$ false **for** $t \leftarrow 0$ **to** B **do**
  $a_t \leftarrow \mathrm{LLMAnswer}(g, E_t)$
  $\ell_{\mathrm{score}} \leftarrow \ell_{\mathrm{score}} \vee [\widehat{\mathrm{FBG}}(g, E_t, B-t, a_{0:t}) < \theta]$
  $\ell_{\mathrm{rep}} \leftarrow \ell_{\mathrm{rep}} \vee [t > 0 \wedge \mathrm{norm}(a_t) = \mathrm{norm}(a_{t-1})]$ **if** $t = B \vee [\ell_{\mathrm{score}} \wedge (\rho = \mathrm{FBG} \vee \ell_{\mathrm{rep}})]$ **then**
    break
  $q_t \leftarrow$ LLM-proposed query given (g, E_t)
  $E_{t+1} \leftarrow \mathrm{dedup}(E_t \cup \mathrm{search}(q_t))$
return $\big(a_t, \mathbf{1}[q_\alpha \neq \perp \wedge \hat{r}(g, E_t, B-t, a_{0:t}) \le q_\alpha]\big)$

Alternative policies change the score and crossing rule within this loop.

Alternative policies and selection. The binary future-improvement policy stops at the first nonterminal prefix whose predicted probability is below its threshold. The error-head gate stops when predicted current-answer error is at most its threshold, and the gold-support policy stops when predicted evidence coverage is at least its threshold. All three may stop at $t = 0$ and default to B if no threshold is crossed.

We select each learned policy's threshold from cross-fitted development predictions, choosing the cheapest candidate within one EM point of full-budget accuracy. If no point enters that band, we choose the highest-EM point and break ties by cost. On the same development questions, we select the fixed comparator as the smallest budget attaining maximum development EM. We call this fixed budget $B^\star$. We also compare with mixtures that assign different simple policies to different questions. Their best accuracy at each cost forms the *simple-policy frontier*. Within each signal family, primary cells share the crossing direction, threshold grid, cross-fitting, and selection code. Different score scales require different grids. Certification follows stopping and never changes T.

3.5 Post-stop risk certification

We use post-stop certification to identify answer subsets whose error rate meets a chosen target. After fixing the rule and error score $\hat{r}$, we use a separate calibration set of m episodes, each contributing its stopped input and error label (x_{T_i}, e_i), with $e_i = 1$ for an incorrect answer. Selection with guaranteed risk (SGR) [15] starts with answers assigned low error scores and considers increasingly permissive cutoffs. It bounds their error rate with an exact binomial upper confidence bound at level $\delta/\lceil \log_2 m \rceil$. It returns a cutoff q_α only when a nonempty subset's bound is at most the target error rate α. Otherwise, we set $q_\alpha = \perp$ and report zero certificate coverage. At each fixed $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$, SGR uses $\delta = 0.10$ and the guarantee below holds separately.

Proposition 1 (Inherited SGR risk bound [15]). *For fixed α, assume independent, identically distributed calibration and test episodes, with the stopping rule and score fixed before calibration labels. With probability at least $1 - \delta$ over calibration, the nonempty subset returned by SGR satisfies*

$$\Pr[e = 1 \mid \hat{r}(x_T) \le q_\alpha] \le \alpha.$$

If no candidate meets the corrected bound, no certificate is returned.

The bound controls the error rate among accepted answers across episodes, with a population-level guarantee rather than a question- or subgroup-level guarantee. The supplement gives a proof sketch, its reliance on the SGR theorem, and certificate acceptance rule.

4 Evaluation Protocol

Evaluation populations. A *cell* is one evaluation setting with a fixed dataset, corpus and index, retrieval system, answer model, question sample, and decoding condition. Table 1 summarizes the settings. Our 26-cell core audit contains 12 call-only cells and 14 fully instrumented cells with prefix-level call, token, and wall-time records. Equal-cost comparisons add four cells with a different answer model and question sample to those 14 (18 total). One external cell uses MultiHopRAG inference questions and a news-article corpus [35].

Within a cell, all stopping policies replay the same full-budget runs, changing only their chosen exit. A *same-question rerun* changes the decoding seed while holding questions fixed. An *independent question draw* resamples both training and test questions and repeats fitting and selection. A *crossed-backbone cell* changes the answer model on a new draw. The external cell is a separately frozen news-corpus draw.

Tasks, retrievers, and backbones. Call-only 2Wiki and MuSiQue cells use per-question distractor contexts. Their instrumented cells use the shared indices in Table 1. We use sparse retrieval over shared sentence or paragraph indices, with a frozen cross-encoder for Wikipedia 25M. The external news cell retains inference queries. We exclude comparison, temporal, and null items before evaluation because they require different answer or abstention targets. The supplement details corpus construction and these eligibility choices.

We use nine frozen answer models, or backbones, from eight providers. Models with reported parameter counts include 3B and 120B backbones. They are Qwen2.5-3B [29], Llama-3.2-3B [24], Command-R-35B [11], Claude Haiku 4.5 and Sonnet 4.5 [4, 5], Nova Lite [2], Kimi K2.5 [22], MiniMax M2.5 [25], and GPT-oss-120B [27]. We train the stopping predictors. The baseline-implementation cell is HotpotQA with Qwen2.5-3B.

Selection and the non-inferiority endpoint. We divide complete training runs into five folds. Each fold receives predictions from a model trained on the other folds. We use these held-out predictions to select the stopping threshold. This is five-fold cross-fitting. Test questions never enter fitting or threshold selection. For the primary $n = 400$ HotpotQA cell, the declared test asks whether guarded FBG saves calls while losing less than 3 EM points. We fixed this *non-inferiority* margin, EM, and exact LLM calls before test evaluation. The endpoint passes only when the paired one-sided 95% lower bound exceeds -3 and mean calls decrease. At observed EM levels of 20 to 30, the 3-EM tolerance permits substantial relative loss. We therefore test equal-cost accuracy and stricter margins of 1 and 2

Table 1: Retrieval settings and evaluation populations for the 26-cell core, crossed, and external analyses.

