EVICAST: FORECASTING RETRIEVAL GAINS THROUGH SHARED RESPONSE CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation extends the knowledge available to large language models with external evidence. However, relevant evidence does not ensure that a generator can use it effectively. Indiscriminate retrieval can degrade task performance and increase retrieval and long-context inference costs. Existing methods neither determine before large-scale execution whether a generator, task distribution, and retriever configuration is worth deploying nor diagnose multiple candidate retrievers under a shared budget. We introduce EviCast, a framework that forecasts retrieval value through interventional evidence-state decomposition. EviCast measures a generator's response to controlled evidence states once, then combines this response profile with the offline evidence-state distribution of each candidate retriever. The resulting decomposition forecasts end-to-end net utility and its 95% lower confidence bound without executing every complete retrieval-augmented generation configuration. In frozen evaluations, EviCast correctly decides all 14 HotpotQA retrieval configurations, achieves a utility forecast mean absolute error of 2.75%, and reduces diagnostic generation calls from 480 to 96, an 80% reduction. Its topology-aware instance captures nonlinear complementarity between terminal and bridge evidence, correctly deciding all 16 configurations formed by two datasets, two generators, and four retrievers with an overall forecast error of 2.81%. These results show that retrieval deployment depends on both evidence availability and the target generator's ability to convert evidence structure into stable net utility. EviCast replaces default retrieval with a measurable, predictable, and reusable decision for resource-adaptive retrieval-augmented generation systems. Code is available at `https://anonymous.4open.science/r/EviCast-ECC2`.

1 INTRODUCTION

Large language models (LLMs) excel at knowledge-intensive question answering, content generation, and complex reasoning. However, their parametric knowledge is slow to update, incomplete, and difficult to trace to source material. Retrieval-augmented generation (RAG) introduces updatable and attributable external knowledge at inference time. It has therefore become a fundamental interface between LLMs and real-world knowledge bases (Lewis et al., 2020).

RAG does not benefit every model or task. Its value depends on both retrieval relevance and the generator's ability to use the retrieved evidence. Irrelevant context can distract LLMs and induce incorrect answers (Shi et al., 2023; Amiraz et al., 2025). Indiscriminate retrieval may therefore harm performance while adding retrieval and long-context inference costs, whereas overly conservative retrieval may forfeit substantial gains from external evidence. Figure 1 illustrates this deployment dilemma. We ask whether the effect of external evidence on a target model can be predicted before deployment and used to enable RAG only when beneficial. We further ask whether one diagnosis can serve multiple candidate retrieval schemes. These questions define retrieval necessity and its associated resource-efficiency problem.

Existing work studies retrieval decisions and utility at three levels. Forward-Looking Active Retrieval augmented generation (FLARE), Self-Reflective Retrieval-Augmented Generation (Self-RAG), and Adaptive-RAG trigger retrieval for an individual query or generation step (Jiang et al., 2023; Asai et al., 2024; Jeong et al., 2024). Retrieval Augmented Generation Assessment (RAGAS)

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

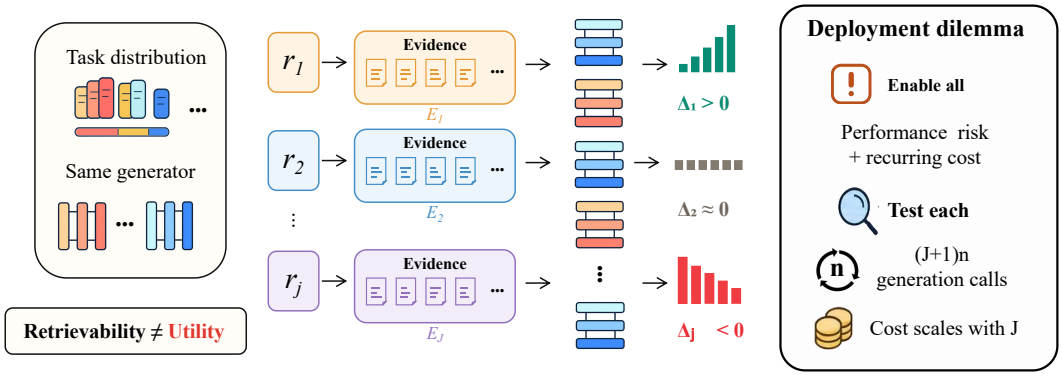

Figure 1: Retrieval utility and diagnosis under a fixed task distribution and generator.

and RAGChecker evaluate RAG systems after execution (Es et al., 2024; Ru et al., 2024). SePer, passage utility models, InfoGain-RAG, and retrieval performance prediction estimate the local utility of a query or document (Dai et al., 2025; Perez-Beltrachini & Lapata, 2025; Wang et al., 2025; Tian et al., 2026). These methods do not determine before large-scale execution whether a complete generator, task distribution, and retriever configuration is worth deploying, nor do they diagnose multiple candidate retrievers under a shared generation budget. We formulate this problem as amortized retrieval necessity diagnosis at deployment time. Given a generator g, task distribution $\mathcal{D}$, and candidate retriever r, the objective is to predict the expected gain $\Delta(g, \mathcal{D}, r)$ over the no-retrieval baseline and make a deployment decision.

Diagnosing retrieval necessity presents two technical challenges. First, the evidence representation must reflect how the generator uses evidence because RAG utility depends jointly on retrieved evidence and the generator's ability to exploit it. Relevance and correct-passage counts are insufficient. In multihop question answering, one top-2 result may contain a bridge passage and a distractor, while another contains a terminal answer passage and a distractor. Both contain one correct passage but may affect the final answer differently. Second, the generator response must transfer reliably across retrievers without executing every candidate RAG configuration, while accounting for uncertainty from small calibration samples. Directly comparing retrieval and no-retrieval performance for each retriever makes generation cost grow linearly with the number of candidates and yields variable small-sample gain estimates.

We address these challenges with EviCast, a retrieval necessity diagnosis framework based on interventional evidence-state calibration. EviCast decomposes candidate retriever utility into a generator response profile and a retriever-specific evidence-state distribution. Controlled evidence interventions estimate the response profile once on a small calibration set, while offline retriever outputs determine each state distribution. Their combination yields the expected net gain and its lower confidence bound (LCB). This decomposition reuses generator response calibration across retrievers, while topology-aware states capture complementary evidence roles in multihop reasoning.

Our frozen evaluation covers multiple RAG mechanisms, two multihop question answering datasets, two generators, and sparse, dense, and fusion retrievers. EviCast achieves 90% decision agreement across 10 task configurations without falsely enabling any configuration that lacks a significant gain. On HotpotQA, 96 shared generations yield 14 correct decisions across 14 retrieval configurations, a 2.75% mean absolute error (MAE), and an 80% reduction in diagnostic generation calls. Across 16 topology configurations spanning datasets and generators, EviCast makes all 16 decisions correctly with an overall MAE of 2.81%. These results show that reusable calibration accurately forecasts retrieval value while substantially reducing the cost of diagnosing multiple configurations.

Our contributions are as follows:

- We formulate deployment-level, amortized retrieval necessity diagnosis to predict gains over the no-retrieval baseline across candidate retrievers under a shared calibration budget. Decisions combine evidence-state separability and the utility LCB.
- We develop EviCast to calibrate generator responses through controlled evidence interventions, map offline retriever outputs into evidence-state distributions, and combine both

components into expected retrieval utility. Its topology-aware instance models nonlinear complementarity between terminal and bridge evidence in multihop question answering.
- We validate EviCast across multiple RAG mechanisms, two multihop question-answering datasets, two generators, and sparse, dense, and fusion retrievers. EviCast achieves 14/14 correct decisions on HotpotQA retrieval configurations and 16/16 on cross-dataset topology configurations while reducing diagnostic generation calls by up to 80%.

2 Related Work

2.1 Retrieval-Augmented Generation And Evidence Use

Retrieval-augmented language models extend parametric knowledge with external memory. Early work established differentiable retrieval, dense passage retrieval, and joint retrieval and generation (Guu et al., 2020; Karpukhin et al., 2020; Lewis et al., 2020). Later systems fuse multiple passages, retrieve from large corpora, or augment frozen models with in-context evidence (Izacard & Grave, 2021; Borgeaud et al., 2022; Ram et al., 2023). Retrieval-Augmented Dual Instruction Tuning (RA-DIT) jointly adapts the generator and retriever, while RankRAG unifies context ranking with answer generation (Lin et al., 2024; Yu et al., 2024). Knowledge Intensive Language Tasks (KILT) provides a shared benchmark for knowledge-intensive tasks and provenance (Petroni et al., 2021).

External evidence is not uniformly beneficial. Irrelevant passages can distract generators, motivating evidence filtering and robustness training (Shi et al., 2023; Yoran et al., 2024; Amiraz et al., 2025). Evidence position and document composition also alter model behavior (Liu et al., 2024; Cuconasu et al., 2024), while retrieval gains vary with the knowledge already stored by the model (Mallen et al., 2023). These findings motivate EviCast to model retrieval supply and generator response separately rather than equating relevance with end-to-end utility.

2.2 Adaptive Retrieval And Selective Augmentation

Adaptive methods trigger retrieval from generation confidence, model self-knowledge, reflection, question complexity, or real-time information needs (Jiang et al., 2023; Wang et al., 2023a; Asai et al., 2024; Jeong et al., 2024; Su et al., 2024). Selective augmentation instead compresses or filters retrieved context before generation (Xu et al., 2024; Wang et al., 2023b). Unlike these query-level or step-level methods, EviCast forecasts configuration-level deployment utility using one generator calibration shared across candidate retrievers.

2.3 RAG Evaluation, Benchmarks, And Utility Estimation

RAGAS, the Automated RAG Evaluation System (ARES), and RAGChecker evaluate executed pipelines through context relevance, faithfulness, answer quality, and component-level diagnostics (Es et al., 2024; Saad-Falcon et al., 2024; Ru et al., 2024). The Retrieval-Augmented Generation Benchmark (RGB), Automatic LLMs' Citation Evaluation (ALCE), and RAGTruth assess robustness, information integration, citation quality, and hallucination (Chen et al., 2024; Gao et al., 2023; Niu et al., 2024). These tools characterize RAG behavior after execution. Utility modeling is closest to our objective. Prior work estimates the value of a query or document through semantic perplexity reduction, learned passage utility, confidence change, and retrieval-quality features (Dai et al., 2025; Perez-Beltrachini & Lapata, 2025; Wang et al., 2025; Tian et al., 2026). EviCast instead decomposes configuration-level net gain into interventional generator responses and retriever-specific evidence-state distributions, enabling amortized deployment diagnosis.

