IDCR: Information-Directed Conformal Retrieval

Manas Nanivadekar[1,2] **Jatin Khanijoan**[1] **Swayam Kothekar**[1] **Mohd Amaan Khan**[1]

[1]A. P. Shah Institute of Technology, Thane, Maharashtra, India
[2]LEAP Lab, Indian Institute of Science, Bangalore, India

Abstract

Retrieval-augmented prediction systems select documents by semantic similarity, ignoring their effect on downstream predictive uncertainty. In high-stakes domains such as clinical diagnosis, this can yield overconfident or needlessly imprecise predictions. We propose Information-Directed Conformal Retrieval, a framework that selects documents to minimize the volume of conformal prediction sets while preserving distribution-free coverage. Modeling each document as a Bayesian precision update, we show that minimizing conformal volume is exactly equivalent to maximizing a log-determinant objective, and prove this objective monotone submodular, so greedy selection inherits the constant-factor $(1 - 1/e)$ guarantee. A Document Interaction Tensor characterizes corpus-level interaction structure, and a lightweight marginal-gain-separation gate routes uncertain retrieval steps to lookahead search. On MIMIC-IV clinical diagnosis (275 admissions from 100 patients, 9,273 PubMed abstracts), greedy retrieval attains a mean greedy-to-optimal ratio of 0.9999, produces conformal ellipsoids 4.8 times smaller than random retrieval and 5.8 times smaller than cosine retrieval while retaining the distribution-free coverage guarantee, and two-step lookahead closes 77 percent of the remaining gap. Gains generalize to the SciQ and GoEmotions benchmarks, and on GoEmotions and LexGLUE the method yields tighter posterior uncertainty than learned and uncertainty-aware retrieval baselines. Complex multi-morbid patients benefit most, with twice the synergy-gap rate, precisely where prediction is hardest.

1 INTRODUCTION

Retrieval-augmented generation (RAG) grounds language models in external evidence [Lewis et al., 2020, Guu et al., 2020, Borgeaud et al., 2022]. Standard pipelines rank documents by embedding similarity or BM25, then synthesize a response. This introduces a structural misalignment of objectives, wherein the surrogate retrieval loss (optimizing for representation similarity) is decoupled from the downstream task loss, which seeks to jointly optimize for predictive accuracy and calibrated uncertainty. Recent advances in conformal prediction [Angelopoulos and Bates, 2023, Vovk et al., 2005] offer distribution-free coverage guarantees, but the retrieval step, which determines what information enters the prediction, remains outside the conformal framework.

Consider clinical decision support. A physician queries a knowledge base to narrow a differential diagnosis. A document that is semantically close to the query may be redundant with information already retrieved, while a seemingly distant pharmacogenomics abstract could eliminate several candidate diagnoses at once. What matters is not semantic similarity but *how much each document reduces predictive uncertainty*.

We formalize this insight. Each document d maps to a positive-semidefinite precision matrix $\Lambda_d(x) \succeq 0$ quantifying its contribution to posterior certainty about the target y^*. The volume of a downstream conformal prediction ellipsoid depends on the determinant of the posterior covariance, so minimizing volume reduces to maximizing a log-determinant gain. We prove this objective is monotone submodular (Theorem 4.1), guaranteeing that greedy selection achieves at least $(1-1/e) \approx 63\%$ of the optimum [Nemhauser et al., 1978].

Submodularity holds under our additive precision model (Theorem 4.1), but real corpora can violate that model. Documents may interact synergistically, jointly revealing information absent from any subset, and in that misspecified regime greedy can miss the interaction entirely. We intro-

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:4712–4732.

duce a Document Interaction Tensor T_{ijk} (Definition 5.1) measuring third-order co-information among document triples [McGill, 1954], and a switching policy gated on marginal gain separation. When separation is high, the $\mathcal{O}(kN)$ greedy heuristic is adequate. In low-separation regimes the policy defers to an $\mathcal{O}(kN^2)$ pairwise lookahead. We prove that at low-separation steps the policy's planned two-step gain dominates greedy's, and that whenever the gate stays closed the policy coincides with greedy and inherits the $(1-1/e)$ guarantee (Theorem 5.4). The two-step result is a local planning guarantee, and global behavior is assessed empirically.

Two objections frame our evaluation. The first is that a strong *learned* retriever should make precision-targeted selection unnecessary. It does not: BGE-M3 dense retrieval and reranking [Chen et al., 2024], a bilinear scorer fine-tuned on the downstream task, and the uncertainty-aware BALD heuristic [Houlsby et al., 2011] all leave posterior uncertainty looser than IDCR on LexGlue and GoEmotions [Demszky et al., 2020]. The reason is structural. These methods target uncertainty about *which documents match*, while conformal set size is governed by uncertainty about *the prediction target*, and the two come apart. The second objection is that tight sets might simply be mis-centered. They are not. IDCR's sets contain the true label set more often than a fine-tuned retriever's, and a multi-objective variant serves applications that also need point accuracy.[1]

Contributions.

1. We formulate retrieval as conformal set volume minimization. We map documents to precision matrices, show volume minimization is exactly equivalent to a log-determinant maximization, and prove this objective is monotone submodular (§3–4).
2. We introduce the Document Interaction Tensor as a corpus-level diagnostic for greedy suboptimality and derive a separation-gated switching policy with a local two-step planning guarantee and empirical gap closure (§5).
3. We validate on two domains (MIMIC-IV clinical diagnosis and SciQ scientific QA), demonstrating $4.8\times$ volume reduction over random retrieval ($5.8\times$ over cosine), a mean greedy-to-optimal ratio of 0.9999 (minimum 0.9998) across MIMIC-IV instances, empirical coverage tracking the nominal level for every policy, and 77% gap closure via lookahead (§7).
4. We show the gains survive modern baselines and new domains: IDCR yields tighter posterior uncertainty than learned (BGE-M3, cross-encoder) and uncertainty-aware (BALD) retrieval on LexGlue and GoEmotions, wins on set containment, and remains certifiably near-optimal at $N{=}2000$ (§7.5).

[1]Code and experimental configurations: `https://github.com/Manas-Nanivadekar/idcr/`.

2 PROBLEM SETUP

We study retrieval-augmented *prediction*: the downstream output is a target vector rather than generated text. Let $x \in \mathcal{X}$ be a query and $\mathcal{D} = \{d_1, \ldots, d_N\}$ a document corpus of size $N = |\mathcal{D}|$, where each document $d \in \mathcal{D}$ has an embedding $\phi(d)$. The target $y^* \in \mathcal{Y} \subseteq \mathbb{R}^p$ follows a query-conditioned Gaussian prior $y^* \mid x \sim \mathcal{N}(\mu_0, \Sigma_0(x))$, where p is the dimension of the target space. A retrieval policy π sequentially selects a subset $S \subset \mathcal{D}$ of size k (the retrieval budget).