Setting	Shared index and retrieval unit	Fit / calibration / test	Provenance
HotpotQA distractor	Ten-paragraph question context, sentence units	60 to 90 / 0 / 100 to 140	9 call-only core cells
HotpotQA corpus	269,558 sentences from 7,405 contexts	90 / 30 / 400 (Haiku: 200)	7 core and 4 crossed-backbone cells
Wikipedia 25M (FEVER 2017)	25,247,888 sentences with reranking	90 / 30 / 400	1 core boundary cell
2WikiMultihopQA	206,743 sentences from 12,576 questions	90 / 0 to 30 / 140 to 400	4 core cells
MuSiQue answerable	21,100 paragraphs from 2,417 questions	60 to 90 / 0 to 30 / 100 to 400	5 core cells
MultiHopRAG news	53,943 sentence units from 609 articles and 49 outlets	96 / 24 / 400	1 prospectively frozen external cell

Table 2: Claim-to-evidence hierarchy. The status identifies the role of each analysis.

Claim	Evidence population	Status
Guarded FBG passes the declared full-search test	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, 14-cell tokens and wall time, answer-once ledger	**Robustness**
Question sampling can flip the endpoint decision	Same-question reruns and four independent question draws	**Prospectively frozen**
Signal role and selection explain the frontier gap	Other heads, calibration, rerolled futures, reselection, oracle	*Exploratory diagnosis*
The certificate refuses unsupported subsets	84 cell and target-risk pairs over 14 fully instrumented core cells	**Robustness**

EM. At the observed paired contrast variance, a normal approximation gives 80% power for a 1.39-EM non-inferiority margin when the true difference is zero. Four independent HotpotQA question draws repeat the whole pipeline. Two same-question reruns isolate decoding variation.

Baselines and comparability. Our matched simple family is $\mathcal{S} = \{\text{fixed-0}, \ldots, \text{fixed-}B, \text{answer repetition}\}$. These policies share each cell's trajectories, retriever, answer prompt, and cost reconstruction. Matched learned policies include FBG-only, guarded FBG, a binary future-improvement classifier, an error-head gate, gold-support classifiers, Stop-RAG [28], and full TASR when native answer-token probabilities are present [21]. A verifier-inspired substitute uses claims from the cell's backbone and public natural language inference (NLI) or relevance cross-encoders in place of HALT's released CUDA-only verifier [31]. Released Self-RAG, FLARE, SKR, and a trained Adaptive-RAG router retain their own retrieval and generation designs [6, 18, 19, 39]. We evaluate these adapted systems in one implementation cell and report them separately from policies replayed on shared runs.

Evidence status. Table 2 states the evidential role of each analysis. We use *prospectively frozen* only when a cell-specific protocol was written before its test output was inspected. Alternative learners and diagnostics without that status provide robustness or exploratory evidence.

Observed tradeoffs and statistical uncertainty. We first compare each learned policy with $\mathcal{S}$ using its measured EM and mean calls. A point is *beaten* if a simple policy has no lower EM and no higher cost, with at least one strict improvement (Pareto dominance). It is *better* when it is undominated by the simple family and Pareto-improves the development-selected $B^\star$: EM is no lower, cost is no higher, and at least one inequality is strict. All other undominated points are *frontier* points, including an exact match to $B^\star$ and accuracy-cost trade-offs that do not dominate it. Frontier membership alone establishes neither statistical significance nor acceptable answer quality. We separately report ϵ-dominance, which treats EM differences within a declared tolerance as equal. Our cell counts characterize the evaluated settings. A general success rate requires representative sampling.

A simple mixture assigns different simple policies to different questions, independently of their content. For example, it can give some questions a smaller budget and others a larger one. Matching average cost tests whether adapting to each question adds value. Geometrically, these mixtures form the upper concave hull of $\mathcal{S}$: the best accuracy obtainable by mixing simple policies. We compare the learned rule with a mixture of adjacent hull policies at its mean cost. The empirical hull is a retrospective comparator, whereas learned thresholds and $B^\star$ are frozen before testing. Beyond the hull, we use the best simple point at no greater cost. Point verdicts retain deterministic policies as deployable references. A *cellwise resolved* interval excludes zero in one cell. Bonferroni and max-t intervals account for testing several eligible cells for the same policy. A corrected result is *resolved* only if its interval excludes zero. An *unresolved* interval contains zero and supports neither direction. Equivalence requires a separate test.

Costs and uncertainty. We report EM as the primary quality measure and token F1 as a secondary measure. Answering before search and making one query and one answer per search uses $1 + 2T$ exact LLM calls. Fully instrumented runs also log input and output tokens, retrieval latency, and wall time. Two searches require five calls when answering at every prefix, versus three calls (two queries and one answer) under counterfactual answer-once accounting, $1 + T$. We derive this counterfactual ledger from the logged prefixes. It represents a different execution schedule, not a separately executed system. The execution frontier excludes the lifecycle cost of collecting full-budget trajectories, producing training prefixes, fitting predictors, and recalibrating after distribution change. Paired bootstrap intervals use 10,000 resamples, keeping both policies' outcomes for each question together. Question-clustered analyses also keep shared questions together across answer models. We apply max-t separately to each learned policy over cells for which the policy and per-question contrast vectors are available. The external cell is separate because several signal families are ineligible there.

5 Results

We test matched-cost accuracy (Section 5.1), robustness to new questions (Section 5.2), signal quality (Section 5.3), and post-stop certification (Section 5.4).