3 Method

EviCast forecasts end-to-end RAG utility before deployment. It measures a generator's response to controlled evidence states, then estimates how often each candidate retriever produces these states in offline results. Combining the shared response profile with each retriever distribution yields an expected net gain, a confidence interval (CI), and a deployment decision. This decomposition avoids complete RAG calibration for every retriever. Figure 2 shows the full data flow.

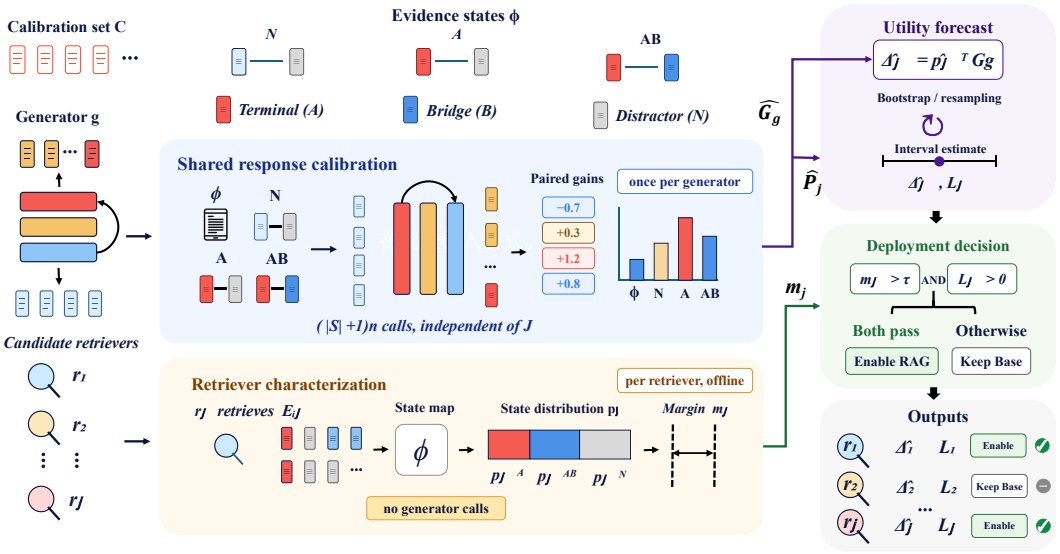

Figure 2: EviCast framework for reusable retrieval utility forecasting and deployment.

3.1 Problem Formulation

We define a fixed deployment environment as $\mathcal{E} = (g, \mathcal{D}, \pi, \mathcal{K}, k)$, where g is the generator, $\mathcal{D}$ the target task distribution, π the prompt template, $\mathcal{K}$ the retrieval corpus, and k the number of returned passages. Given candidate retrievers $\mathcal{R} = \{r_1, \ldots, r_J\}$, retriever r_j returns $E_j(x) = r_j(x; \mathcal{K}, k)$. Higher values of the task metric $M(\hat{y}, y)$ indicate better predictions.

The true deployment utility of r_j relative to no retrieval is:

$$\Delta_j = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[M(g_\pi(x, E_j(x)), y) - M(g_\pi(x, \emptyset), y) \right]. \tag{1}$$

The ideal deployment action is:

$$a_j^* = \mathbf{1}[\Delta_j > 0], \tag{2}$$

where one enables retrieval and zero retains no-retrieval inference.

Directly computing Δ_j requires separate RAG generations for every retriever. EviCast instead uses a labeled calibration set $\mathcal{C} = \{(x_i, y_i, \mathcal{P}_i, \rho_i)\}_{i=1}^n$ disjoint from the final test set. Here $\mathcal{P}_i$ contains candidate evidence, and ρ_i contains relevance or role annotations used only for calibration. EviCast returns $\widehat{\Delta}_j$, its CI, and $\widehat{a}_j$. The generator, task distribution, prompt, corpus, retrieval depth, and evidence format remain fixed during diagnosis, while only the offline rankings vary. The objective is to recover the sign and magnitude of Δ_j with generation cost independent of J.

3.2 Evidence States And The Topology-Aware Instantiation

EviCast uses an *evidence state* as its central intermediate variable rather than a relevance score or correct-passage count. For query x and evidence set E, the state mapping is:

$$s = \phi(x, E; \rho) \in \mathcal{S}, \tag{3}$$

where $\mathcal{S}$ is a mutually exclusive and exhaustive state space. Each state should preserve the evidence properties that determine generator utility. Within task stratum h, the response should be approximately invariant to the retriever that produced the evidence. Let $\delta(x, E) = M(g_\pi(x, E), y) - M(g_\pi(x, \emptyset), y)$. The desired response-invariance condition is:

$$\mathbb{E}[\delta(x, E) \mid h, s, r_j] \approx \mathbb{E}[\delta(x, E) \mid h, s]. \tag{4}$$

Retrievers may change state probabilities but should not systematically change the response within a state. This condition allows one generator response calibration to be reused across retrievers.

Multihop question answering requires evidence with distinct, complementary roles. Let e^A denote the terminal passage that directly contains answer information, and let $\mathcal{B}$ denote bridge passages that

identify an intermediate entity or connect the reasoning chain. For a fixed top-2 evidence set E, the topology-aware state mapping is defined as follows:

$$\phi(E) = \begin{cases} N, & e^A \notin E, \\ A, & e^A \in E,\ E \cap \mathcal{B} = \emptyset, \\ AB, & e^A \in E,\ E \cap \mathcal{B} \neq \emptyset. \end{cases} \tag{5}$$

State N lacks terminal evidence, state A contains only terminal evidence, and state AB contains terminal evidence with at least one bridge passage. This topology distinguishes a bridge passage paired with a distractor from a terminal passage paired with a distractor, although both sets contain one annotated supporting passage.

For each calibration example, EviCast constructs length-matched interventions for all three states. The N intervention excludes terminal evidence, prioritizes a bridge passage, and fills the remaining position with a distractor. The A intervention contains terminal evidence and one distractor, while the AB intervention contains terminal and bridge evidence. A prespecified anchor retriever selects and orders candidates after their roles are fixed. Evidence states, task strata, and ordering rules are frozen before calibration generation.

3.3 Reusable Interventional Response Calibration

Observed retrieval alone cannot distinguish missing evidence from a generator's failure to use available evidence. EviCast therefore intervenes on evidence states to measure the generator response separately. For item i, the no-retrieval score, controlled score, and paired gain are:

$$\begin{aligned} b_i &= M(g_\pi(x_i, \emptyset), y_i), \\ q_{i,s} &= M(g_\pi(x_i, \widetilde{E}_{i,s}), y_i), \quad d_{i,s} = q_{i,s} - b_i, \end{aligned} \tag{6}$$

where $\widetilde{E}_{i,s}$ is the controlled evidence for state s. Pairing reduces question-difficulty variation.

Let $\mathcal{C}_h = \{i : h_i = h\}$ and $n_h = |\mathcal{C}_h|$. The response of generator g in stratum h and state s is:

$$\widehat{G}_{g,h,s} = \frac{1}{n_h} \sum_{i \in \mathcal{C}_h} d_{i,s}. \tag{7}$$

All retrievers share the baseline and controlled generations. Thus, $\widehat{G}$ is computed once for a fixed generator, prompt, and evidence format. Changing retrievers requires no new response generation.

3.4 Retriever Evidence-State Characterization

Each candidate retriever runs offline on the shared calibration set. Let $E_{ij} = r_j(x_i; \mathcal{K}, k)$ and $s_{ij} = \phi(x_i, E_{ij}; \rho_i)$. The state probability for retriever r_j within stratum h is:

$$\widehat{p}_{j,h,s} = \frac{1}{n_h} \sum_{i \in \mathcal{C}_h} \mathbf{1}[s_{ij} = s]. \tag{8}$$

This mapping places heterogeneous retrievers in a common state space and combines them with the shared response profile without additional generator calls.

EviCast also tests whether a retrieval representation separates task-relevant evidence from distractors. Let A_j be a retrievability statistic and A_0 its uninformative baseline. The retrievability margin is defined as follows:

$$m_j = A_j - A_0. \tag{9}$$

For ranked question answering, A_j is the pairwise area under the receiver operating characteristic curve (AUC) for ranking supporting passages above distractors, and $A_0 = 0.5$. For case retrieval, A_j is held-out prediction performance based on retrieval keys, and A_0 is the majority baseline. This excess statistic provides a common separability gate across retrieval mechanisms.

3.5 Utility Forecast And Deployment Decision

The response profile quantifies the value of each evidence state, while the retriever distribution gives its frequency. The law of total expectation yields:

$$\widehat{\Delta}_j = \sum_{h \in \mathcal{H}} \widehat{w}_h \sum_{s \in \mathcal{S}} \widehat{p}_{j,h,s} \widehat{G}_{g,h,s}, \qquad \widehat{w}_h = \frac{n_h}{n}. \tag{10}$$

The utility forecast weights the generator response to each state by the probability that the retriever supplies that state. Under Equation 4, a new retriever requires only a new estimate of $\widehat{p}$, not $\widehat{G}$.

Both terms are uncertain when estimated from a small calibration set. EviCast propagates their joint uncertainty through a paired stratified bootstrap. Each replicate samples calibration items with replacement within h and recomputes the response profile, state probabilities, and utility. Given B replicates, the utility LCB is:

$$L_j = \mathrm{Quantile}_{0.025}\left(\left\{\widehat{\Delta}_j^{*(b)}\right\}_{b=1}^{B}\right). \tag{11}$$

The final deployment rule combining evidence separability with utility confidence is:

$$\widehat{a}_j = \mathbf{1}[m_j > \tau]\mathbf{1}[L_j > 0]. \tag{12}$$

Retrieval is enabled only when the representation accesses task-relevant evidence and the expected gain has a positive 95% LCB. Forecast $\widehat{\Delta}_j$ supports retriever ranking and resource planning.