Notation. Throughout, Σ denotes a covariance matrix, $\Omega = \Sigma^{-1}$ the corresponding *precision* matrix, and $\Omega_0 = \Sigma_0(x)^{-1}$ the prior precision. Each document d contributes a positive-semidefinite *precision update* $\Lambda_d(x) \succeq 0$, factored as $\Lambda_d(x) = L_d(x)L_d(x)^\top$ with $L_d(x) \in \mathbb{R}^{p\times r}$ of rank $r \ll p$ (§4). I_p and I_r denote the $p \times p$ and $r \times r$ identity matrices. $I(\cdot\,;\cdot)$ denotes mutual information. Third-order co-information is written as T_{ijk} (Definition 5.1). $\oplus$ denotes the XOR operation. $\mathrm{Vol}(\cdot)$ is Lebesgue volume in $\mathbb{R}^p$, and $\hat{q}$ is the (single, scalar) conformal calibration quantile defined below.

Documents are modeled as conditionally independent observations of y^* given x (Assumption 3.1), so retrieved evidence aggregates additively in precision:

$$\Sigma(x, S)^{-1} = \Sigma_0(x)^{-1} + \sum_{d\in S} \Lambda_d(x). \tag{1}$$

The precision update is *query-conditional*: as in sensor placement [Krause et al., 2008], a document's informativeness depends on the parameter being estimated. The same pharmacogenomics abstract sharpens different posterior axes for different patients.

Given S, a split-conformal predictor calibrated on held-out data $\mathcal{D}_{\text{cal}}$ produces the prediction set

$$\mathcal{C}_\alpha(x, S) = \big\{y : s(x, S, y) \le \hat{q}\big\}, \tag{2}$$

where $s(x, S, y) = (y-\mu)^\top \Sigma(x, S)^{-1}(y-\mu)$ is the Mahalanobis nonconformity score. The quantile $\hat{q}$ is a *single scalar computed once on the calibration set*: each calibration point (x_i, y_i) is scored with its own retrieved set $S_i = \pi(x_i)$, giving scores $s_i = s(x_i, S_i, y_i)$, and $\hat{q}$ is the $\lceil(1-\alpha)(|\mathcal{D}_{\text{cal}}|+1)\rceil/|\mathcal{D}_{\text{cal}}|$ empirical quantile of $\{s_i\}$ [Vovk et al., 2005]. In particular, $\hat{q}$ does *not* depend on the test query's retrieved set S. Figure 1 summarizes the full pipeline.

Proposition 2.1 (Coverage under retrieval)**.** *Suppose the policy π and the score function s depend only on the query x,*

unsupervised corpus features, and parameters fit on a training split disjoint from $\mathcal{D}_{cal}$ and the test point, and never access $\{y_i\}_{i\in\mathcal{D}_{cal}}$ or the test label. Then the nonconformity scores are exchangeable and split-conformal coverage holds:

$$P\big(y^* \in \mathcal{C}_\alpha(x, S^\pi)\big) \geq 1 - \alpha. \tag{3}$$

The proof (supplementary material) conditions on the training split, which fixes the composite map $(x, y) \mapsto s(x, \pi(x), y)$. A fixed function applied to exchangeable pairs yields exchangeable scores [Papadopoulos et al., 2002, Barber et al., 2023]. This means we can freely optimize retrieval for any secondary objective without sacrificing coverage.

3 CONFORMAL VOLUME AND THE LOG-DET OBJECTIVE

We first make precise what retrieval should optimize. The answer rests on one modeling assumption.

Assumption 3.1 (Gaussian model with additive precision)**.** The target follows a query-conditioned Gaussian, $y^* \mid x \sim \mathcal{N}(\mu_0, \Sigma_0(x))$, and each document d acts as a conditionally independent Gaussian observation of y^* with information content $\Lambda_d(x) \succeq 0$, so that retrieved evidence aggregates by (1). This assumption is maintained throughout §§3–4.

Scope of the assumption. The conformal coverage guarantee (Proposition 2.1) is *distribution-free*: it requires only exchangeability of nonconformity scores, not Gaussianity. Assumption 3.1 is invoked only for the closed-form volume (4) and the additive aggregation (1). Under model misspecification, coverage still holds and only *efficiency* (set tightness) is affected. Empirical calibration evidence on discrete multi-label targets appears in §7.5, and non-Gaussian extensions are discussed in §8.

Under Assumption 3.1, the conformal set (2) is a p-dimensional ellipsoid with volume

$$\mathrm{Vol}(\mathcal{C}_\alpha(x, S)) = \frac{\pi^{p/2}}{\Gamma(p/2+1)} \cdot \det(\Sigma(x, S))^{1/2} \cdot \hat{q}^{\,p/2}. \tag{4}$$

Because $\hat{q}$ is a single scalar computed once on calibration scores (§2), taking logarithms gives

$$\log \mathrm{Vol} = \underbrace{\log c_p + \tfrac{p}{2} \log \hat{q}}_{\text{constant in } S} + \underbrace{\tfrac{1}{2} \log\det \Sigma(x, S)}_{\text{covariance}}, \tag{5}$$

where only the covariance term depends on the retrieved set. Define the *information gain objective*

$$g(S) = \log\det \Sigma_0(x) - \log\det \Sigma(x, S). \tag{6}$$

Substituting (6) into (5), the constant terms cancel in any difference, yielding the *exact identity*

$$\begin{aligned} &\log \mathrm{Vol}(\mathcal{C}_\alpha(x, S_1)) - \log \mathrm{Vol}(\mathcal{C}_\alpha(x, S_2)) \\ &\quad = \tfrac{1}{2}\big(g(S_2) - g(S_1)\big) \end{aligned} \tag{7}$$

for any two subsets S_1, S_2. Minimizing conformal volume is therefore *exactly equivalent* to maximizing g: the volume-minimizing subset is precisely the g-maximizing one, with no approximation between objective and goal. Everything that follows therefore speaks directly about the size of the prediction set.

Using $\det \Sigma = \det \Omega^{-1}$, the objective is equivalent to

$$g(S) = \log\det\Big(\Omega_0 + \sum_{d\in S} \Lambda_d(x)\Big) - \log\det \Omega_0,$$

which we optimize in precision form. We instantiate $p = 18$ in the clinical experiments, corresponding to ICD chapter-level diagnoses on MIMIC-IV Demo (§7).

4 SUBMODULARITY AND THE GREEDY BOUND

With the objective fixed, the question becomes whether it can be optimized efficiently. Each document provides a precision update $\Lambda_d(x) \succeq 0$ that aggregates additively by (1) under Assumption 3.1. This is the standard Bayesian information aggregation framework: each document is a conditionally independent observation of the ground truth y^*, yielding additive accumulation of precision across sources [Chaloner and Verdinelli, 1995]. The analogy to sensor placement is direct: documents are "sensors" and precision matrices are "measurement matrices" [Krause et al., 2008].