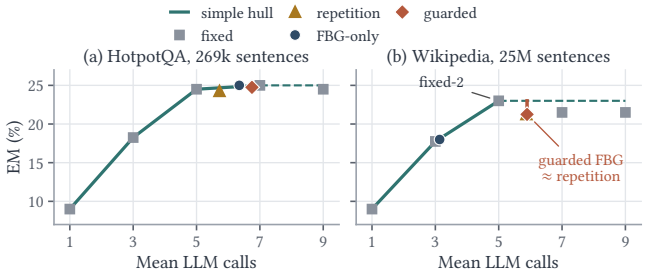

Figure 2: EM versus mean LLM calls. Solid lines join simple-policy hull vertices. Dashed plateaus retain the best simple point at no greater cost. Vertical gaps are point-estimate comparisons.

5.1 RQ1: Does learning improve the frontier?

Full-budget savings overstate adaptive value. On the nine-backbone HotpotQA panel with call accounting, FBG-only removes 27 to 57% of searches relative to the full budget. Its paired EM differences range from −2.1 to +2.0 points, and all nine intervals contain zero. The relevant comparison is with a tuned smaller budget, which can also save searches. Across the core 26 cells, FBG-only is dominated in 10 cells, lies on the frontier without improving $B^\star$ in 13, and improves $B^\star$ in 3. Guarded FBG is dominated in 15 of 26 cells, is frontier-only in 6, and improves $B^\star$ in 5. Table 3 separates the 12 call-only and 14 fully instrumented cells. Here "vs. full" compares against full search. Risk NA means unavailable raw features, and 'refuse' means zero certificate coverage.

Figure 2 illustrates why the comparator matters. In the primary HotpotQA cell, guarded FBG reaches 24.75 EM at 6.75 calls versus 24.50 at 9.00 for the full budget. FBG-only is cheaper and has slightly higher point accuracy than the guard. Fixed-2 remains competitive on the same frontier. In the Wikipedia 25M cell, fixed-2 reaches 23.00 EM at 5.00 calls and strictly dominates guarded FBG at 21.25 EM and 5.90 calls.

We measure the available gain with a hindsight oracle that stops at the first correct prefix and charges the full budget if none is correct. It reveals available stopping opportunities but requires knowing which answers are correct, information unavailable during deployment. Table 4 reports this ceiling for the 14 fully instrumented core cells.

The oracle dominates full search in all fourteen cells, with gains from +1.75 to +8.00 EM and a mean gain of +4.84. In the primary cell, it reaches 29.50 EM at 3.14 searches and 7.28 calls, compared with 24.50 EM at four searches and nine calls for full search. Guarded FBG reaches 24.75 EM in that cell. We then test how much of this gain learned rules capture from information available at the stopping decision.

We quantify uncertainty in equal-cost accuracy gaps with paired intervals. Table 5 summarizes eligible cells, point verdicts, equal-call gaps, and intervals excluding zero before and after correction. NA marks unavailable per-question results.

Joint uncertainty changes the verdict. Policy-specific eligibility explains why TASR and Stop-RAG do not cover all 18 cells. Neither FBG policy has a positive cellwise equal-call interval, and each has two negative intervals. The binary future-improvement classifier has three positive intervals and one negative interval. Its sole Bonferroni-positive comparison is +0.11 EM with interval [+0.03, +0.19]. The error-head gate has one positive estimate of +0.14 EM with Bonferroni interval [+0.02, +0.33]. Neither remains positive under its max-t simultaneous band. Bonferroni uses bootstrap percentiles, while max-t scales differences by their estimated uncertainty and bounds them jointly. A result can therefore pass one correction without passing the other. The two corrected gains remain below 0.2 EM and are operationally small. The gold-support classifier lacks the per-question vectors required for corrected intervals, so its Bonferroni and max-t entries are NA.

We repeat the hull comparison at matched calls, input tokens, and wall time. Across the 14 fully instrumented cells, guarded FBG differs from the matched hull by −0.59, −0.64, and −0.60 EM, with no positive FBG interval in any of these cost measures. Under an answer-once ledger, guarded FBG is dominated in 17 of 18 cells and repetition in all 18. Under this accounting, repetition is competitive only when prefix answers are already available.

Precision of the matched-cost comparisons. Across the 18-cell family, the saved cellwise 95% equal-call intervals have median width 2.75 EM points for guarded FBG (range 0.06 to 3.75) and 2.98 for FBG-only (range 0.00 to 8.48). These intervals hold policies and the hull fixed. Identical outcomes can yield zero width despite uncertainty on new questions. The 1.39-EM power calculation concerns the full-search comparison, where full search and guarded FBG differ in correctness on only 44 of 2,000 question-draw records. Equal-cost comparisons involve different pairs of outcomes, and their intervals still allow practically useful gains.

Shared-run policies and adapted systems. On the pooled frontier of all simple and learned policies in the 14 fully instrumented cells, FBG-only is undominated in 11, fixed-2 in 12, and the binary future-improvement classifier in 13. Against simple policies on the 18-cell family, Stop-RAG is beaten in 9 of 14 eligible cells with three negative but no positive equal-call intervals, and full TASR in 5 of 12. An NLI-augmented FBG variant is the cheapest policy within 1 EM of full search in only 1 of 14 cells. The supplement reports the adapted systems and verifier-inspired substitutes. Their different generation designs provide system-level context rather than isolating the stopping decision.

Key finding. Choosing exits with hindsight raises accuracy by 4.84 percentage points on average. Simple rules Pareto-dominate guarded FBG in 15 of 26 settings: no worse accuracy or cost, with at least one strict improvement. Learning must capture useful exits at competitive cost.