3.6 Amortized Diagnostic Cost

With n calibration items, $|\mathcal{S}|$ evidence states, and J candidate retrievers, the diagnostic generation costs are:

$$C_{\mathrm{EviCast}} = (|\mathcal{S}| + 1)n, \qquad C_{\mathrm{Direct}} = (J + 1)n. \tag{13}$$

For the top-2 topology instance, $|\mathcal{S}| = 3$, so EviCast uses $4n$ generator calls for a fixed generator regardless of J. Adding a retriever requires offline retrieval and state counting but no additional generation. Shared calibration is therefore amortized whenever generator calls dominate diagnostic cost and $J > |\mathcal{S}|$. Appendix A.1 provides the complete notation and diagnostic algorithm.

4 Experiments

We evaluate EviCast along four dimensions. First, we assess retrieval necessity decisions across tasks and RAG paradigms. The cross-task settings cover Retrieval-Augmented Emotion Reasoning (RAER), Retrieval Augment for Missing Modality Multimodal Emotion Recognition (RAMER), and k-nearest-neighbor (kNN) label evidence on Explainable Multimodal Emotion Recognition (EMER), the Multimodal EmotionLines Dataset (MELD), and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Here A, V, and L denote acoustic, visual, and language modalities. Second, we test whether response calibration transfers across candidate retrievers and reduces diagnostic cost. Third, we evaluate generalization across datasets and generators. Finally, state-representation ablations and controlled-response analyses examine the role of evidence topology. All experiments use disjoint calibration and test data, and every deployment decision is frozen before observing the final test results.

4.1 Evaluation Protocol

Scope and Protocol. The evaluation comprises ten cross-task units spanning case-based emotion reasoning, missing-modality augmentation, label evidence, and dense passage RAG (Wen et al., 2025; Fan et al., 2024; Ma et al., 2024; Lewis et al., 2020). We evaluate shared calibration across 14 HotpotQA retrieval configurations and topology-aware diagnosis on MuSiQue and HotpotQA with two generators and four dense, sparse, and fusion retrievers. A separately frozen transfer evaluates Qwen2.5 text-only generators at 3B and 7B scales. Every decision is fixed before independent testing, and calibration and test examples are disjoint. Within each comparison, all methods use the

Table 1: EviCast retrieval decisions and independent test gains across ten task configurations.

Mechanism	Dataset / setting	Sep. margin (%)	Utility LCB (%)	Decision	Test gain [95% CI]	Agree
RAER	EMER	+5.56	+0.00	Disable	-0.25 $[-0.76, 0.00]$	✓
RAER	MELD	-12.05	+0.00	Disable	0.00 $[0.00, 0.00]$	✓
kNN Label	RAVDESS	+62.50	+18.46	**Enable**	$+27.50$ $[+19.50, +35.50]$	✓
RAMER	EMER / A	+0.00	-12.18	Disable	$+3.69$ $[-2.01, +9.53]$	✓
RAMER	EMER / V	+2.78	-7.34	Disable	$+1.32$ $[-2.36, +5.01]$	✓
RAMER	EMER / L	+48.61	-3.80	Disable	$+6.71$ $[+1.54, +11.60]$	×
RAMER	EMER / AV	+2.78	+0.41	Disable	-0.13 $[-2.86, +2.57]$	✓
RAMER	EMER / AL	+44.44	-4.34	Disable	-3.37 $[-9.55, +2.74]$	✓
RAMER	EMER / VL	+16.67	-8.73	Disable	-2.85 $[-5.94, +0.16]$	✓
Dense RAG	HotpotQA / BGE-M3	+34.38	+22.49	**Enable**	$+41.15$ $[+34.96, +47.34]$	✓
					Overall	**9/10 agreement, 0 false enables**

same queries, prompts, retrieval depth, evidence format, and answer parser. Appendices A and G provide complete splits, evidence-role rules, inference settings, and external-method adaptations.

Evaluation. We use accuracy for classification and normalized token F1 for open-ended question answering. We evaluate utility forecasts by their MAE against independent test gains and by Pearson and Spearman correlations. A configuration has a significant positive gain only when the 95% LCB of its paired test utility exceeds zero. We report agreement with this criterion, false enabling decisions, and diagnostic generation calls. Call counts exclude final test inference common to all methods. Appendix A details baselines, state ablations, bootstrapping, and metrics.

4.2 Retrieval Necessity Across Tasks And RAG Paradigms

Table 1 reports the frozen EviCast decisions for ten task configurations. EviCast agrees with the independent test conclusion in 9 of 10 cases, yielding 90.0% decision accuracy with no false enabling decisions. It correctly enables kNN label evidence on RAVDESS and Dense Passage RAG on HotpotQA, which yield final gains of 27.50% and 41.15%, respectively. EviCast disables the other eight configurations, seven of which lack stable positive test gains. These results span different objectives, evidence forms, and injection mechanisms, showing that EviCast diagnoses the end-to-end gain from external evidence rather than a local property of one RAG implementation. Figure 4 shows a retrospective EMER case in which the configuration-level decision retains the correct Base answer instead of the extra emotion label produced by retrieval.

Figure 3 projects the ten configurations onto the two-gate decision plane in Equation 12. RAMER core with AV has a slightly positive utility LCB but a retrieval-separability margin below 5%, so the separability gate prevents deployment based on this small estimate. RAMER core with AL has a 44.44% margin but a utility LCB of -4.34%, so the utility gate rejects a retrievable signal that does not yield actual gain. RAVDESS and HotpotQA satisfy both conditions and achieve large improvements on independent test sets. These cases show that evidence retrievability and model utility impose distinct deployment requirements whose conjunction defines an executable decision.

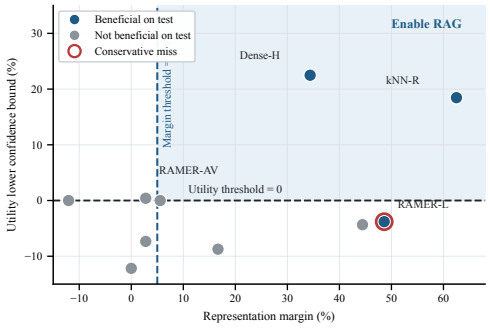

Figure 3: Two-gate decision plane for ten cross-task configurations.

4.3 Reusable Calibration Across Retrievers

Table 2 evaluates the accuracy and efficiency of shared calibration. Across 14 HotpotQA retrieval configurations, one EviCast calibration uses 96 generation calls and makes 14 of 14 correct deployment decisions, with a 2.75% MAE and Pearson and Spearman correlations of 0.983 and 0.890. Relative to Direct, EviCast reduces forecast error by 71.3% and generation calls by 80.0%. Table 3 compares EviCast with retrieval performance prediction (RPP), unnormalized Weighted Information Gain (U-WIG), and BERT-based Query Performance Prediction (BERT-QPP) under matched budgets. All scalar baselines are linearly calibrated to F1 gain on the calibration split. Across 50

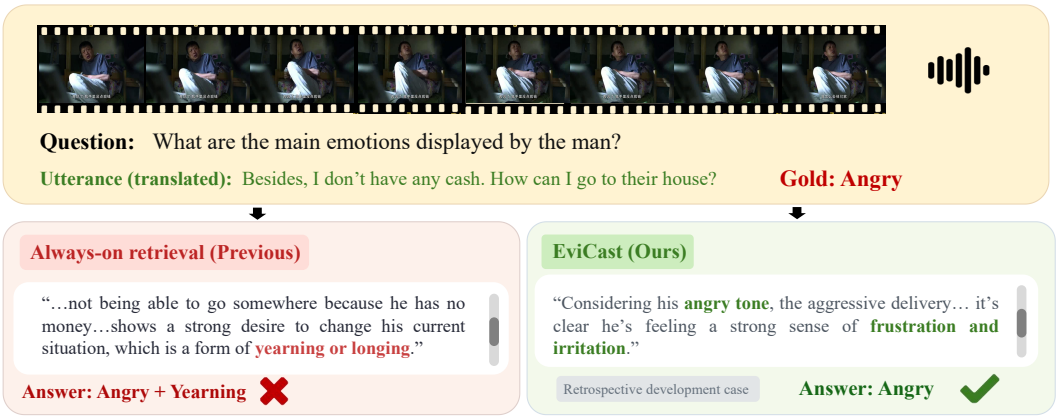

Figure 4: EMER example with video frames, recorded generator reasoning excerpts, and the EviCast configuration-level decision.

calibration subsamples and 512 disjoint test queries, EviCast achieves the lowest MAE at both low-budget settings. At 160 and 320 generations, it reduces error by 16.4% and 10.7% over the strongest competing result reported in Table 3 at each budget, and by 62.4% and 25.3% over RPP. At 512 generations, it also remains more accurate than Direct, RPP, U-WIG, and BERT-QPP.

Table 2: Shared calibration on 14 HotpotQA configurations with MAE in percent and generation calls.

Method	Agree	Pearson	Spearman	MAE (%) ↓	Calls ↓
EviCast	**14/14**	**0.983**	**0.890**	**2.75**	**96**
Direct	14/14	0.818	0.332	9.59	480

Across 12 standard deployable configurations, EviCast obtains 12 of 12 correct decisions and a 2.22% MAE, compared with 10.20% for Direct, while reducing generation calls by 76.92%. Among eight configurations added under a frozen protocol, EviCast achieves a lower absolute forecast error in seven. These results support the decomposition in Figure 2, where one generator response profile serves dense, sparse, and fusion retrievers through their state distributions. Appendix C and Figure 6 provide complete per-configuration results and forecasts. Appendix H reports the full budget analysis, while Figure 7 shows how diagnostic savings scale with retriever count.

Table 3: Matched-budget MAE (%) on HotpotQA with Qwen2.5-3B-Instruct and four retrievers.

Method	160	320	512
EviCast (Ours)	**5.73**	**4.34**	**3.86**
Direct	6.85	4.89	3.92
RPP (Tian et al., 2026)	15.25	5.81	4.77
U-WIG (Dado et al., 2026)	7.04	4.89	3.94
BERT-QPP (Arabzadeh et al., 2021)	6.99	4.86	3.87

Table 4: Decision agreement and MAE (%) across datasets and generators.