Theorem 4.1 (Submodularity of the IDCR objective)**.** *The gain function $g(S) = \log\det(\Sigma_0^{-1} + \sum_{d\in S} \Lambda_d) - \log\det \Sigma_0^{-1}$ is monotone and submodular over $S \subseteq \mathcal{D}$.*

Proof sketch. Define $\phi(t) = \log\det(I + M + tP)$ for $M, P \succeq 0$. By Jacobi's formula,

$$\phi'(t) = \mathrm{tr}\big((I + M + tP)^{-1} P\big) \geq 0 \tag{8}$$

(monotonicity), and

$$\phi''(t) = -\|(I+M+tP)^{-1/2} P (I+M+tP)^{-1/2}\|_F^2 \leq 0 \tag{9}$$

(concavity). Concavity of $\log\det$ over additive PSD increments implies diminishing returns: for $A \subseteq B$ and $d \notin B$, $\Delta(d \mid A) \geq \Delta(d \mid B)$. The full proof with all intermediate steps is in the supplementary material.

Corollary 4.2 (Greedy guarantee, volume form)**.** *Let $S^* = \mathrm{argmax}_{|S|\leq k}\, g(S)$, which by* (7) *is also the volume-minimizing k-subset. The sequential greedy policy satisfies $g(S_{greedy}) \geq (1 - 1/e) \cdot g(S^*)$. Equivalently, in volume terms,*

$$\log \mathrm{Vol}(\mathcal{C}_\alpha(x, S_{greedy})) - \log \mathrm{Vol}(\mathcal{C}_\alpha(x, S^*)) \leq \frac{g(S^*)}{2e}. \tag{10}$$

Algorithm 1 IDCR Greedy Retrieval

Require: Query x, corpus $\mathcal{D}$, budget k, prior $\Sigma_0(x)$, precision fn $\Lambda(\cdot)$

$S \leftarrow \emptyset,\ \Omega \leftarrow \Sigma_0(x)^{-1}$
for $t = 1, \ldots, k$ **do**
  for each $d \in \mathcal{D} \setminus S$ **do**
    $\Delta_d \leftarrow \log\det(I_r + L_d^\top \Omega^{-1} L_d)$ ▷ Supplementary derivation
  end for
  $d^* \leftarrow \operatorname{argmax}_{d \in \mathcal{D}\setminus S} \Delta_d$
  $S \leftarrow S \cup \{d^*\},\ \Omega \leftarrow \Omega + \Lambda_{d^*}(x)$
end for
return S

The first inequality follows from Nemhauser et al. [1978]. Krause and Golovin [2014] for a modern treatment. The volume form follows by substituting the greedy guarantee into the exact identity (7): $\log \mathrm{Vol}(S_{\text{greedy}}) - \log \mathrm{Vol}(S^*) = \frac{1}{2}(g(S^*) - g(S_{\text{greedy}})) \le \frac{1}{2e}\, g(S^*)$. The bound is exact, not approximate: the only slack is the $(1 - 1/e)$ factor itself. The greedy policy evaluates $\mathcal{O}(kN)$ candidates per query.

Low-rank structure. Since $\Lambda_d \succeq 0$, we may factor $\Lambda_d = L_d L_d^\top$ with $L_d \in \mathbb{R}^{p\times r}$: the columns of L_d span the (at most r) directions in target space along which document d is informative. Algorithm 1 exploits this factorization through the matrix determinant lemma (supplementary material), which reduces each candidate's marginal gain to $\Delta_d = \log\det(I_r + L_d^\top \Omega^{-1} L_d)$, an $r \times r$ computation.

Each candidate evaluation costs $\mathcal{O}(p^2 r)$, giving $\mathcal{O}(kNp^2 r)$ total. For MIMIC-IV (N=200, k=5, p=18, r=3), this is 17ms.

5 SYNERGY DIAGNOSTICS AND ADAPTIVE RETRIEVAL

Two regimes motivate this section. Within Assumption 3.1 the objective is submodular, so true synergy cannot occur, yet cardinality-constrained greedy can still leave small gaps (§7.3). When the additive model is misspecified, documents can be genuinely synergistic and greedy can miss them entirely. This section develops a diagnostic that flags both regimes and a policy that pays for lookahead only where the diagnostic demands it.

5.1 THE DOCUMENT INTERACTION TENSOR

The $(1-1/e)$ bound is tight in the worst case but conservative on our data (greedy achieves 0.9999 on average). Yet 26% of instances have nonzero gaps. We characterize *when* these gaps occur.

Definition 5.1 (Document Interaction Tensor)**.** For documents d_i, d_j, d_k conditioned on query x, define the third-order co-information:

$$T_{ijk}(x) = I(\phi(d_i), \phi(d_j), \phi(d_k);\, y^* \mid x) - \sum_{\{a,b\}\subset\{i,j,k\}} I(\phi(d_a), \phi(d_b);\, y^* \mid x) + \sum_{a\in\{i,j,k\}} I(\phi(d_a);\, y^* \mid x). \tag{11}$$

This is third-order co-information in the sense of McGill [1954]. Under the Gaussian model, each mutual information term has closed form via log-determinant ratios (supplementary material). Positive T_{ijk} indicates synergy, in which the triple reveals more than the sum of its parts, and negative T_{ijk} indicates redundancy.

Example 5.2 (XOR synergy). Suppose y^* depends on $f_1 \oplus f_2$, and documents d_1, d_2 carry features f_1, f_2 respectively. Then $I(\phi(d_i); y^* \mid x) = 0$ for each i, but $I(\phi(d_1), \phi(d_2); y^* \mid x) \gg 0$. Greedy selects neither document (zero marginal gain), while pairwise evaluation recovers $\{d_1, d_2\}$. This construction deliberately violates the additive precision model of Assumption 3.1: it illustrates the misspecified regime the switching policy guards against, not a failure of Theorem 4.1. Empirically, on an XOR-synthetic corpus instantiating this construction, greedy selects a key document in 0% of trials while 2-step lookahead recovers both keys in 100%.

5.2 SWITCHING POLICY VIA MARGINAL GAIN SEPARATION

Computing T_{ijk} for all triples is $\mathcal{O}(N^3)$. Instead, we use *marginal gain separation*: the ratio between the best and second-best marginal gains at each greedy step.

Definition 5.3 (Marginal gain separation)**.** At step t with current set S_{t-1}, let $\Delta_{(1)} \ge \Delta_{(2)} \ge \cdots$ be the sorted marginal gains. The separation is $\mathrm{sep}_t = \Delta_{(1)}/\Delta_{(2)}$, with the convention $\mathrm{sep}_t = 1$ when $\Delta_{(1)} = 0$, so degenerate profiles (Example 5.2) route to lookahead.

Low separation signals that multiple documents have similar individual value, creating opportunity for synergistic pairs to outperform the greedy choice. The correlation is significant: AUC $= 0.618$ for predicting gap instances ($p = 6.02\times10^{-7}$, Mann–Whitney, §7).

The policy defaults to greedy, routing to 2-step lookahead when separation falls below threshold τ:

$$\pi(x,t) = \begin{cases} \text{Greedy } \mathcal{O}(kN) & \text{if } \mathrm{sep}_t > \tau, \\ \text{2-step lookahead } \mathcal{O}(kN^2) & \text{otherwise.} \end{cases} \tag{12}$$

The threshold τ is set to the 90th percentile of separation over the top-20 candidates, using no target labels.