5.2 RQ2: Does non-inferiority persist?

New questions can change the test outcome. On 400 untouched HotpotQA questions, guarded FBG reaches 24.75 EM versus 24.50 for the full budget. The paired difference is +0.25 with a 95% interval of [−0.75, 1.50] and a one-sided lower bound of −0.75. Calls fall from 9.00 to 6.75, input tokens fall by 30.7%, output tokens by 27.1%, and wall time by 23.8%. The policy passes the declared non-inferiority test: it costs less, and the one-sided bound rules out a loss of 3 EM

Table 3: Complete FBG point audit across 12 call-only and 14 fully instrumented core cells. EM / mean calls use half-up rounding. Verdicts use unrounded values and the definitions in Section 4.

		Full search	Selected budget $B^\star$	Repetition	FBG-only			Guarded FBG			Risk
Cell	n	EM	EM/calls (searches)	EM/calls	EM/calls	vs. full [CI]	Verdict	EM/calls	vs. full [CI]	Verdict	certificate
HotpotQA / Haiku 4.5	140	52.14	52.14/5.00 (2)	50.00/5.47	50.71/4.76	-1.43 [-5.00,2.14]	frontier	51.43/5.71	-0.71 [-3.57,1.43]	beaten	NA
HotpotQA / Sonnet 4.5	140	52.14	52.14/7.00 (3)	52.86/5.93	53.57/6.70	+1.43 [0.00,3.57]	better	53.57/6.77	+1.43 [0.00,3.57]	better	NA
HotpotQA / GPT-oss-120B	140	57.14	59.29/5.00 (2)	55.00/4.69	55.71/5.97	-1.43 [-4.29,1.43]	beaten	55.71/6.19	-1.43 [-4.29,1.43]	beaten	NA
HotpotQA / Nova Lite	140	54.29	54.29/7.00 (3)	54.29/5.04	52.86/6.87	-1.43 [-3.57,0.00]	beaten	53.57/6.96	-0.71 [-2.14,0.00]	beaten	NA
HotpotQA / Qwen2.5-3B	140	38.57	40.00/7.00 (3)	39.29/5.87	39.29/6.11	+0.71 [-3.57,5.00]	beaten	40.00/6.73	+1.43 [-2.14,5.00]	better	NA
HotpotQA / Llama-3.2-3B	140	39.29	37.86/5.00 (2)	37.86/5.39	37.86/5.43	-1.43 [-6.43,3.57]	beaten	39.29/6.31	+0.00 [-4.29,4.29]	frontier	NA
HotpotQA / Command-R-35B	100	51.00	52.00/7.00 (3)	54.00/5.06	53.00/6.76	+2.00 [0.00,5.00]	beaten	53.00/6.78	+2.00 [0.00,5.00]	beaten	NA
HotpotQA / Kimi K2.5	140	59.29	59.29/5.00 (2)	56.43/4.67	57.86/4.94	-1.43 [-5.00,1.45]	frontier	57.86/5.31	-1.43 [-5.00,1.45]	beaten	NA
HotpotQA / MiniMax M2.5	140	55.71	55.71/9.00 (4)	45.00/4.20	53.57/4.41	-2.14 [-7.14,2.86]	frontier	53.57/5.53	-2.14 [-6.43,2.14]	frontier	NA
2Wiki / Qwen2.5-3B	140	28.57	30.71/7.00 (3)	29.29/5.70	28.57/6.20	+0.00 [-2.86,2.86]	beaten	29.29/6.87	+0.71 [-1.43,2.86]	beaten	NA
MuSiQue / Qwen2.5-3B	140	13.57	12.86/5.00 (2)	12.86/6.27	12.14/5.09	-1.43 [-4.29,1.43]	beaten	13.57/6.39	+0.00 [0.00,0.00]	frontier	NA
MuSiQue / Llama-3.2-3B	100	18.00	18.00/7.00 (3)	17.00/5.86	16.00/5.86	-2.00 [-7.00,3.00]	beaten	18.00/6.88	+0.00 [-3.00,3.00]	better	NA
HotpotQA A, seed 7	400	24.50	25.00/7.00 (3)	24.25/5.73	25.00/6.35	+0.50 [-0.75,1.75]	better	24.75/6.75	+0.25 [-0.75,1.50]	frontier	refuse
HotpotQA A, seed 11	400	25.00	25.25/7.00 (3)	25.00/5.70	25.25/5.74	+0.25 [-2.00,2.50]	beaten	25.25/6.51	+0.25 [-1.00,1.50]	beaten	refuse
HotpotQA A, seed 19	400	26.00	26.00/7.00 (3)	25.50/5.72	25.25/6.38	-0.75 [-2.00,0.50]	beaten	25.50/6.68	-0.50 [-1.50,0.50]	beaten	refuse
HotpotQA B, seed 7	400	31.25	31.25/9.00 (4)	27.75/5.75	29.50/6.62	-1.75 [-3.50,-0.25]	frontier	29.25/6.71	-2.00 [-3.50,-0.50]	frontier	refuse
HotpotQA C, seed 7	400	26.50	24.75/7.00 (3)	23.25/5.90	13.50/2.67	-13.00 [-17.00,-9.25]	frontier	23.25/5.90	-3.25 [-5.25,-1.25]	frontier	refuse
HotpotQA D, seed 7	400	25.50	25.50/9.00 (4)	26.50/5.89	26.25/5.42	+0.75 [-1.50,3.00]	better	27.00/6.36	+1.50 [0.00,3.00]	better	refuse
HotpotQA E, seed 7	400	34.25	34.25/9.00 (4)	32.00/5.71	33.25/7.51	-1.00 [-2.25,0.00]	frontier	34.25/7.76	+0.00 [-0.75,0.75]	better	refuse
2Wiki corpus (seed 7, draw 1)	400	23.25	24.25/7.00 (3)	23.50/5.50	20.25/3.11	-3.00 [-6.25,0.25]	frontier	23.50/5.50	+0.25 [-2.00,2.50]	beaten	refuse
2Wiki corpus (seed 7, draw 2)	400	25.25	25.25/9.00 (4)	23.25/5.51	19.00/3.84	-6.25 [-9.25,-3.50]	frontier	23.50/5.59	-1.75 [-3.75,0.00]	beaten	refuse
2Wiki corpus (seed 7, draw 3)	400	25.50	26.25/5.00 (2)	25.25/5.61	24.25/3.12	-1.25 [-4.75,2.25]	frontier	25.75/5.66	+0.25 [-1.75,2.25]	beaten	refuse
MuSiQue corpus (seed 7, draw 1)	400	9.25	7.25/7.00 (3)	7.50/6.53	3.75/3.03	-5.50 [-8.25,-3.00]	frontier	7.50/6.54	-1.75 [-3.25,-0.50]	beaten	refuse
MuSiQue corpus (seed 7, draw 2)	400	10.25	10.25/9.00 (4)	9.25/6.45	2.50/1.44	-7.75 [-10.50,-5.25]	frontier	9.25/6.50	-1.00 [-2.50,0.25]	beaten	refuse
MuSiQue corpus (seed 7, draw 3)	400	12.00	10.75/7.00 (3)	10.75/6.43	0.25/1.00	-11.75 [-15.00,-8.75]	frontier	10.75/6.43	-1.25 [-2.50,-0.25]	beaten	refuse
Wikipedia 25M	400	21.50	23.00/5.00 (2)	21.25/5.88	18.00/3.14	-3.50 [-6.00,-1.00]	frontier	21.25/5.90	-0.25 [-2.00,1.25]	beaten	refuse