Setting	Agree	MAE ↓
MuSiQue / Ours + Qwen (Xu et al., 2025)	**4/4**	**1.53**
MuSiQue / Ours + Phi-4 (Abouelenin et al., 2025)	**4/4**	**1.55**
HotpotQA / Ours + Qwen (Xu et al., 2025)	**4/4**	4.59
HotpotQA / Ours + Phi-4 (Abouelenin et al., 2025)	**4/4**	**3.58**
2Wiki / Ours + Qwen (Xu et al., 2025)	3/4	4.70
2Wiki / Ours + Phi-4 (Abouelenin et al., 2025)	**4/4**	**1.15**
Three-dataset total	**23/24**	**2.85**
Ours + Qwen2.5-Text (Yang et al., 2024) (3B+7B)	**8/8**	**3.67**

4.4 Generalization Across Datasets And Generators

Table 4 evaluates the complete EviCast pipeline on MuSiQue and HotpotQA with two generators. EviCast makes 16 of 16 correct decisions with a 2.81% MAE. We then conduct a separately frozen prospective transfer to 2WikiMultiHopQA. Across two generators and four retrievers, EviCast makes 7 of 8 correct decisions with no false enabling decisions, achieves a Pearson correlation of 0.955 between forecast and test gains, and uses 20% fewer diagnostic calls than Direct calibration. Across all three datasets, it makes 23 of 24 frozen deployment decisions correctly with a 2.85% MAE. We also evaluate four retrievers with Qwen2.5 text-only generators at 3B and 7B scales. Across eight configurations, each tested on 512 independent queries, EviCast makes 8 of 8 correct decisions with no false enables and a 3.67% aggregate MAE. These results show that the response-calibration and state-distribution mechanism transfers across datasets, generator families, and model scales. Appendices D.5 and D.6 report the complete prospective and text-only results.

4.5 Ablation And Mechanism Analysis

Evidence-State Representation. Table 5 varies only the evidence-state representation while fixing datasets, retrievers, generators, and retrieval statistics. Linear interpolation between fully supportive and fully misleading endpoints yields an overall MAE of 8.30%. Count distinguishes nonlinear responses to zero, one, and two correct passages, reducing the MAE to 3.87%. The complete EviCast representation further separates terminal and bridge roles and reduces the MAE to 2.81%. EviCast lowers forecast error by 66.1% relative to Linear and 27.4% relative to Count while preserving 16 of 16 correct deployment decisions.

Direct generates Actual calibration outputs separately for each candidate retriever and obtains an overall MAE of 2.77%. EviCast achieves comparable forecast accuracy with a 2.81% MAE while requiring 512 rather than 640 diagnostic calls per generator and dataset setting, a reduction of 20.0%. Topology states therefore obtain their accuracy from structured, reusable response calibration rather than a larger generation budget.

Table 5: Decision agreement and MAE (%) for evidence-state ablations and Direct calibration.

Method	MuSiQue agree	MuSiQue MAE	HotpotQA agree	HotpotQA MAE	Overall agree	Overall MAE	Gen. calls
Linear	8/8	7.24	8/8	9.35	16/16	8.30	384
Count	8/8	3.49	8/8	4.25	16/16	3.87	512
EviCast	**8/8**	**1.54**	**8/8**	**4.08**	**16/16**	**2.81**	**512**
Direct calibration reference							
Direct	8/8	1.79	8/8	3.75	16/16	2.77	640

Why Evidence Topology Matters. Figure 5 and Appendix Table 12 show why correct-passage counts are insufficient. States N and A each contain one correct supporting passage, but N provides only bridge information, whereas A directly contains terminal answer information. Their F1 gains are 4.55% and 50.95% for Qwen and 2.58% and 45.65% for Phi-4. Thus, one correct passage does not define a utility-homogeneous state. Reusable profiles must distinguish terminal and bridge roles.

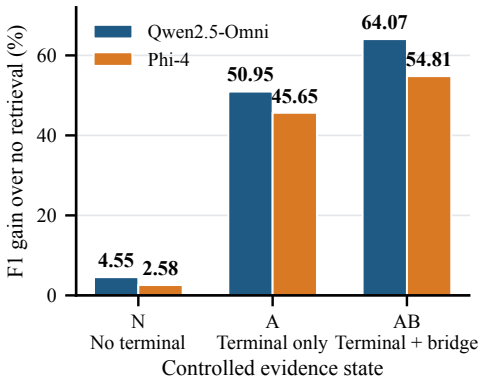

Figure 5: Controlled-state F1 gains on MuSiQue.

The controlled responses also explain the gap between Count and EviCast in Table 5. Count merges N and A because both contain one correct passage, forcing them to share a response value. EviCast estimates their responses separately and combines them using each retriever's probabilities of returning N, A, and AB. It therefore distinguishes retrievers that return the same number of correct passages but different proportions of terminal and bridge evidence. This distinction explains the lower forecast error of the topology-aware representation and links the states to deployment decisions. The comparison between Count and EviCast is particularly informative because both use the same number of diagnostic calls. Their accuracy difference therefore isolates the value of distinguishing evidence roles, rather than a larger calibration budget or more evidence states.

5 Conclusion

Retrieval deployment should depend on how a target generator uses evidence, not only on its availability. EviCast forecasts utility by combining shared interventional response calibration with retriever-specific evidence-state distributions. Experiments show that response calibration transfers across retrievers and that terminal and bridge roles explain utility differences hidden by passage counts. On HotpotQA, EviCast makes all 14 deployment decisions correctly while reducing diagnostic calls by 80%. These results shift RAG from default retrieval toward utility-guided deployment, with one calibration shared by multiple retrievers. Future work will extend evidence states to variable depths, iterative retrieval, and multimodal knowledge environments.

AI USE STATEMENT

We used generative artificial intelligence (AI) tools to assist with code implementation and debugging, experimental result analysis, literature searches, manuscript translation, and preliminary figure drafts. AI-generated suggestions, code, and text served as supporting material and did not replace experimental validation or the authors' scientific judgment. The authors verified the methods, code, references, result interpretations, and figures in the final manuscript and take responsibility for its content and accuracy.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. The distracting effect: Understanding irrelevant passages in rag. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18228–18258, 2025.

Negar Arabzadeh, Maryam Khodabakhsh, and Ebrahim Bagheri. Bert-qpp: contextualized pre-trained transformers for query performance prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 2857–2861, 2021.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pp. 9112–9141, 2024.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.

Xuejun Chang, Debabrata Mishra, Craig Macdonald, and Sean MacAvaney. Neural passage quality estimation for static pruning. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 174–185, 2024.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 17754–17762, 2024.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 719–729, 2024.

Or Dado, David Carmel, and Oren Kurland. Rag performance prediction for question answering. *arXiv preprint arXiv:2604.07985*, 2026.

Lu Dai, Yijie Xu, Jinhui Ye, Hao Liu, and Hui Xiong. Seper: Measure retrieval utility through the lens of semantic perplexity reduction. In *International Conference on Learning Representations*, volume 2025, pp. 78977–79007, 2025.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, pp. 150–158, 2024.

Qi Fan, Hongyu Yuan, Haolin Zuo, Rui Liu, and Guanglai Gao. Leveraging retrieval augment approach for multimodal emotion recognition under missing modalities. *arXiv preprint arXiv:2410.02804*, 2024.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pp. 874–880, 2021.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 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: Human language technologies (volume 1: Long papers)*, pp. 7036–7050, 2024.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 7969–7992, 2023.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6769–6781, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. Ra-dit: Retrieval-augmented dual instruction tuning. In *International Conference on Learning Representations*, volume 2024, pp. 19138–19162, 2024.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15747–15760, 2024.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, 2024.

Laura Perez-Beltrachini and Mirella Lapata. Uncertainty quantification in retrieval augmented question answering. *arXiv preprint arXiv:2502.18108*, 2025.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. Kilt: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2523–2544, 2021.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems*, 37:21999–22027, 2024.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 338–354, 2024.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International conference on machine learning*, pp. 31210–31227. PMLR, 2023.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 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 (Volume 1: Long Papers)*, pp. 12991–13013, 2024.

Yongquan Tao and Shengli Wu. Query performance prediction by considering score magnitude and variance together. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pp. 1891–1894, 2014.

Fangzheng Tian, Debasis Ganguly, and Craig Macdonald. Predicting retrieval utility and answer quality in retrieval-augmented generation. In *European Conference on Information Retrieval*, pp. 368–385. Springer, 2026.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. Self-knowledge guided retrieval augmentation for large language models. In *Findings of the association for computational linguistics: EMNLP 2023*, pp. 10303–10315, 2023a.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*, 2023b.

Zihan Wang, Zihan Liang, Zhou Shao, Yufei Ma, Huangyu Dai, Ben Chen, Lingtao Mao, Chenyi Lei, Yuqing Ding, and Han Li. Infogain-rag: Boosting retrieval-augmented generation through document information gain-based reranking and filtering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7190–7204, 2025.

Zhuofan Wen, Zheng Lian, Shun Chen, Hailiang Yao, Longjiang Yang, Bin Liu, and Jianhua Tao. Listen, watch, and learn to feel: Retrieval-augmented emotion reasoning for compound emotion generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11313–11327, 2025.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with context compression and selective augmentation. In *International Conference on Learning Representations*, volume 2024, pp. 43478–43502, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, volume 2024, pp. 29862–29883, 2024.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *Advances in neural information processing systems*, 37:121156–121184, 2024.

APPENDIX

A COMPLETE EXPERIMENTAL PROTOCOL AND IMPLEMENTATION DETAILS

This appendix expands the aggregate retrieval-augmented generation (RAG) results in the main paper and documents the data splits, controlled interventions, retriever configurations, statistical procedures, and frozen-decision audit. Except for the mechanism studies, every deployment decision was recorded before inference on the independent test set or construction of the final retrieval cache. Test answers, supporting evidence, and evidence roles were not used for calibration or deployment decisions.

A.1 NOTATION, INTERVENTIONS, AND DIAGNOSTIC ALGORITHM

Table 6 defines the EviCast notation. In the topology-aware instantiation, each calibration example receives length-matched interventions. State N excludes terminal evidence, prioritizes a bridge passage, and fills the remaining position with a distractor. State A contains terminal evidence and one distractor, while state AB contains terminal and bridge evidence. A prespecified anchor retriever selects and orders candidates after their roles are fixed. Evidence states, task strata, and ordering rules are frozen before calibration.