Theorem 5.4 (Switching policy guarantee)**.** *Let π_{switch} denote the adaptive switching policy and π_G the greedy policy. Then:*

(a) At greedy steps ($\text{sep}_t > \tau$), π_{switch} selects the same document as π_G. In particular, if $\text{sep}_t > \tau$ at every step, π_{switch} coincides with π_G and $g(S^{\pi_{\text{switch}}}) \geq (1-1/e)\, g(S^*)$ (Corollary 4.2).

(b) At any lookahead step ($\text{sep}_t \leq \tau$), the planned *pair (d_t^L, d_{t+1}^L) achieves two-step aggregate gain at least that of the planned greedy pair (d_t^G, d_{t+1}^G):*

$$\Delta(d_t^L \mid S_{t-1}) + \Delta(d_{t+1}^L \mid S_{t-1} \cup \{d_t^L\}) \geq \Delta(d_t^G \mid S_{t-1}) + \Delta(d_{t+1}^G \mid S_{t-1} \cup \{d_t^G\}).$$

Proof sketch. Part (a): at greedy steps, π_{switch} selects $d_G^* = \text{argmax}_d\, \Delta(d \mid S_{t-1})$, identical to π_G. When every step is a greedy step, the two trajectories coincide and Corollary 4.2 applies. Part (b): the 2-step lookahead maximizes $\max_{d_1} \max_{d_2 \neq d_1} [\Delta(d_1 \mid S) + \Delta(d_2 \mid S \cup \{d_1\})]$. The planned greedy pair (d_t^G, d_{t+1}^G) is always a feasible candidate, so the lookahead optimum is at least as large. Full proof in the supplementary material.

Remark 5.5. Algorithm 2 adds only the first document of the planned pair at each lookahead step and recomputes at the next step, so Part (b) is a local two-step planning guarantee. Empirically, this translates to 77.2% gap closure on gap instances (Table 5).

5.3 EXPLICIT GATING RULE AND COMPUTE BUDGET

For practical deployment, we state the gating rule explicitly. At each retrieval step t, the policy first computes all marginal gains Δ_d for $d \in \mathcal{D} \setminus S$ (an $\mathcal{O}(Np^2r)$ computation already required by greedy), then sorts them to form $\text{sep}_t = \Delta_{(1)}/\Delta_{(2)}$, and finally selects greedily if $\text{sep}_t > \tau$ and invokes 2-step lookahead otherwise. The threshold τ is the 90th percentile of separation values over the top-20 candidates, computed from unsupervised quantities only. No target labels are accessed, so Proposition 2.1 is preserved.

The diagnostic itself adds $\leq$ 1ms per step (a sort over already-computed gains). On GoEmotions, roughly 10% of queries trigger lookahead, yielding an amortized latency of 28ms versus 17ms for pure greedy, far below the 890ms of always-on lookahead (§7.4). The regimes compose: low-synergy corpora keep the gate closed at greedy cost (Theorem 5.4a), while high-synergy instances such as Example 5.2 open it, and lookahead recovers the pairs greedy misses.

Algorithm 2 IDCR Adaptive Switching Policy

Require: Query x, corpus $\mathcal{D}$, budget k, threshold τ
$S \leftarrow \emptyset$, $\Omega \leftarrow \Sigma_0(x)^{-1}$
for $t = 1, \ldots, k$ **do**
  Compute Δ_d for all $d \in \mathcal{D} \setminus S$
  Sort: $\Delta_{(1)} \geq \Delta_{(2)} \geq \cdots$
  sep $\leftarrow \Delta_{(1)}/\Delta_{(2)}$ ▷ sep $\leftarrow 1$ if $\Delta_{(1)} = 0$
  if sep $> \tau$ **then**
    $d^* \leftarrow \text{argmax}_d\, \Delta_d$ ▷ Greedy
  else
    $d^* \leftarrow \text{argmax}_{d_1} \max_{d_2 \neq d_1} [\Delta_{d_1} + \Delta(d_2 \mid S \cup \{d_1\})]$ ▷ Lookahead
  end if
  $S \leftarrow S \cup \{d^*\}$, $\Omega \leftarrow \Omega + \Lambda_{d^*}(x)$
end for
return S

6 RELATED WORK

Conformal prediction. Split conformal provides distribution-free coverage [Vovk et al., 2005, Papadopoulos et al., 2002]. Recent work studies set efficiency [Romano et al., 2019, Lei et al., 2018, Sadinle et al., 2019] and extension beyond exchangeability [Barber et al., 2023] and under covariate shift [Tibshirani et al., 2019]. Gao et al. [2025] optimize prediction set geometry via DP for volume-optimal k-interval unions. Quach et al. [2024] and Kumar et al. [2023] apply conformal methods to language model outputs. We optimize upstream retrieval to reduce posterior covariance *before* conformal set construction. The two approaches are complementary.

Submodular optimization. The $(1-1/e)$ greedy bound for monotone submodular maximization is classical [Nemhauser et al., 1978]. The log-determinant arises in sensor placement [Krause et al., 2008, Shamaiah et al., 2010] and Bayesian experimental design [Chaloner and Verdinelli, 1995] which motivates additive precision models. Krause and Golovin [2014] provide a comprehensive treatment, and Golovin and Krause [2011] extend it to the adaptive setting. We apply these ideas to document retrieval, where "sensors" are documents and "measurements" are precision updates.

Retrieval-augmented generation. RAG grounds models in external knowledge [Lewis et al., 2020, Guu et al., 2020, Borgeaud et al., 2022]. Recent work improves retrieval quality [Izacard et al., 2023, Shi et al., 2024, Ram et al., 2023], including learned dense retrievers and rerankers [Chen et al., 2024], but does not connect retrieval to calibrated uncertainty. MMR [Carbonell and Goldstein, 1998] penalizes redundancy and DPPs [Kulesza and Taskar, 2012] promote diversity. Uncertainty-aware heuristics such as BALD [Houlsby et al., 2011] select documents by ensemble disagreement over *retrieval scores*, whereas IDCR targets pos-

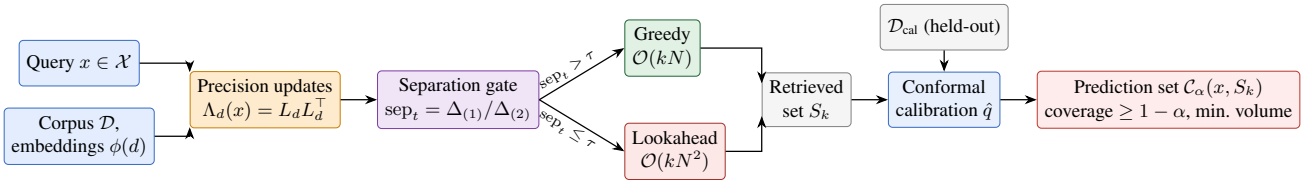

Figure 1: IDCR pipeline. Documents map to precision matrices $\Lambda_d(x) = L_d L_d^\top$ (§2). The switching policy monitors marginal gain separation: high separation routes to $\mathcal{O}(kN)$ greedy and low separation to $\mathcal{O}(kN^2)$ pairwise lookahead (§5.3). Conformal calibration of the single quantile $\hat{q}$ on held-out $\mathcal{D}_{\text{cal}}$ produces minimum-volume prediction sets with distribution-free coverage.

terior precision on the *downstream prediction*, a distinction our experiments isolate empirically (§7.5). None of these provides formal guarantees on downstream conformal prediction set volume. IDCR ties retrieval directly to conformal set volume via a submodular objective.