Table 4: Hindsight-oracle ceiling for fourteen fully instrumented cells. Never-correct questions incur the full budget.

	Full search	Hindsight oracle			
Run	EM	EM	Gain	Searches	Calls
HotpotQA A, seed 7	24.50	29.50	+5.00	3.14	7.28
HotpotQA A, seed 11	25.00	30.25	+5.25	3.12	7.24
HotpotQA A, seed 19	26.00	30.25	+4.25	3.13	7.25
HotpotQA sample B	31.25	37.00	+5.75	3.00	7.00
HotpotQA sample C	26.50	30.25	+3.75	3.23	7.46
HotpotQA sample D	25.50	32.25	+6.75	3.12	7.24
HotpotQA sample E	34.25	39.00	+4.75	2.94	6.88
2Wiki draw 1	23.25	30.25	+7.00	3.12	7.23
2Wiki draw 2	25.25	31.25	+6.00	3.15	7.31
2Wiki draw 3	25.50	33.50	+8.00	3.04	7.08
MuSiQue draw 1	9.25	11.25	+2.00	3.78	8.55
MuSiQue draw 2	10.25	12.50	+2.25	3.71	8.43
MuSiQue draw 3	12.00	13.75	+1.75	3.69	8.39
Wikipedia 25M	21.50	26.75	+5.25	3.19	7.38

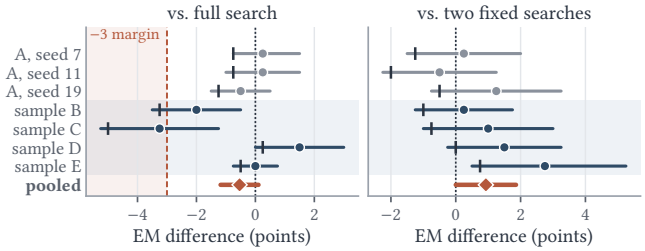

Figure 3: Guarded FBG minus full search or fixed-2 (EM points). Bars: paired 95% intervals. Ticks: one-sided lower bounds. Dashed line: −3-EM margin. Fixed-2 is not cost matched.

Table 5: Cross-signal comparison over the principal 18-cell family. Cells give point / matched-cost eligibility. Verdicts count beaten / frontier-only / improves selected budget. Wins / losses count positive / negative cellwise intervals.

		Verdict	Gap range	Wins/	Corrected wins	
Signal	Cells	counts	(EM)	losses	Bonferroni	max-t
FBG-only	18/18	2/14/2	−2.38, 1.86	0/2	0	0
Guarded FBG	18/18	11/5/2	−1.89, 0.85	0/2	0	0
Future-improvement clf.	18/18	1/15/2	−2.40, 2.61	3/1	1	0
Error-head gate	18/18	9/9/0	−1.50, 0.70	1/0	1	0
Gold-support classifier	18/18	8/8/2	−2.00, 0.47	0/1	NA	NA
Full TASR	12/12	5/6/1	−1.75, 0.53	0/0	0	0
Stop-RAG	14/14	9/4/1	−3.00, 0.50	0/3	0	0

points or more. The two-sided interval includes zero, so this result does not establish an accuracy gain.

Figure 3 separates changes in answer generation from changes in the questions. Grey rows rerun sample A, blue rows use new question samples, and rust marks the question-clustered pooled estimate. With the same 400 questions, decoding seeds 7, 11, and 19 give guarded differences of +0.25, +0.25, and −0.50 EM. All pass the predeclared rule. Four new question draws give −2.00, −3.25, +1.50, and 0.00 EM. Their one-sided lower bounds are −3.25, −5.00, +0.25, and −0.50, so two of four pass and two fail. Across all five question samples, three of five pass. The two negative cellwise intervals include zero after max-t adjustment across the comparison family.

The draw-level mean is −0.70 EM with one-sided lower bound −2.51. Pooled contrasts against full search and fixed-3 are −0.54 [−1.18, 0.11] and −0.32 [−1.03, 0.37] EM, and both contain zero. Contrasts favoring the guard over fixed-2 or repetition use 1.67 and 0.90 more calls. Thus every favorable pooled contrast costs more, while every equal-cost or lower-cost contrast remains unresolved.