For ranked question answering, the retrievability statistic A_j is the pairwise area under the receiver operating characteristic curve (AUC) between supporting passages and distractors, with $A_0 = 0.5$. For case retrieval, A_j is the held-out prediction performance from retrieval keys, with majority-class performance as A_0. We use $B = 10{,}000$ paired stratified bootstrap replicates and fix the separability threshold at $\tau = 0.05$.

Table 6: Core notation used by EviCast.

Symbol	Meaning
$\mathcal{E}, \mathcal{C}$	Fixed deployment environment and calibration set
$\mathcal{R}, \mathcal{S}$	Candidate retrievers and evidence states
$\phi(x, E; \rho)$	Evidence state mapping
$\widehat{G}_{g,h,s}$	Generator response in stratum h and state s
$\widehat{p}_{j,h,s}$	State probability produced by retriever r_j
m_j, τ	Retrievability margin and its threshold
$\widehat{\Delta}_j, L_j$	Utility forecast and its 95% lower confidence bound (LCB)
$\widehat{a}_j$	Final retrieval deployment decision

Algorithm 1 calibrates the shared generator response before evaluating any number of retrievers. Retrieved evidence is used only to map outputs to states and estimate their distribution. It is not provided to the generator during response calibration.

Algorithm 1 EviCast retrieval necessity diagnosis

Input Environment $\mathcal{E}$, calibration set $\mathcal{C}$, retrievers $\mathcal{R}$, state map ϕ, and threshold τ
1 Generate the no-retrieval baseline and controlled output for every $s \in \mathcal{S}$.
2 Compute the shared response profile $\widehat{G}$ with Equation 7.
3 For every $r_j \in \mathcal{R}$, run offline retrieval and map each result to a state.
4 Estimate the state distribution $\widehat{p}_j$ and retrievability margin m_j.
5 Bootstrap $\widehat{\Delta}_j$ and its LCB L_j.
6 Set $\widehat{a}_j$ with Equation 12.
Output $\{(\widehat{\Delta}_j, L_j, \widehat{a}_j)\}_{j=1}^{J}$

A.2 Data Splits And Isolation Rules

Table 7 summarizes the calibration and test boundaries. The cross-paradigm settings cover Retrieval-Augmented Emotion Reasoning (RAER), Retrieval Augment for Missing Modality Multimodal Emotion Recognition (RAMER), and k-nearest-neighbor (kNN) label evidence on Explainable Multimodal Emotion Recognition (EMER), the Multimodal EmotionLines Dataset (MELD), and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). A, V, and L denote acoustic, visual, and language modalities. These experiments use at most 32 calibration examples. MuSiQue v4 excludes the 544 examples used by earlier Linear, Count, and cross-generator analyses. HotpotQA v4 excludes the 288 examples used by the initial diagnosis and cross-retriever study. Both v4 studies use newly selected, nonoverlapping calibration and test sets.

Table 7: Data splits and frozen boundaries. Store denotes a case base or training pool that is disjoint from the final queries. Excluding prior examples removes every development identifier (ID) used in earlier analyses before the new split is selected.

Experiment	Dataset	Task	Store	Cal.	Test	Stratification	Isolation rule
RAER-core	EMER	Multi-label emotion	72	32	132	Coarse emotion	Store/query disjoint
RAER-core	MELD	Emotion recognition	332	32	241	Emotion class	Store/query disjoint
kNN Label-Evidence	RAVDESS	Speech emotion	Train pool	32	200	Six coarse emotions	Cal. from 300 non-test samples
RAMER-core	EMER	Missing-modality emotion	40	32	133	Coarse emotion	Six settings share test queries
Dense Passage RAG	HotpotQA	Open multihop question answering (QA)	10 passages/query	32	256	Type×difficulty	Zero cal./test overlap
Retriever reuse	HotpotQA	Open multihop QA	10 passages/query	32	256	Type×difficulty	14 settings share one split
Topology v4	MuSiQue	2/3/4-hop QA	∼20 passages/query	128	512	Question hops	544 prior IDs excluded
Topology v4	HotpotQA	Two-hop QA	10 passages/query	128	512	Type×difficulty	288 prior IDs excluded
Prospective transfer	2WikiMultiHopQA	Two-hop QA	10 passages/query	128	512	Question type	Zero cal./test overlap

The initial HotpotQA experiment uses the official distractor validation set. Of its 7,405 examples, 7,345 contain ten candidate passages and exactly two supporting titles. The 32-example calibration set contains 26 bridge-hard and 6 comparison-hard questions, while the 256-example test set contains 206 and 50. The MuSiQue v4 calibration set contains 66, 40, and 22 questions with two, three, and four hops. The corresponding test counts are 265, 161, and 86. The HotpotQA v4 calibration set contains 105 bridge-hard and 23 comparison-hard questions, while its test set contains 420 and 92. The 2WikiMultiHopQA calibration set contains 98 compositional and 30 inference questions, while its independent test set contains 393 and 119. All splits use prespecified seeds and deterministic ordering by example ID.

A.3 Models, Retrievers, And Inference

The multimodal classification experiments use a Stage0a instruction-tuned Qwen2.5-Omni-7B model. RAER-core separately normalizes 384-dimensional Sentence-BERT (SBERT) features, 4,096-dimensional Contrastive Language-Image Pre-training (CLIP) features from eight frames, and 768-dimensional Audio Spectrogram Transformer (AST) features. Their concatenation forms the retrieval key with fixed $k = 5$. RAMER-core uses the same modalities and retrieves neighbors from the observed modalities to fill missing ones with normalized neighbor features. RAVDESS retrieves $k = 1$ nearest-neighbor label evidence using emotion2vec acoustic representations.

The multihop question-answering experiments use Qwen2.5-Omni-7B thinker and Phi-4-multimodal-instruct in text-only mode. Both use deterministic greedy decoding, a batch size of 4, and at most 64 generated tokens. We use the following fixed system prompt.

```
Answer the question using the supplied evidence when
present.  Return only the concise answer, without
explanation or formatting.
```

The no-retrieval prompt is formatted as `Question: q; Answer:`. Evidence conditions place numbered full passages before the question. The answer parser selects the first nonempty generated line and mechanically removes an optional `Answer:` prefix. All raw generations are retained for audit.

The four base retrievers return the top two full passages. BGE-M3 encodes the question and each title concatenated with its passage, then ranks candidates by cosine similarity. BM25 uses $k_1 = 1.5$ and $b = 0.75$. Term frequency inverse document frequency (TF-IDF) uses cosine similarity over lowercased word unigrams and bigrams. Reciprocal Rank Fusion (RRF) combines BGE-M3 and

BM25 rankings with $c = 60$. Ties follow the original candidate order. Title and lead-sentence variants alter only the ranking text. The generator always receives the full passage.

A.4 Baselines And Evidence-State Representations

Base answers without retrieval and defines the net-utility reference. Always-on RAG injects retrieved results for every test query. Direct generates Actual calibration outputs for each retriever and estimates utility from paired Actual and Base differences. It requires new generator calls for every candidate configuration.

The state-representation ablation compares Linear, Count, and the full EviCast representation. Linear measures fully supportive and misleading endpoints, then interpolates by the proportion of correct evidence slots. Count partitions fixed top-2 results into states with zero, one, or two correct passages. EviCast uses the topology-aware $N/A/AB$ states in Equation 5. All representations share calibration examples and retrieval statistics, isolating the state representation. Non-multihop tasks and the 14-configuration reuse experiment use annotation-aligned Supportive and Misleading states. MuSiQue and HotpotQA use the topology-aware representation.

A.5 Metrics, Confidence Intervals, And Generation Calls

Multi-label emotion recognition uses mean example-level F1, single-label classification uses accuracy, and open-ended question answering uses normalized token F1 with exact match as a secondary metric. Per-example gain is RAG performance minus Base performance. We use 10,000 paired stratified bootstrap replicates that jointly recompute evidence-state responses, retriever state distributions, and expected utility. A configuration has a significant positive gain when the 95% LCB of its independent test gain exceeds zero. Each estimated gain includes a 95% confidence interval (CI).

Diagnostic calls count only generator invocations required for deployment decisions and exclude shared final-test inference. Given calibration size n, state count $|\mathcal{S}|$, and J candidate retrievers, shared response calibration requires $(|\mathcal{S}| + 1)n$ calls. Direct calibration requires $(J + 1)n$ calls. Offline retrieval, state mapping, and bootstrap estimation require no generator calls.

B Complete Cross-Task Results For The Two-Gate Diagnosis

B.1 Calibration Statistics

Table 8 expands the calibration statistics in Table 1. Here ρ is the mean proportion of supportive evidence slots in the Actual retrieval output, while G^+ and G^- are the gains of Supportive and Misleading evidence over Base. Actual serves only as a comparator and is not used by EviCast.

Table 8: Complete calibration results for ten cross-task configurations. All values are percentages. Brackets following $\widehat{\Delta}$ and Actual contain 95% CIs.

Paradigm / dataset	Setting	Margin	ρ	G^+	G^-	$\widehat{\Delta}$ [CI]	Actual cal. gain [CI]
RAER-core / EMER	N/A	+5.56	30.63	0.00	0.00	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
RAER-core / MELD	N/A	−12.05	27.50	0.00	0.00	0.00 [0.00, 0.00]	+3.13 [0.00, +9.38]
kNN Label-Evidence / RAVDESS	N/A	+62.50	68.75	+46.88	+3.13	+33.20 [+18.46, +48.44]	+40.63 [+25.00, +56.25]
RAMER-core / EMER	A	0.00	50.00	−11.60	+10.10	−0.75 [−12.18, +10.88]	−2.81 [−15.97, +10.73]
RAMER-core / EMER	V	+2.78	54.38	−4.60	+5.03	−0.21 [−7.34, +6.51]	+3.93 [−5.63, +12.98]
RAMER-core / EMER	L	+48.61	46.25	+1.09	+4.43	+2.89 [−3.80, +9.58]	+10.73 [+2.86, +19.26]
RAMER-core / EMER	AV	+2.78	51.88	−0.48	+11.00	+5.05 [+0.41, +10.04]	+2.86 [−1.67, +8.16]
RAMER-core / EMER	AL	+44.44	52.50	−3.27	+16.13	+5.94 [−4.34, +16.47]	+4.25 [−7.22, +16.53]
RAMER-core / EMER	VL	+16.67	47.50	−5.30	−0.87	−2.97 [−8.73, +2.60]	−5.04 [−13.53, +3.08]
Dense Passage RAG / HotpotQA	BGE-M3	+34.38	68.75	+50.22	+5.16	+36.14 [+22.49, +49.39]	+24.18 [+7.83, +41.35]

The results expose two failure modes of single-signal rules. RAMER-AV has a positive utility LCB, but its separability margin is only 2.78% and its test gain is not significant. RAMER-AL has a 44.44% margin, but its utility CI crosses zero and its final gain is again not significant. RAVDESS and HotpotQA instead combine strong retrievability with a positive generator response, improving test performance by 27.50% and 41.15%.