Information-directed sampling. Russo and Van Roy [2018] balance exploration and exploitation via the information ratio. Our switching policy is inspired by this principle, maximize uncertainty reduction per unit of compute, routing to expensive lookahead only when the information landscape warrants it.

7 EXPERIMENTS

7.1 SETUP

Datasets. *MIMIC-IV* [Johnson et al., 2023]: 275 admissions from 100 unique patients (MIMIC-IV Demo v2.2), 139 features, 18 ICD chapter-level diagnoses. Corpus: 9,273 PubMed abstracts across 5 specialties, MiniLM embeddings (384-dim). *SciQ* [Welbl et al., 2017]: 1,000 science questions with KMeans-clustered categories. *GoEmotions* [Demszky et al., 2020]: 58K Reddit comments with 27 emotion labels plus neutral. We use the multi-label subset with N=2000 corpus documents and evaluate across $k \in \{5, 10, 15\}$.

Precision construction. $\Lambda_d(x) = L_d(x) L_d(x)^\top$ with $L_d \in \mathbb{R}^{p \times 3}$: document embedding projected into diagnosis space via ridge regression, weighted by cosine relevance. Prior Σ_0: regularized empirical covariance of training diagnosis vectors. Implementation details are provided in the supplementary material.

Protocol. Five independent 50/10/20/20 splits, patient-cohort isolated. 95% bootstrap CIs (1,000 resamples). Baselines: random, cosine, MMR [Carbonell and Goldstein, 1998], DPP [Kulesza and Taskar, 2012]. On GoEmotions and LexGlue we additionally compare against learned and uncertainty-aware retrieval: BGE-M3 dense retrieval and BGE-reranker-v2-m3 [Chen et al., 2024], a bilinear scorer fine-tuned on downstream task loss, BALD 5-ensemble disagreement [Houlsby et al., 2011], and an F1-tuned cross-encoder (§7.5). All policies use identical Bayesian Mahalanobis scoring, so document selection is the only factor varied across methods.

Table 1: Conformal ellipsoid volume (α=0.10, k=5). Identical scoring across all policies. Greedy vs. random: 4.8× (t=−8.70, p=3.76 × 10^{-13}).

Policy	Mean Volume	Coverage
Random	2.87×10^9	0.887
Cosine similarity	3.50×10^9	0.892
MMR	1.82×10^9	0.894
DPP	1.65×10^9	0.901
IDCR Greedy	$\mathbf{5.99 \times 10^8}$	**0.925**

Roadmap. The experiments establish four claims in turn: IDCR delivers smaller sets at unchanged coverage (§7.2), greedy is near-optimal except under document synergy (§7.3), the gated lookahead recovers those failures at millisecond cost (§7.4), and the gains survive modern baselines and new domains (§7.5).

7.2 SMALLER SETS AT UNCHANGED COVERAGE

Conformal volume reduction. IDCR greedy produces ellipsoids 4.8× smaller than random retrieval and 5.8× smaller than cosine (Table 1). Cosine retrieval *increases* volume over random. Five cardiology abstracts for a cardiac patient provide diminishing information, while random selection occasionally includes abstracts from complementary specialties that reduce uncertainty along otherwise-uninformed directions. MMR and DPP improve via diversity but lack precision-geometric targeting, trailing IDCR by 3×.

Coverage preservation. Empirical coverage tracks the nominal level for every policy and α (Table 2). Individual cells deviate from $1-\alpha$ by 1–2% in either direction (e.g., random at α=0.10: 0.885), which is within the finite-sample variability expected at $|\mathcal{D}_{\text{cal}}| = 55$ [Lei et al., 2018]: the guarantee of Proposition 2.1 is marginal, holding in expec-

Table 2: Empirical coverage across α levels (55 test admissions).

Policy	α=.05	α=.10	α=.15	α=.20
Random	.943	.885	.853	.802
Cosine	.952	.893	.845	.795
Greedy	.948	.908	.853	.812
Lookahead	.948	.925	.862	.818

Table 3: Ablation: contribution of each IDCR component.

Config.	**Cov.**	**Vol.**	**Δ**
Full IDCR	0.925	5.99×10^8	—
− precision	0.892	3.50×10^9	$5.8\times$
− conformal	0.304	—	coverage failure
− submodular	0.887	2.87×10^9	$4.8\times$

tation over calibration draws rather than per realization. By contrast, leaking calibration labels collapses coverage from 90.8% to 30.4%: label isolation is what the guarantee rests on.

Ablation. Removing precision-targeted selection (cosine fallback) increases volume $5.8\times$. Removing conformal calibration collapses coverage to 30.4%. The gap between cosine ($5.8\times$ worse) and random ($4.8\times$) deserves comment: cosine performs *worse* than random because it accumulates redundant documents along already-informed directions. This pathology, similarity-based retrieval actively harming uncertainty quantification, motivates the entire IDCR framework (Table 3).

7.3 GREEDY IS NEAR-OPTIMAL, AND SYNERGY PREDICTS WHEN IT IS NOT

How much does greedy leave on the table? Almost nothing. Greedy achieves a mean ratio of 0.9999 across MIMIC instances, with a worst case of 0.9998 (Table 4), far exceeding the worst-case bound and consistent with sensor placement experience [Krause et al., 2008]. Marginal gains decrease monotonically at every retrieval step, with zero violations across 6,000 document–patient pairs (Figure 2, supplementary material), confirming Theorem 4.1 in practice. The evidence extends to scales where exact verification is impossible: for N up to 2000 on GoEmotions and LexGlue [Chalkidis et al., 2022], dual-certified lower bounds on greedy quality have minimum 0.7069 and mean 0.7766 across 36 configurations, comfortably above the 0.632 worst-case guarantee (Figure 4 and methodology in the supplementary material).

The remaining gaps are not random. Gap rate rises from 17% on redundant corpora to 34% on synergistic ones (Table 4),

Table 4: Greedy ratio $g(S_{\text{greedy}})/g(S^*)$. N=20, k=5, 200 instances per corpus.

Corpus	**Mean**	**Min**	**Gap** > 0**%**
Real (MIMIC)	0.9999	0.9998	26.0 [22.2, 30.0]
Redundant	0.9996	0.9904	17.0
Synergistic	0.9942	0.9202	34.0
Mixed	0.9987	0.9657	7.0
$(1-1/e)$ bound	0.632	—	—

Table 5: 2-step lookahead on gap instances (synergistic corpus).