Our replication panel shows greater sensitivity to question samples than to decoding seeds. A precise variance estimate requires more than these five draws and three seeds. Full-budget and guarded correctness differ on 44 of 2,000 questions, allowing a small item set to determine a draw-level decision.

Backbone crosses and Wikipedia-derived datasets. Three Claude Haiku 4.5 draws with $n = 200$ and one Llama-3.2-3B draw with $n = 400$ all pass the 3-EM margin. Guarded FBG reaches the frontier formed by all simple and learned policies in none, and its mean equal-call difference is +0.14 EM. Changing the backbone therefore changes the accuracy level more than the deployment verdict.

On the 25,247,888-sentence Wikipedia 25M index, guarded FBG has paired difference −0.25 EM with interval $[-2.00, 1.25]$ and lowers calls from 9.00 to 5.90. Fixed-2 still dominates it. Three 2Wiki draws lower calls to between 5.5 and 5.7, but fixed-2 dominates the guard in every draw. On three MuSiQue draws, answer repetition or fixed-2 also dominates the guard. Domination thus recurs across several Wikipedia-derived benchmarks that share sparse retrieval and multi-hop questions.

One external news-corpus cell. We froze one MultiHopRAG inference-query cell before fitting or test evaluation. It contains 400 questions over 609 English news articles and uses Qwen2.5-3B with the same four-search cap. Table 6 separates the frozen FBG endpoint from three exploratory signal policies. We compare each policy with a random mixture of simple policies at the same mean call count.

Table 6: External news-cell policies at equal calls to the simple-policy hull.

Policy	EM	Calls	Gain vs. simple	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	$[0.41, 2.06]$

Guarded FBG reaches 73.25 EM at 7.40 calls, compared with 72.75 EM at 9.00 calls for full search. It passes the declared endpoint, but the matched-cost advantage remains unresolved. The exploratory binary future-improvement and error-head policies have positive intervals on this draw. These gains make the two policies candidates for replication on independent news-corpus samples.

Key finding. Rerunning answer generation on the same questions consistently passes the allowed accuracy-loss test. Across five question samples, three pass. Repeating generation on one sample therefore does not establish reliability on new questions.

5.3 RQ3: Which signals help?

Current error provides the stronger risk ordering. We rank stopped answers from lowest to highest predicted error. *Coverage* is the fraction retained. Risk-coverage area averages their error rates over coverage from 0.05 to 1.00, with lower values indicating lower risk. We compare with 20 random orderings. Expected calibration error (ECE) measures the gap between predicted probabilities and observed frequencies. Area under the receiver operating characteristic curve (AUROC) measures how well scores separate the two classes. Table 7 shows that our current-error head improves on random in every cell, while FBG does so in 7 of 18. This gap separates the ability to rank current errors from the ability to allocate searches that will correct them.

Table 7: Signal and trajectory diagnostics for RQ3. Test labels evaluate the frozen outputs and never select a policy.

Diagnostic	Scope	Measured result
FBG risk ordering	18 cells	Better than random in 7 cells
Error-head ordering	18 cells	Better than random in all 18 cells
Error calibration	18 cells	Test-prefix ECE 0.019 to 0.087, median 0.043
Gold-support coverage	18 cells	Cross-fitted AUROC 0.652 to 0.946
Late answer reversals	1,220 records	79 harmful and 14 helpful
Rerolled futures	300 stops	Beat logged best in 10.0%, changed correctness in 5.7%



Figure 4: Refitting and reselection over 400 training resamples per run (percentages). Beaten means dominated by a reselected simple policy. Improving the selected budget can coexist with domination.

Better prediction need not improve stopping. Our same-feature gold-support classifier separates evidence-coverage states with AUROC from 0.652 to 0.946. As a stopping rule, it is beaten in 8 cells, frontier-only in 8, and better in 2, with no cellwise equal-call win. The future-improvement classifier is dominated in only one cell, but its three positive intervals disappear under max-t correction.

Current-error prediction identifies wrong answers, whereas FBG must also anticipate a useful later reversal. Threshold choice and population shift further separate ranking from policy value. Our diagnostics separate prediction quality from selection stability. Identifying their causal effects requires controlled interventions.

Late retrieval leaves avoidable errors. We record 79 correct-to-wrong and 14 wrong-to-correct changes after late retrieval, with a 95% ratio interval of 3.2 to 11.1 and $p < 10^{-3}$. On the nine-backbone call-only panel, FBG-only realizes 24 beneficial and 32 harmful flips. The policy trades avoided late errors against new errors from its own exits.

Where does instability arise? Held-out-prefix FBG correlations range from 0.26 to 0.60 and absolute errors from 0.12 to 0.19. To measure how training variation affects decisions, we resample complete training runs 400 times, repeating fitting and selection for both learned and simple policies. Questions, retrieval, preprocessing, and decoding stay fixed.

Figure 4 shows how often each verdict survives refitting. Columns 7, 11, and 19 use decoding reruns of HotpotQA sample A. A policy can improve on the reselected fixed budget while another simple policy dominates it, so the corresponding heatmap rows can overlap. The logged threshold is reselected in only 3 to 17% of resamples on six of seven HotpotQA conditions. A simple policy dominates the guard in 51 to 100% of resamples in 13 of 14 runs, all except sample E.

Two fresh continuations from each of 300 sampled primary-cell stops beat the recorded best answer in 10.0% of cases and change correctness in 5.7%. At temperature 0.8, the former is an upper bound on support bias that also includes decoding noise.

The original clipped target variants move mean guarded EM by at most 0.2 points. The logged guarded policies have beaten/frontier/better counts 9/3/2 across the 14 instrumented cells. Refitting the full feature set and reselecting the threshold gives 11/2/1, whereas removing step, evidence-count, and stability features gives 11/1/2. The tallies refer to different fitted policies: 8 of 14 refits recover the logged threshold. Across 154 protocol variants, simple policies dominate 114 and the guard Pareto-improves $B^\star$ in 17.