B.2 Two-Gate Rule Ablation

Table 9 removes either gate across the same ten configurations. Separability alone falsely enables RAER-EMER, RAMER-AL, and RAMER-VL. Utility alone falsely enables RAMER-AV and misses RAMER-L. The complete rule eliminates false enables while retaining both strongly beneficial configurations.

Table 9: Ablation of the two-gate decision rule. A false enable deploys RAG when the final gain is not significant. A conservative reject retains Base when RAG has a significant positive gain.

Decision rule	Agreement	False enable	Conservative reject
Separability only, $m_j > 0.05$	7/10	3	0
Utility only, $L_j > 0$	8/10	1	1
EviCast two-gate rule	**9/10**	**0**	1

C Per-Configuration Results For Cross-Retriever Reuse

C.1 Fourteen Retrieval Configurations

This experiment fixes the HotpotQA questions, generator, and prompt while varying the retriever or ranking text. The base group contains BGE-M3, BM25, TF-IDF, and RRF fusion of BGE-M3 and BM25. The extended group applies dense, sparse, and fusion retrieval to titles, lead sentences, or text truncated to 128 tokens. Oracle prioritizes the two annotated supporting passages, while Anti-BGE reverses the BGE-M3 ranking. These controls define upper and lower evidence-quality conditions rather than deployable retrievers.

Table 10 reports forecasts and independent test results for all configurations. EviCast makes 14/14 correct deployment decisions, enabling all 13 beneficial configurations and rejecting Anti-BGE because its permutation CI crosses zero. Its forecasts achieve Pearson and Spearman correlations of 0.983 and 0.890 with a mean absolute error (MAE) of 2.75%. Direct calibration obtains 0.818, 0.332, and 9.59%, respectively.

Table 10: Complete reuse results for 14 HotpotQA retrieval configurations. Values are percentages, and brackets contain 95% CIs. Decisions use only Margin and the EviCast LCB. They never use the test gain.

Retrieval setting	Margin	ρ	EviCast $\widehat{\Delta}$ [CI]	Direct $\widehat{\Delta}$	Final gain [CI]	Decision
BGE-M3	34.38	68.75	36.14 [22.49, 49.39]	24.18	41.15 [35.02, 47.17]	Enable
BM25	25.59	60.94	32.62 [18.83, 46.65]	32.10	35.70 [29.41, 41.87]	Enable
TF-IDF	26.37	53.12	29.10 [15.81, 42.39]	25.22	30.89 [24.54, 37.19]	Enable
RRF (BGE+BM25)	34.08	62.50	33.32 [20.01, 46.14]	23.66	36.88 [30.99, 42.88]	Enable
Oracle	50.00	100.00	50.22 [33.13, 66.79]	50.22	58.41 [53.01, 63.71]	Enable
Anti-BGE	−50.00	0.00	5.16 [−8.65, 19.27]	5.16	1.43 [−3.62, 6.65]	**Keep Base**
BGE-title	26.17	57.81	31.21 [17.14, 44.76]	16.89	31.05 [24.82, 37.30]	Enable
BGE-lead	34.38	68.75	36.14 [21.76, 50.31]	27.31	35.49 [29.40, 41.65]	Enable
BGE-128	31.45	62.50	33.32 [20.51, 45.71]	16.47	35.80 [30.05, 41.55]	Enable
BM25-title	22.27	56.25	30.51 [15.86, 45.44]	29.69	28.59 [22.59, 34.73]	Enable
BM25-lead	23.63	50.00	27.69 [12.67, 42.68]	15.54	28.95 [22.70, 35.22]	Enable
TF-IDF-title	21.97	57.81	31.21 [16.61, 45.57]	26.56	29.53 [23.30, 35.68]	Enable
TF-IDF-lead	27.54	57.81	31.21 [16.69, 45.89]	21.16	30.27 [24.16, 36.42]	Enable
RRF (BGE+TF-IDF)	32.71	60.94	32.62 [19.32, 45.67]	22.10	36.74 [30.69, 43.01]	Enable

Figure 6 visualizes every forecast. EviCast remains close to the ideal diagonal across dense, sparse, fusion, and control configurations, while Direct shows larger magnitude and ranking errors. The consistent trend across configurations confirms that a small subset does not drive the aggregate correlation.

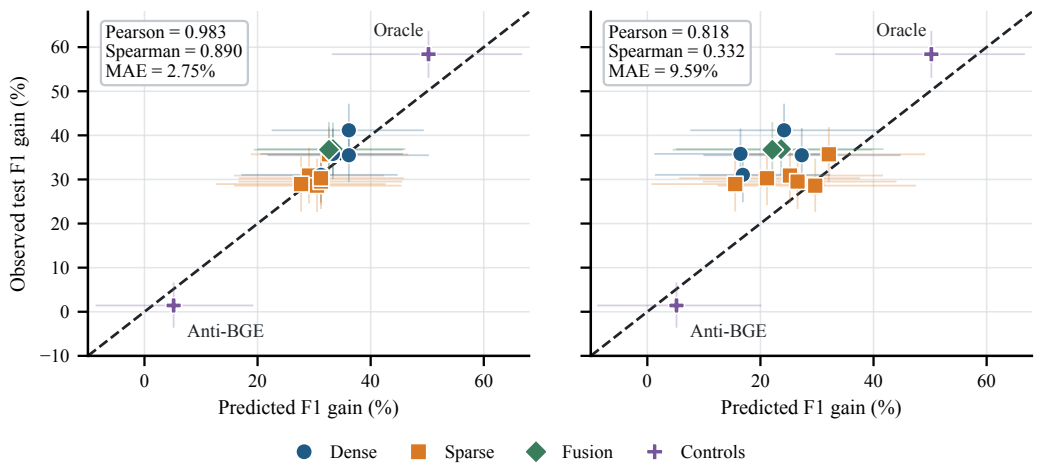

Figure 6: Forecast utility and independent test utility for 14 HotpotQA retrieval configurations. The left panel shows EviCast, and the right panel shows Direct calibration for each retriever. The dashed line denotes an ideal forecast. Error bars show 95% CIs.

C.2 Diagnostic Savings From Reuse

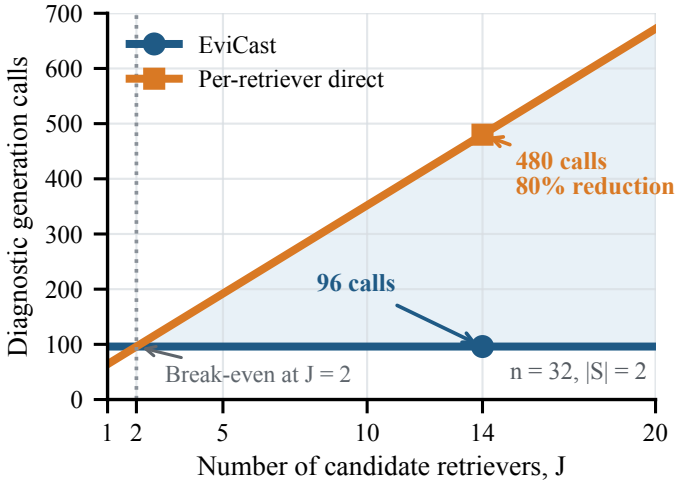

Figure 7: Diagnostic generation calls versus the number of candidate retrievers for shared EviCast and per-retriever Direct calibration.

All 14 configurations share the Base, Supportive, and Misleading responses, so EviCast requires $3 \times 32 = 96$ calibration generations. Direct requires Base and one Actual condition per configuration, yielding $15 \times 32 = 480$ generations. EviCast thus reduces diagnostic calls by 80.0%. Among the 12 deployable retrievers, it retains 12/12 correct decisions with 96 calls, reducing calls by 76.92% relative to Direct's 416 and forecast error from 10.20% to 2.22%.

These values measure only the one-time diagnostic cost. Rejecting an unhelpful configuration also saves recurring retrieval and context processing. Under our HotpotQA prompts, the average input grows from 58.30 tokens without retrieval to 290.04 tokens with top-2 RAG.

D Complete Validation Of Topology-Aware Evidence States

D.1 From Linear Slots To Topology-Aware States

The initial Linear instantiation treats each correct passage in a top-2 result as additive. Count instead calibrates nonlinear responses to zero, one, and two correct passages, but merges all cases containing exactly one correct passage. Cross-generator analysis shows that this aggregation obscures the functional distinction between terminal and bridge evidence. Topology-aware EviCast therefore defines N as missing terminal evidence, A as terminal evidence alone, and AB as terminal and bridge evidence together.

Table 11 summarizes this progression. On two new, larger frozen splits, Topology makes all 16 decisions correctly and reduces aggregate MAE to 2.81%.

Table 11: Evolution and independent validation of the evidence-state design.

Stage	Data and generator	Frozen settings	Agreement	Linear MAE	Count MAE	Topology MAE	Direct MAE
Initial Linear validation	MuSiQue / Qwen	4	4/4	9.33	N/A	N/A	2.86
Prospective Count validation	MuSiQue / Qwen	4	4/4	4.84	1.64	N/A	3.92
Cross-generator Count check	MuSiQue / Phi-4	4	2/4	1.02	6.54	N/A	11.02
Topology v4	MuSiQue / Qwen+Phi-4	8	**8/8**	7.24	3.49	**1.54**	1.79
Topology v4	HotpotQA / Qwen+Phi-4	8	**8/8**	9.35	4.25	**4.08**	3.75

D.2 Generator Evidence Response Curves

Table 12 reports controlled responses for both generators in every stratum. Count-1 pools N and A, while the topology columns expose their distinct effects. On two-hop MuSiQue, Qwen responses to $N/A/AB$ are $4.55\%/50.95\%/64.07\%$. Phi-4 shows the same step pattern at $2.58\%/45.65\%/54.81\%$. Thus, one correct passage does not define a unique response state, even for a fixed generator.