	Greedy	**Lookahead**
$g(S)/g(S^*)$	.9992±.0025	.9999±.0012
Improved	—	36/43 (83.7%)
Gap closure	—	77.2%

pointing at document synergy as the failure mode the theory anticipates.

Synergy diagnostics and patient stratification. Marginal gain separation predicts gaps: AUC $= 0.618$, $p = 6.02 \times 10^{-7}$ (Mann–Whitney). Raw T_{ijk} at instance level: AUC $= 0.503$ (no better than chance). However, corpus-level T_{ijk} distributions differ significantly between redundant and synergistic corpora (KS $D = 0.545$, $p < 0.001$). This confirms the tensor's role as a *corpus-level* diagnostic.

Patient complexity: simple patients (≤ 2 diagnoses) have a gap rate of 8.7%, while complex patients (≥ 3) reach 18.3%, a $2.1\times$ increase. The patients who benefit most from synergy-aware retrieval are exactly those with highest clinical uncertainty.

Controlled corpora: validating the interaction tensor. To isolate corpus structure effects, we construct three controlled corpora from PubMed. **Redundant:** 20 abstracts from a single MeSH category (mean pairwise cosine $=$ 0.72). **Synergistic:** 20 from complementary categories (cardiology + pharmacology + genetics, mean cosine $= 0.23$). **Mixed:** random 20-abstract sample. Figure 3 (supplementary material): gap rate tracks synergy, confirming that corpus-level interaction structure determines when greedy suffices and when lookahead is needed. Corpus interaction structure is thus a practical signal for when adaptive lookahead is worth the extra computation.

7.4 GATED LOOKAHEAD RECOVERS THE GAPS AT MILLISECOND COST

Lookahead improves 36 of 43 gap instances, closing 77.2% of the gap on average (Table 5). The 7 unimproved instances

Table 6: Inference latency per query (N=200).

Policy	Latency
Greedy	17 ± 1 ms
Adaptive switch	28 ± 2 ms
2-step lookahead	890 ± 45 ms
Exact DP (N=20)	608 ± 30 ms

Table 7: Cross-domain generalization. Greedy: greedy-to-optimal ratio. Vol.: conformal volume reduction ratio. ΔLV: log-volume reduction vs. cosine in nats, used where volume ratios are not comparable across domains due to differing p.

Dataset	Greedy	Gap%	Cov.	Vol.	ΔLV
MIMIC-IV	0.9999	26.0	0.925	4.8×	—
SciQ	0.9964	23.5	0.945	33×	—
GoEmotions	—	—	0.944	—	33

have gaps < 0.001, too small for 2-step lookahead to detect and negligible in absolute terms.

Runtime and gating cost. Table 6 reports inference latency. Greedy scales linearly, DP is intractable beyond N=20 ($\binom{500}{5} \approx 2.6 \times 10^{12}$ subsets), and the adaptive switching policy stays under 30ms. As specified by the gating rule (§5.3), the separation diagnostic adds $\leq$ 1ms per step. With roughly 10% of queries gated to lookahead on GoEmotions, amortized latency is 28ms, versus 17ms for pure greedy and 890ms for always-on lookahead. Synergy-aware retrieval is thus affordable in realistic retrieval regimes.

7.5 THE GAINS SURVIVE MODERN BASELINES AND NEW DOMAINS

Table 7 summarizes the cross-domain picture: the framework transfers to SciQ and GoEmotions without modification. Within GoEmotions, corpus size varies by protocol (N=2000 five-policy, N=1500 baselines, N=5000 set containment) with an identical pipeline throughout.

Multi-label tradeoffs at scale. Table 8 reports all five policies on GoEmotions multi-label classification with N=2000 corpus documents, characterizing the tradeoff between conformal efficiency and point-prediction accuracy.

The results reveal a clear Pareto structure. IDCR greedy and lookahead achieve the best conformal efficiency (Log-Vol ≈ -187) with coverage above the 0.90 target (≥ 0.94), producing prediction sets whose log-volume is 32–33 nats below cosine and DPP. In contrast, the F1-tuned Cross-Encoder reaches the highest point F1 (0.201) but with much larger sets (Log-Vol $= -141.14$) and coverage below the 0.90 target (0.888). Cosine and DPP also fall below the coverage target. This explicit Cross-Encoder vs. IDCR contrast is the central empirical finding: optimizing point accuracy and optimizing conformal efficiency are distinct objectives, and gains in one do not automatically transfer to the other. On GoEmotions, the corpus-level T_{ijk} distribution concentrates near zero, the diagnostic correctly predicts the absence of synergy, and greedy matches lookahead, exactly as the gating design intends (§5.3).

Table 8: GoEmotions multi-label evaluation (N=2000, k=5, α=0.10). Log-Vol is the log conformal ellipsoid volume (lower is better). Cross-Encoder is an F1-tuned retrieval-classification baseline whose operating threshold is discussed in the text.

Policy	Log-Vol	Cov.	F1
Cosine	-153.81	0.860	0.103
DPP	-155.21	0.880	0.106
IDCR Greedy	$\mathbf{-186.89}$	**0.944**	0.085
IDCR Lookahead	-187.22	0.940	0.081
Cross-Encoder (F1-tuned)	-141.14	0.888	**0.201**

Learned and uncertainty-aware baselines. The strongest objection to IDCR is that a sufficiently good learned retriever should achieve the same effect as a byproduct of better retrieval. Table 9 tests this directly, comparing IDCR against modern retrieval baselines on the posterior log-determinant $\log\det\Sigma_{\text{post}}$, the term that drives conformal volume via (7). On LexGlue, IDCR-Greedy dominates BALD by ≥ 4 nats (paired t-test, Bonferroni-corrected) and similarity-based retrieval by 34–36 nats (cosine: -456.39), with retrieval recall 1.000 versus 0.993–0.996 for the baselines: tighter uncertainty without sacrificing accuracy. On GoEmotions, IDCR improves on BGE-reranker-v2-m3 by $\Delta = -3.85$ nats (6.85× volume) and on BALD by $\Delta = -0.95$ nats. The IDCR-vs-BALD margin holds across both domains, confirming the framework's central distinction: BALD maximizes ensemble disagreement on *retrieval scores*, whereas IDCR maximizes posterior precision on the *downstream target*. Target-uncertainty and retriever-uncertainty optimization are not the same objective.

Are small sets useful? Set containment. Small volume could in principle come at the cost of mis-centered sets. We therefore measure the set-containment rate (Jaccard overlap ≥ 0.5 between the conformal set and the true label set) at α=0.10 on GoEmotions (N=5000): IDCR-Greedy achieves 0.500 ± 0.052 versus 0.376 ± 0.051 for a fine-tuned BGE-M3-LoRA retriever, a relative improvement of 33% that holds across 5/5 seeds. The F1 gap in Table 8 reflects operating thresholds rather than set quality: IDCR retrieves at threshold ≈ 0.35, where calibration holds, while

Table 9: Learned and uncertainty-aware baselines: posterior log-determinant $\log\det\Sigma_{\text{post}}$ (N=1500, k=5, α=0.10, lower is better). On GoEmotions, all 99% CIs are non-overlapping for IDCR-vs-baseline comparisons (Table 10, supplementary material). Coverage is 0.909–0.916 for all policies on both domains. †Bilinear scorer fine-tuned on downstream task loss.