Across 1,400 refits, increasing training from 30 to 90 episodes lowers domination by 6.1 points and raises mean EM by 0.14, although improvement is not monotonic. Pooling training runs from same-environment cells (up to 421 episodes) or all 14 cells (up to 959), excluding target test questions, cuts guarded domination from 11 per-cell refits to 8 and 9 cells, and FBG-only domination from 10 per-cell refits to 4 and 6. Mean matched-cost gaps remain -0.22 to -0.53 EM. Of 56 paired intervals, 1 favors learning and 6 the mixture (Supp. S15).

We test larger training sets on the same question population with 4,480 fresh trajectories: 2,880 for training and 1,600 new test questions, combined with 400 logged test questions. From 180 to 2,880 training runs, FBG-only gains 0.39 to 0.76 EM over the matched-call mixture. Conditional simultaneous bands exclude zero, with a lower bound of 0.01 EM at 2,880. Guarded FBG remains unresolved, and neither rule resolves an accuracy gain over $B^\star$. We establish this gain for one backbone and one four-search cell (Supp. S15).

Target clipping and simple controls. On the same 14 cells, a signed best-future target uses the original features, five-fold training-chain cross-fitting, and threshold grid. The guarded signed refit has beaten/frontier/better counts 11/3/0, compared with 11/2/1 for a freshly refitted clipped target. Mean EM changes from 22.13 to 22.05, while mean calls fall from 6.44 to 6.26. This isolates whether retaining negative future gains helps at the same search budget. We also select a deterministic evidence-growth rule on training chains: after at least one search, stop when newly retained units are at most a threshold in $\{0, 1, 2, 3\}$. Adding this rule to the simple family gives beaten/frontier/better counts 11/2/1 for logged guarded FBG and 3/9/2 for FBG-only. Across calls, tokens, and a wall-time proxy, mean matched-cost gaps are negative and all 84 conditional simultaneous bands include zero (supplement). A training-selected query-repetition rule matches $B^\star$ EM at 0.23 fewer calls on average. Adding it also leaves all 84 bands unresolved (Supp. S21). We extend Haiku 4.5 to eight searches with 120 fresh training and 200 test chains (90 fit the policies). FBG-only reaches 55.5 EM at 8.19 calls. Its equal-call gap is +0.50 EM with band $[-2.22, 3.22]$. All 12 bands across horizons four and eight include zero (Supp. S22).

Key finding. A wrong answer does not tell us whether another search will help. In one setting, FBG-only achieves a small but statistically significant accuracy gain at the same call cost with 180 or more training runs.

5.4 RQ4: Can we certify stopped answers?

For each of 14 fully instrumented cells, we reserve 30 of 120 development episodes, fit the error scorer on 90, and evaluate one guarded-policy stop per calibration and test episode. After correcting for its cutoff search, SGR certifies no answer subset in any of the 84 cell and target-risk pairs (six α values in each of 7 HotpotQA corpus, 3 2WikiMultihopQA, 3 MuSiQue, and 1 Wikipedia 25M cells), so no test answers are certified. Two factors explain this result: the small calibration set and the high answer error rates. With $m = 30$, each binomial test runs at level $0.10/\lceil \log_2 30 \rceil = 0.02$, so even 30 calibration answers with no errors cannot certify a risk below 12.2%, and smaller candidate sets have looser bounds. Guarded EM in these cells ranges from 7.50 to 34.25 (Table 3), so most stopped answers are wrong and no score-ranked subset meets even $\alpha = 0.6$.

Key finding. Saving searches does not guarantee reliable answers. In the 14 core instrumented cells, no subset meets the certification rule with 30 calibration examples per cell. Certification requires sufficient calibration data and a low enough error rate.

6 Discussion and Scope

Our results distinguish detecting a wrong answer from deciding whether another search will help. With more training, we resolve a small FBG-only gain at matched call cost in one setting. The gain over the selected fixed budget remains unresolved. This comparison matters because a shorter fixed budget can also save searches. We assess answer reliability separately through certificate coverage and accepted-answer risk.

Costs depend on the surrounding system. Repetition needs intermediate answers, so its advantage must be reassessed when answers are generated only once. Our records exclude training, maintenance, billing, energy, retries, and failures. Supp. S3 maps the measured regimes to policy choices.

We study nine features and four searches on static English corpora, then extend training size and the search budget within individual settings. Training resamples test fitting variability, whereas new questions test whether an advantage repeats. New retrievers, languages, hidden states, or joint query selection need frozen development-selected rules, declared costs, and independent tests with paired and simultaneous intervals.

7 Conclusion

We establish the distinction between useful stopping opportunities and competitive stopping policies. Simple policies dominate guarded FBG in 15 of 26 core settings, while our larger training study resolves a small FBG-only gain in one cell. We recommend deployment tests against tuned simple policies on independent questions, with the cost of intermediate answers included.

Acknowledgments

We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, for supporting this study.

Ethical Considerations

Errors from premature stopping. A controller can reduce resource use while withholding evidence that would correct an answer. This risk may be concentrated on topics that are rare in the development data, including questions involving underrepresented regions, languages, or communities. We therefore evaluate quality and cost jointly and recommend deployment only after an independent, task-specific frontier test. Evaluating multilingual equity requires extending this English-only evidence.

Limits of certification. The certificate in Proposition 1 bounds the error rate among accepted answers when calibration and test episodes are independent and drawn from the same distribution. It does not bound risk for an individual question or subgroup and must not be presented as a per-answer probability of correctness. Corrected SGR returns no certificate at any evaluated target in the 14 core instrumented cells. Deployment interfaces should mark these answers as uncertified, recalibrate by task family, and monitor subgroup risk after distribution shift.