Table 12: Complete controlled evidence responses on MuSiQue and HotpotQA, reported as F1 gain over Base in percent. Count-0/1/2 groups examples by the number of correct passages. $N/A/AB$ groups them by evidence topology.

Dataset	Generator	Stratum	Count-0	Count-1	Count-2	N	A	AB
MuSiQue	Qwen	2-hop	2.99	20.95	64.07	4.55	50.95	64.07
MuSiQue	Qwen	3-hop	−0.11	15.90	31.50	3.41	49.48	57.91
MuSiQue	Qwen	4-hop	−8.32	0.91	6.66	−9.87	31.34	26.05
MuSiQue	Phi-4	2-hop	−0.26	17.72	55.07	2.58	45.65	54.81
MuSiQue	Phi-4	3-hop	0.84	8.46	25.83	3.19	36.74	49.92
MuSiQue	Phi-4	4-hop	−1.34	−3.85	19.20	−3.28	22.22	28.58
HotpotQA	Qwen	bridge-hard	3.74	24.41	61.48	9.68	58.95	61.48
HotpotQA	Qwen	comparison-hard	11.84	8.54	26.40	−5.95	27.64	26.40
HotpotQA	Phi-4	bridge-hard	0.60	16.96	55.75	4.42	42.49	55.75
HotpotQA	Phi-4	comparison-hard	3.70	11.24	38.65	0.61	23.68	38.65

D.3 Retriever Distributions Over Topology States

The response curves depend on the generator and task distribution, while the retriever determines each state's frequency. Table 13 reports the $N/A/AB$ distribution for each retriever. Because it is generator-independent, the same distribution combines with either the Qwen or Phi-4 response profile.

Table 13: Test distributions of four retrievers over topology states in percent.

Dataset	Retriever	N	A	AB	Margin
MuSiQue	BGE-M3	66.41	14.84	18.75	29.79
MuSiQue	BM25	83.59	7.03	9.38	15.37
MuSiQue	TF-IDF	80.47	6.25	13.28	18.44
MuSiQue	RRF	73.44	10.16	16.41	24.85
HotpotQA	BGE-M3	44.53	14.06	41.41	37.65
HotpotQA	BM25	39.06	24.22	36.72	30.91
HotpotQA	TF-IDF	50.78	17.19	32.03	30.13
HotpotQA	RRF	42.19	14.84	42.97	37.43

D.4 Per-Configuration Forecasts For Sixteen Frozen Settings

Tables 14 and 15 report the 16 configurations summarized in the main paper. On MuSiQue, Topology obtains a 1.54% MAE, compared with 7.24% for Linear, 3.49% for Count, and 1.79% for Direct. On HotpotQA, the corresponding values are 4.08%, 9.35%, 4.25%, and 3.75%. Overall, Topology makes 16/16 correct deployment decisions with a 2.81% MAE. Its error is lower than Linear in all 16 configurations and Count in 11.

Table 14: Eight frozen MuSiQue v4 configurations. The table reports forecast gains from Linear, Count, Topology, and Direct together with independent test gains in percent.

Generator	Retriever	Margin	Linear	Count	Topology [CI]	Direct	Final gain [CI]
Qwen	BGE-M3	29.79	24.64	21.38	18.73 [13.29, 24.19]	18.98	19.56 [16.12, 23.06]
Qwen	BM25	15.37	15.28	14.49	9.94 [4.46, 15.46]	7.74	12.24 [9.09, 15.56]
Qwen	TF-IDF	18.44	19.05	16.88	12.17 [6.21, 18.04]	11.47	12.40 [9.28, 15.64]
Qwen	RRF	24.85	21.84	20.47	15.53 [9.83, 21.36]	15.59	18.28 [14.76, 21.96]
Phi-4	BGE-M3	29.79	23.19	17.77	16.32 [11.39, 21.59]	15.35	15.11 [11.84, 18.52]
Phi-4	BM25	15.37	16.76	11.67	8.53 [3.99, 13.23]	7.43	7.46 [4.69, 10.25]
Phi-4	TF-IDF	18.44	20.69	13.95	10.67 [5.81, 15.66]	10.91	8.31 [5.40, 11.23]
Phi-4	RRF	24.85	21.70	16.53	13.38 [8.16, 19.06]	14.59	11.84 [8.72, 14.96]

Table 15: Eight frozen HotpotQA v4 configurations. Column definitions follow Table 14.

Generator	Retriever	Margin	Linear	Count	Topology [CI]	Direct	Final gain [CI]
Qwen	BGE-M3	37.65	37.18	32.96	31.78 [24.48, 39.15]	33.77	32.86 [28.78, 36.93]
Qwen	BM25	30.91	38.75	32.51	35.02 [27.64, 42.33]	33.71	26.55 [22.38, 30.79]
Qwen	TF-IDF	30.13	34.86	30.31	29.06 [22.23, 36.36]	29.24	22.92 [19.05, 26.73]
Qwen	RRF	37.43	38.79	34.43	33.43 [25.97, 40.74]	36.13	30.74 [26.77, 34.79]
Phi-4	BGE-M3	37.65	33.52	29.53	27.94 [21.47, 34.59]	27.82	28.53 [24.44, 32.56]
Phi-4	BM25	30.91	34.57	28.32	29.45 [22.78, 36.39]	29.47	22.00 [18.12, 25.88]
Phi-4	TF-IDF	30.13	32.08	26.19	24.39 [18.05, 30.79]	21.36	20.80 [16.88, 24.78]
Phi-4	RRF	37.43	35.97	30.63	29.18 [22.82, 35.85]	27.99	26.50 [22.55, 30.46]

For each dataset and generator pair, Topology generates four shared conditions consisting of Base, N, A, and AB. It therefore requires $4\times128 = 512$ generations, while Direct requires $5\times128 = 640$. Across two datasets and two generators, calls decrease from 2,560 to 2,048, a 20.0% reduction. Relative savings increase with the number of candidate retrievers.

D.5 Prospective Transfer To 2WikiMultiHopQA

We prospectively transfer the topology-aware diagnosis to 2WikiMultiHopQA under a protocol frozen before producing retrieval rankings or generator outputs. The study uses 128 shared calibration examples and 512 disjoint test examples. It evaluates two generators with BGE-M3, BM25, TF-IDF, and RRF under fixed top-2 retrieval. Across eight configurations, EviCast makes 7/8 correct decisions with no false enables. Forecast gains achieve Pearson and Spearman correlations of

0.955 and 0.905 with independent test gains. Direct makes 6/8 correct decisions and requires 640 rather than 512 generation calls per generator.

Table 16: Prospective 2WikiMultiHopQA results. Calls are counted per generator.

Method	Agree	False enable	MAE (%)	Pearson	Spearman	Calls
Linear	**8/8**	**0**	13.03	**0.970**	**0.952**	**384**
Count	7/8	**0**	2.63	0.964	0.905	512
EviCast	**7/8**	**0**	2.92	**0.955**	**0.905**	**512**
Direct	6/8	**0**	**2.61**	0.957	0.905	640

Table 17: Per-configuration EviCast forecasts and independent 2WikiMultiHopQA test gains in percent.

Generator	Retriever	Margin	Forecast [95% CI]	Decision	Test gain [95% CI]	Test conclusion
Qwen	BGE-M3	39.11	21.79 [14.44, 29.29]	Enable	26.87 [22.89, 30.95]	Beneficial
Qwen	BM25	25.42	7.46 [1.28, 13.78]	Enable	12.27 [8.74, 15.87]	Beneficial
Qwen	TF-IDF	14.82	1.19 [−4.52, 6.97]	Keep Base	6.59 [3.81, 9.43]	Beneficial
Qwen	RRF	35.01	13.87 [7.44, 20.60]	Enable	17.37 [13.74, 21.18]	Beneficial
Phi-4	BGE-M3	39.11	22.84 [16.79, 29.25]	Enable	26.76 [23.04, 30.52]	Beneficial
Phi-4	BM25	25.42	11.33 [6.32, 16.60]	Enable	11.33 [8.39, 14.28]	Beneficial
Phi-4	TF-IDF	14.82	6.15 [1.55, 10.84]	Enable	5.84 [3.32, 8.40]	Beneficial
Phi-4	RRF	35.01	16.43 [11.29, 21.87]	Enable	16.79 [13.46, 20.23]	Beneficial

The only conservative reject is Qwen with TF-IDF, whose positive forecast has a 95% LCB below zero. EviCast obtains a 2.92% MAE, compared with 2.63% for Count and 2.61% for Direct. This prospective transfer extends the evidence for deployment decisions and relative utility forecasting to a new dataset.

D.6 Transfer To Text-Only Generators Across Model Scales

We evaluate frozen topology-aware diagnosis on standard text-only generators at different scales. The study uses Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct on HotpotQA with BGE-M3, BM25, TF-IDF, and RRF. For each generator, EviCast shares 128 calibration examples and fixes the top-2 evidence states, anchor, separability threshold, and deployment rule. Each retriever is evaluated on 512 disjoint test queries.

Table 18: Frozen transfer to text-only generators at two model scales.

Generator	Params.	Configs.	Agree	MAE (%) ↓	Pearson	Spearman
Qwen2.5-Instruct	3B	4	**4/4**	3.86	−0.145	0.200
Qwen2.5-Instruct	7B	4	**4/4**	**3.49**	0.543	0.200
Overall	3B and 7B	8	**8/8**	**3.67**	0.498	0.524

Across both scales, EviCast makes all eight deployment decisions correctly with no false enables or rejects. Its aggregate MAE is 3.67%, compared with 6.28% for Linear and 3.79% for Count. These results extend the shared intervention and evidence-state decomposition to text-only generators at two parameter scales.

E Threshold Robustness, Cost, And Frozen Audit

E.1 Retrievability Threshold

The main experiments fix the separability threshold at $\tau = 0.05$. Holding the utility gate $L_j > 0$ fixed, we sweep τ over the ten cross-task configurations. Table 19 shows 9/10 agreement with no false enables throughout $0.05 \leq \tau \leq 0.30$, demonstrating robustness to the selected threshold.