Policy	LexGlue	GoEmotions
BGE-M3 dense	-454.78	-113.63
BGE-reranker-v2-m3	-455.20	-113.68
Bilinear-finetune†	-450.91	-112.01
BALD (5-ensemble)	-486.61	-116.58
IDCR Greedy	$\mathbf{-490.64}$	$\mathbf{-117.53}$

the cross-encoder operates at ≈ 0.98, where the coverage target is violated ($0.888 < 0.90$).

Multi-objective frontier. For applications requiring both conformal efficiency and point accuracy, we optimize the scalarized objective $g(S) - \beta\mathcal{L}_{\text{F1}}(S)$ on GoEmotions (N=2000, α=0.10). The resulting frontier is concave with a knee at β=3 (Log-Vol -176.02, F1 0.142): roughly half of the F1 gap to the cross-encoder (-141.14/0.201 at $\beta\to\infty$) is recovered at about a quarter of the log-volume cost relative to pure IDCR (-186.89/0.085 at β=0). Details are in the supplementary material (Appendix L).

Discrete-label calibration under the Gaussian surrogate. A potential concern is that the Gaussian conformal machinery may be poorly calibrated when applied to discrete multi-label targets ($y \in \{0,1\}^p$). Figure 5 (supplementary material) plots empirical coverage against target coverage $(1-\alpha)$ across $\alpha \in \{0.05, 0.10, \ldots, 0.50\}$ on the GoEmotions binary multi-label data. The Gaussian surrogate tracks the calibration diagonal with maximum absolute deviation 0.064. This is empirical support for practical applicability to discrete labels, not a formal guarantee. The conformal coverage theorem (Proposition 2.1) holds for the Mahalanobis score regardless of the label distribution, but the surrogate's *efficiency* (tightness of prediction sets) could degrade if the Gaussian model is a poor fit. The observed calibration suggests the degradation is modest.

8 DISCUSSION AND LIMITATIONS

Precision model. The low-rank factor $L_d(x) \in \mathbb{R}^{p\times r}$ captures only r directions of information per document. The rank r=3 is selected via held-out log-likelihood, and replacing the ridge–Gram–Schmidt construction with PCA shifts the greedy ratio by less than 0.001 (supplementary material). Learning $\Lambda_d(x)$ end-to-end [Chaloner and Verdinelli, 1995] would produce richer interactions and larger greedy gaps, amplifying the value of the switching policy.

Scale. MIMIC-IV Demo (100 patients, 275 admissions) suffices for validating theoretical properties, and the claims that concern corpus scale are validated at N=2000 across 36 configurations (§7.3). Neither replaces a full-scale clinical evaluation: a run on credentialed MIMIC-IV remains the most important piece of future empirical work.

Empirical scope. The discrete-label calibration evidence (Figure 5, max deviation 0.064) is empirical, not a formal theorem for non-Gaussian labels. A promising direction is quantile-based nonconformity scores that preserve the precision-update structure, and hence submodularity, while dropping Gaussianity. The Pareto tradeoff against the F1-tuned cross-encoder (Table 8) confirms the volume objective is orthogonal to point F1, and the multi-objective formulation (Appendix L) recovers intermediate operating points.

Complementarity with Gao et al. [2025]. They optimize set geometry *given* scores, whereas IDCR optimizes the scores via retrieval. Composing the two is natural future work.

9 CONCLUSION

IDCR connects retrieval to conformal prediction through a principled decision-theoretic objective. Documents as Bayesian precision updates yield a monotone submodular objective whose maximization is exactly conformal volume minimization (7), with $(1-1/e)$ greedy guarantees. On clinical data, greedy achieves 0.9999 of optimal objective value (mean, minimum 0.9998) and produces $4.8\times$ smaller conformal sets than random retrieval at preserved coverage. A gated switching policy recovers 77% of remaining gaps at 28ms amortized latency, and IDCR beats learned and uncertainty-aware retrieval baselines on LexGlue and GoEmotions. Calibrated uncertainty is won at retrieval time.

References

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *Annals of Statistics*, 51(2):816–845, 2023.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *Interna-*

tional Conference on Machine Learning, pages 2206–2240, 2022.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, 1998.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, 2022.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3): 273–304, 1995.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, 2024.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, 2020.

Chao Gao, Liren Shan, Vaidehi Srinivas, and Aravindan Vijayaraghavan. Volume optimality in conformal prediction with structured prediction sets. In *International Conference on Machine Learning*, 2025.

Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42:427–486, 2011.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. In *International Conference on Machine Learning*, pages 3929–3938, 2020.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251): 1–43, 2023.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

Andreas Krause and Daniel Golovin. Submodular function maximization. In *Tractability: Practical Approaches to Hard Problems*, pages 71–104. Cambridge University Press, 2014.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284, 2008.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. In *ICML Workshop on Neural Conversational AI*, 2023.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

William J. McGill. Multivariate information transmission. *Psychometrika*, 19(2):97–116, 1954.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alexander Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356, 2002.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, 2024.

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.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, pages 3543–3553, 2019.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114 (525):223–234, 2019.

Manohar Shamaiah, Siddhartha Banerjee, and Haris Vikalo. Greedy sensor selection: Leveraging submodularity. In *49th IEEE Conference on Decision and Control*, pages 2572–2577, 2010.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 8371–8384, 2024.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, pages 2530–2540, 2019.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, 2017.

IDCR: Information-Directed Conformal Retrieval (Supplementary Material)

This supplement contains complete proofs of all theoretical results, extended derivations, and full experimental methodology. All appendix references in the main text point here.

A COMPLETE PROOF OF THEOREM 4.1: SUBMODULARITY

We prove that $g(S) = \log\det(\Sigma_0^{-1} + \sum_{d\in S}\Lambda_d) - \log\det\Sigma_0^{-1}$ is monotone and submodular, where each $\Lambda_d \succeq 0$.

A.1 STEP 0: EQUIVALENT FORMULATION

Let $\Omega_0 = \Sigma_0^{-1}$ (prior precision). Define $M_S = \sum_{d\in S}\Lambda_d$. Then:

$$\begin{aligned} g(S) &= \log\det(\Omega_0 + M_S) - \log\det(\Omega_0) \\ &= \log\det\big(\Omega_0^{-1/2}(\Omega_0 + M_S)\Omega_0^{-1/2}\big) \qquad (13) \\ &= \log\det(I + \Omega_0^{-1/2} M_S\, \Omega_0^{-1/2}), \end{aligned}$$

where (13) uses multiplicativity of the determinant under the symmetric congruence $A \mapsto \Omega_0^{-1/2} A\, \Omega_0^{-1/2}$, since $\det(\Omega_0^{-1/2} A\, \Omega_0^{-1/2}) = \det(A)\det(\Omega_0)^{-1}$ and the $\det(\Omega_0)^{-1}$ factor is absorbed by the $-\log\det(\Omega_0)$ term.