Resource claims and operational telemetry. Our fully instrumented runs measure LLM calls, input and output tokens, and wall time. Energy, invoice cost, and the marginal environmental effect of a provider call lie outside these measurements. Historical execution logs omit timeout and retry failure codes. The fresh eight-search extension records zero failed requests and a list-price cost estimate of $2.03, rather than an invoice. The resource findings apply to calls, tokens, and time. Assessing energy, billed cost, and service reliability requires separate measurements. We train the small stopping predictors on CPU and keep the backbones frozen.

Data, access, and dual use. HotpotQA, 2WikiMultihopQA, MuSiQue, FEVER, and MultiHopRAG are public research datasets used under their respective terms. The evaluated material includes Wikipedia-derived text and one English news snapshot. We use no private user or production data and access no live websites. The artifact at https://github.com/lexuanbach/wsdm2027-learned-stopping-audit releases code and saved outputs. Its trajectories may repeat public biographical or news facts and should be handled under the source datasets' licenses. The same stopping machinery could be tuned for cheap but unreliable answering. Publishing selection outcomes and certificate coverage lets users inspect that accuracy-cost tradeoff.

References

[1] Pranjal Aggarwal, Aman Madaan, Ankit Anand, Srividya Pranavi Potharaju, Swaroop Mishra, Pei Zhou, Aditya Gupta, Dheeraj Rajagopal, Karthik Kappaganthu, Yiming Yang, Shyam Upadhyay, Manaal Faruqui, and Mausam. 2024. AutoMix: Automatically Mixing Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[2] Amazon Artificial General Intelligence. 2024. *The Amazon Nova Family of Models: Technical Report and Model Card.* Technical Report. Amazon. https://www.amazon.science/publications/the-amazon-nova-family-of-models-technical-report-and-model-card

[3] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2024. Conformal Risk Control. In *International Conference on Learning Representations (ICLR)*.

[4] Anthropic. 2025. *System Card: Claude Haiku 4.5.* Technical Report. Anthropic. https://www.anthropic.com/claude-haiku-4-5-system-card

[5] Anthropic. 2025. *System Card: Claude Sonnet 4.5.* Technical Report. Anthropic. https://www.anthropic.com/claude-sonnet-4-5-system-card

[6] 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)*.

[7] Jihwan Bang, Juntae Lee, Seunghan Yang, and Sungha Choi. 2025. Think Straight, Stop Smart: Structured Reasoning for Efficient Multi-Hop RAG. arXiv:2510.19171 NeurIPS 2025 Workshop.

[8] Lingjiao Chen, Matei Zaharia, and James Zou. 2024. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=cSimKw5p6R

[9] Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. Do NOT Think That Much for 2+3=? On the Overthinking of Long Reasoning Models. In *Proceedings of the 42nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 267)*. 9487–9499.

[10] Yuan Shih Chow, Herbert Robbins, and David Siegmund. 1971. *Great Expectations: The Theory of Optimal Stopping*. Houghton Mifflin.

[11] Cohere. 2024. Command R: Retrieval-Augmented Generation at Scale. Cohere documentation changelog. https://docs.cohere.com/changelog/command-r-retrieval-augmented-generation-at-production-scale

[12] Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, Nicholas Thumiger, Aditya Desai, Ion Stoica, Ana Klimovic, Graham Neubig, and Joseph E. Gonzalez. 2025. The Danger of Overthinking: Examining the Reasoning-Action Dilemma in Agentic Tasks. *arXiv preprint arXiv:2502.08235* (2025).

[13] 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

[14] Zhengru Fang, Senkang Forest Hu, Zhonghao Chang, Yu Guo, Yihang Tao, Hongyao Liu, Mengzhe Ruan, Jun Huang, and Yuguang Fang. 2026. Inference-Time Budget Control for LLM Search Agents. arXiv:2605.05701 [cs.AI]

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

[16] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*.

[17] Ronald A. Howard. 1966. Information Value Theory. *IEEE Transactions on Systems Science and Cybernetics* 2, 1 (1966), 22–26.

[18] Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

[19] 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)*.

[20] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Ö. Arık, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. In *Conference on Language Modeling (COLM)*. https://openreview.net/forum?id=Rwhi91ideu

[21] 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.

[22] Kimi Team. 2026. Kimi K2.5: Visual Agentic Intelligence. *arXiv preprint arXiv:2602.02276* (2026).

[23] 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.

[24] Meta AI. 2024. Llama 3.2: Revolutionizing Edge AI and Vision with Open, Customizable Models. Meta AI blog. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

[25] MiniMax AI. 2026. MiniMax Model Release Notes: M2.5. MiniMax platform documentation. https://platform.minimax.io/docs/release-notes/models Model release listed in February 2026.

[26] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332* (2021).

[27] OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv preprint arXiv:2508.10925* (2025).

[28] 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.

[29] Qwen Team. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115* (2024).
[30] Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. 2023. Robots That Ask For Help: Uncertainty Alignment for Large Language Model Planners. In *Conference on Robot Learning (CoRL)*.
[31] 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.
[32] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems (NeurIPS)*.
[33] Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. 2022. Confident Adaptive Language Modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*.
[34] Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. DRAGIN: Dynamic Retrieval Augmented Generation based on the Real-time Information Needs of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.
[35] Yixuan Tang and Yi Yang. 2024. MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries. In *Conference on Language Modeling*. arXiv:2401.15391 https://arxiv.org/abs/2401.15391
[36] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.
[37] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World*. Springer.
[38] Abraham Wald. 1947. *Sequential Analysis*. John Wiley & Sons.
[39] Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023. Self-Knowledge Guided Retrieval Augmentation for Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP*.
[40] 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
[41] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*.
[42] Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Weichuan Liu, Lei Hou, and Juanzi Li. 2025. SeaKR: Self-aware Knowledge Retrieval for Adaptive Retrieval Augmented Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 27022–27043. doi:10.18653/v1/2025.acl-long.1312