Table 19: Robustness to the separability margin threshold.

Range of τ	Agreement	False enable	Conservative reject
$0.000 \leq \tau \leq 0.025$	8/10	1	1
$0.050 \leq \tau \leq 0.300$	**9/10**	**0**	1
$0.350 \leq \tau \leq 0.500$	8/10	**0**	2

E.2 Data Isolation And Frozen Decisions

Table 20 lists audit boundaries for the principal experiments. The cross-retriever study adds candidates without replacing the generator response, while each v4 study reconstructs the profile on a new split. Deployment decisions are frozen before final evaluation, without feeding final answers, rankings, or states back into the thresholds.

Table 20: Frozen-decision and data-isolation audit. Pass indicates that the decision precedes final testing and that neither test answers nor test retrieval rankings are used for calibration.

Experiment	Cal.	Test	Prior IDs excluded	Overlap	Audit
Initial HotpotQA diagnosis	32	256	N/A	0	Pass
HotpotQA retriever reuse	32	256	N/A	0	Pass
MuSiQue Count v3	32	256	288	0	Pass
MuSiQue Topology v4	128	512	544	0	Pass
HotpotQA Topology v4	128	512	288	0	Pass
2Wiki prospective transfer	128	512	N/A	0	Pass
HotpotQA text-only scale transfer	128	512	288	0	Pass

Artifact checks verify disjoint example IDs, aligned ordering across inference arms, complete record counts, consistency between predictions and frozen decisions, and restriction to prespecified result files. All 17 tests pass for MuSiQue v3, its cross-generator study, MuSiQue v4, HotpotQA v4, and the retrievability diagnosis.

F Additional Mechanism Evidence And Qualitative Analysis

F.1 Evidence Quality And Utility Frontier

To measure how useful evidence coverage affects end-to-end gain, we use frozen per-example results from Qwen2.5-Omni, MiniCPM-o-4.5, and Phi-4 multimodal (Phi-4-MM). We construct mixtures of Wrong and Gold evidence in 5% increments and draw 2,000 replicates per ratio without new generator calls.

Table 21: Evidence utility frontiers for three generators. Scores are percentages. Break-even is the minimum Gold proportion at which the Gold/Wrong mixture reaches Base performance.

Generator	Wrong	Random	Base	Gold	Break-even
Qwen2.5-Omni	69.11	72.43	76.20	82.10	54.5
MiniCPM-o-4.5	62.43	62.41	67.88	69.75	74.3
Phi-4-MM	15.09	32.23	36.82	77.02	35.1

Table 21 shows distinct Gold-evidence requirements across generators. MiniCPM-o-4.5 requires approximately 74.3% Gold evidence to offset Wrong evidence, while Phi-4-MM reaches Base performance at approximately 35.1%. Retriever relevance alone cannot capture this generator-specific response, which directly supports the EviCast decomposition.

F.2 Cross-Task Extension To Contextual Sarcasm Understanding

We compare Target-only, Context+Target, text case retrieval, and emotion case retrieval on 480 frozen queries from the Multimodal Sarcasm Detection Dataset (MUStARD++). Their accuracies are 53.33%, 52.08%, 50.83%, and 47.71%. Grouping by case quality reveals the mechanism. Emotion retrieval gains $+3.42\%$ $[-3.08, 9.93]$ over Context when the nearest-neighbor label is correct ($n = 292$), but loses -16.49% $[-25.00, -7.98]$ when it is incorrect ($n = 188$). It also increases agreement between model predictions and neighbor labels from 52.08% to 60.63%, showing that both evidence directions affect the generator. Text neighbors follow the same pattern, with gains of $+1.81\%$ $[-4.71, 8.33]$ for correct cases and -5.39% $[-13.24, 2.45]$ for incorrect cases. This result illustrates how retrieval changes model decisions but yields positive net utility only with sufficiently reliable evidence states.

F.3 Qualitative Cases

Table 22 contrasts a helpful EMER retrieval with a distracting retrieval from the Multimodal Deception Dataset with Personality and Emotional Characteristics (MDPE). Both insert a high-similarity neighbor, yet have opposite end-to-end effects because utility depends on label alignment rather than similarity alone.

Table 22: Qualitative examples of helpful and distracting retrieval evidence.

Task	Query ID	Gold	Retrieved evidence	Base prediction / score	RAG prediction / score	Effect
EMER	sample_00003211	Angry (C)	Acoustic sim. 0.99, label Angry	A,C,E,F / 0.40	C / 1.00	Helpful
MDPE	060-8-24	Yes	Acoustic sim. 1.00, label No	Yes / 1.00	No / 0.00	Distracting

G Adaptation Of External RAG Methods

G.1 Adaptation Scope

We adapt the methodological core of external methods to a shared evaluation interface spanning multiple evidence forms and injection mechanisms. Table 23 summarizes the preserved components and their implementation in our experiments.

Table 23: Adaptation of external retrieval augmentation mechanisms.

Mechanism	Preserved core	Adaptation in this work	Status
RAER-core (Wen et al., 2025)	Retrieve similar cases with multimodal emotion features and use them as generation evidence	Use the shared Stage0a generator and frozen feature pipeline with $k = 5$ case augmentation	Complete
RAMER-core (Fan et al., 2024)	Retrieve neighbors from observed modalities and use their features to complete missing modalities	Retain normalized neighbor weighting and six missing-modality settings under a shared model interface	Complete
kNN Label-Evidence (Ma et al., 2024)	Retrieve nearest-neighbor labels with self-supervised speech emotion representations	Inject the $k = 1$ neighbor label as textual evidence for the generator	Complete
Dense Passage RAG (Lewis et al., 2020; Karpukhin et al., 2020)	Retrieve top-k external passages and generate jointly with the question	Use top-2 full passages and a fixed prompt on HotpotQA and MuSiQue candidate pools	Complete

H Utility Forecasting Under Matched Generation Budgets

H.1 Comparison Protocol And Baseline Adaptation

We analyze matched generation budgets on the HotpotQA split using Qwen2.5-3B-Instruct, 128 calibration queries, 512 disjoint test queries, and four retrievers consisting of BGE-M3, BM25, TF-IDF, and RRF. Budgets, sampling seeds, and estimators are fixed before evaluation. For each of 50 repetitions, seed $20260914 + t$, with $t \in \{0, \ldots, 49\}$, produces a label-independent permutation of calibration IDs. All methods and budgets use prefixes of the same permutation.

For answer-generation budget C, EviCast uses $\lfloor C/4 \rfloor$ queries for Base and three topology states. Direct and comparison predictors use $\lfloor C/5 \rfloor$ queries for the shared Base and four Actual conditions. EviCast recomputes stratum-conditioned responses and state frequencies on each subset, while Direct estimates the paired mean gain. We first average absolute forecast errors across the four retrievers and then average these MAEs across repetitions. The target is each retriever's mean F1 gain on the fixed test set.

RPP adapts the 30-feature pre-generation utility predictor of Tian et al. (2026) to the ten-candidate HotpotQA pool and fixed top-2 generator evidence. Features include retrieval-score statistics, embedding geometry and coherence, query performance prediction based on Bidirectional Encoder Representations from Transformers (BERT-QPP), readability, QualT5, and teacher-forced generator context likelihood. Local Normalized Query Commitment (NQC) normalizes the top-two score deviation by the absolute mean candidate score, while coherence compares the top five and bottom five candidates. We use E5-small-v2, qt5-small, and the released mean average precision (MAP)-trained BERT-QPP checkpoint. Each retriever uses ordinary least squares with an intercept and calibration-only feature standardization. All 30 features are retained without test-based tuning or prediction clipping.

U-WIG averages the two retrieved-passage scores, while U-SMV combines their magnitude and dispersion (Tao & Wu, 2014; Dado et al., 2026). BERT-QPP supplies a supervised query-passage performance score (Arabzadeh et al., 2021), and QualT5 uses the maximum learned passage-quality score among the same passages (Chang et al., 2024). All scalar baselines use the same calibration-only linear map from their native scores to F1 gain.

Budgets count answer generations. RPP additionally uses context teacher-forcing, embedding and quality-model inference, and central processing unit (CPU) feature extraction and fitting. EviCast uses offline retrieval and annotated state mapping.

H.2 Complete Budget Results

Table 24 reports all budgets. The main paper shows the low, intermediate, and maximum settings, while the 480-generation result appears here. At budget 512, EviCast uses all 128 calibration queries. Comparison predictors use 102 queries and 510 generations. EviCast has no subsampling variability at this budget. Its advantage is largest under smaller budgets, with errors converging as the budget increases.

Table 24: Complete answer-generation budget comparison. MAE values are in percent.

Method	160	320	480	512
EviCast	**5.73**	**4.34**	3.91	3.86
Direct	6.85	4.89	3.94	3.92
RPP	15.25	5.81	4.88	4.77
U-WIG	7.04	4.89	3.95	3.94
U-SMV	6.99	4.83	3.84	3.79
BERT-QPP	6.99	4.86	3.89	3.87
QualT5	6.77	4.68	**3.70**	**3.70**

Table 25 reports variability across calibration subsamples. EviCast has lower macro MAE than Direct in 32, 34, 28, and 27 of 50 paired repetitions at budgets 160, 320, 480, and 512. Against RPP, the corresponding counts are 47, 35, 38, and 38.

Table 25: Calibration-subsampling quantiles of configuration-level MAE (%).

Budget	Method	5th percentile	Median	95th percentile
160	EviCast	2.98	5.28	9.25
	Direct	2.54	6.36	13.22
	RPP	6.86	15.55	25.71
	U-SMV	2.44	6.16	12.51
	QualT5	2.37	6.00	14.46
320	EviCast	2.89	4.40	6.04
	Direct	2.68	4.88	8.02
	RPP	3.05	5.29	9.76
	U-SMV	2.24	4.51	7.75
	QualT5	2.37	4.53	7.76
480	EviCast	3.50	3.89	4.40
	Direct	2.43	3.88	5.55
	RPP	2.59	4.50	7.77
	U-SMV	2.31	3.81	5.54
	QualT5	2.18	3.56	5.25
512	EviCast	3.86	3.86	3.86
	Direct	2.72	3.92	5.22
	RPP	2.72	4.51	7.41
	U-SMV	2.52	3.84	4.84
	QualT5	2.36	3.65	4.84