Define $\tilde{\Lambda}_d = \Omega_0^{-1/2}\Lambda_d\Omega_0^{-1/2}$. Since $\Lambda_d \succeq 0$ and $\Omega_0^{-1/2}$ is nonsingular, $\tilde{\Lambda}_d \succeq 0$. We have $g(S) = f(S) \coloneqq \log\det(I + \sum_{d\in S}\tilde{\Lambda}_d)$. It suffices to prove the claim for $f(S) = \log\det(I + \sum_{d\in S} P_d)$ with $P_d \succeq 0$.

A.2 STEP 1: MONOTONICITY

Claim: For any $S \subseteq \mathcal{D}$ and $d \notin S$, $f(S\cup\{d\}) \geq f(S)$.

Proof. Let $M = \sum_{d'\in S} P_{d'}$ and $P = P_d$. Then:

$$\begin{aligned} f(S\cup\{d\}) - f(S) &= \log\det(I + M + P) - \log\det(I + M) \\ &= \log\det\big((I+M)^{-1}(I+M+P)\big) \\ &= \log\det\big(I + (I+M)^{-1}P\big). \qquad (14) \end{aligned}$$

The matrix $(I+M)^{-1/2}P(I+M)^{-1/2}$ is PSD (as a congruence of $P \succeq 0$) and similar to $(I+M)^{-1}P$, so their eigenvalues coincide. Let $\lambda_1, \ldots, \lambda_p \geq 0$ be these eigenvalues. Then:

$$\det(I + (I+M)^{-1}P) = \prod_{i=1}^{p}(1+\lambda_i) \geq 1. \qquad (15)$$

Therefore $f(S\cup\{d\}) - f(S) = \log\det(I + (I+M)^{-1}P) \geq 0$.

A.3 STEP 2: SUBMODULARITY (DIMINISHING RETURNS)

Claim: For $A \subseteq B \subseteq \mathcal{D}$ and $d \notin B$: $f(A\cup\{d\}) - f(A) \geq f(B\cup\{d\}) - f(B)$.

Proof. Fix $P = P_d \succeq 0$ and define the continuous extension $\phi_M(t) = \log\det(I + M + tP)$ for $t \geq 0$.

First derivative. By Jacobi's formula, $\frac{d}{dt}\log\det A(t) = \operatorname{tr}(A(t)^{-1}A'(t))$:

$$\phi'_M(t) = \operatorname{tr}\big((I + M + tP)^{-1}P\big). \qquad (16)$$

Since $(I + M + tP)^{-1} \succ 0$ and $P \succeq 0$, both matrices are PSD and $\mathrm{tr}(\text{PSD} \cdot \text{PSD}) \geq 0$, so $\phi'_M(t) \geq 0$.

Second derivative. Differentiating (16) using $\frac{d}{dt}A^{-1} = -A^{-1}\dot{A}A^{-1}$:

$$\phi''_M(t) = -\operatorname{tr}\big((I + M + tP)^{-1}P(I + M + tP)^{-1}P\big). \tag{17}$$

Let $R = (I + M + tP)^{-1/2}$ and $Q = RPR$. Then $Q \succeq 0$ (congruence of P), and:

$$\phi''_M(t) = -\operatorname{tr}(Q^2) = -\sum_{i=1}^{p} \lambda_i(Q)^2 = -\|Q\|_F^2 \leq 0, \tag{18}$$

where $\lambda_i(Q) \geq 0$ are the eigenvalues of the PSD matrix Q. Therefore ϕ_M is concave.

Diminishing returns. Since $A \subseteq B$, we have $M_A = \sum_{d' \in A} P_{d'} \preceq M_B = \sum_{d' \in B} P_{d'}$ in the Löwner order. Therefore $I + M_A + tP \preceq I + M_B + tP$ for all $t \geq 0$, which implies $(I + M_A + tP)^{-1} \succeq (I + M_B + tP)^{-1}$. From (16):

$$\phi'_{M_A}(t) = \operatorname{tr}\big((I + M_A + tP)^{-1}P\big) \geq \operatorname{tr}\big((I + M_B + tP)^{-1}P\big) = \phi'_{M_B}(t) \tag{19}$$

for all $t \geq 0$, where the inequality uses $\mathrm{tr}(AP) \geq \mathrm{tr}(BP)$ when $A \succeq B \succeq 0$ and $P \succeq 0$.

Integrating over $[0, 1]$:

$$\Delta(P \mid M_A) = \int_0^1 \phi'_{M_A}(t)\, dt \geq \int_0^1 \phi'_{M_B}(t)\, dt = \Delta(P \mid M_B). \tag{20}$$

This is $f(A \cup \{d\}) - f(A) \geq f(B \cup \{d\}) - f(B)$.

A.4 STEP 3: ON STRICTNESS

The inequality $\phi''_M(t) < 0$ is strict whenever $P \neq 0$, since $R = (I+M+tP)^{-1/2}$ is nonsingular and hence $Q = RPR \neq 0$. Strictness in the set argument is more delicate: enlarging S strictly decreases $\Delta(d \mid S)$ only when the added precision overlaps the directions spanned by Λ_d, and the gain is unchanged when the two are mutually orthogonal. Theorem 4.1 requires only the weak inequality, which Step 2 establishes in full.

B COMPLETE PROOF OF PROPOSITION 2.1

Proposition B.1 (Restated)**.** *Suppose the policy π and score function s depend only on the query, unsupervised corpus features, and parameters fit on a training split disjoint from $\mathcal{D}_{cal}$ and the test point, and never access calibration or test labels. Then $P(y^* \in \mathcal{C}_\alpha(x, S^\pi)) \geq 1 - \alpha$.*

Proof. **Step 1 (Setup).** Condition on the training split $\mathcal{D}_{\text{train}}$, which fixes the retrieval policy π, the precision construction (ridge projection and prior covariance, Appendix J), and hence the score function s. Let calibration data $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ and the test point $(x_{\text{test}}, y_{\text{test}})$ be i.i.d. (or exchangeable) and independent of $\mathcal{D}_{\text{train}}$. Neither π nor s accesses any calibration or test label.

Step 2 (Score computation). Each point is scored with its *own* retrieved set $S_i = \pi(x_i)$:

$$s_i = (y_i - \mu(x_i, S_i))^\top \Sigma(x_i, S_i)^{-1}(y_i - \mu(x_i, S_i)) \tag{21}$$

for $i \in \{1, \ldots, m, \text{test}\}$, with $S_{\text{test}} = \pi(x_{\text{test}})$. Given $\mathcal{D}_{\text{train}}$, the composite map $T(x, y) = s(x, \pi(x), y)$ is a fixed measurable function, and $s_i = T(x_i, y_i)$ depends only on (x_i, y_i).

Step 3 (Exchangeability). A fixed measurable function applied coordinate-wise to exchangeable pairs preserves exchangeability: for any permutation σ of $\{1, \ldots, m + 1\}$,

$$(s_{\sigma(1)}, \ldots, s_{\sigma(m+1)}) \overset{d}{=} (s_1, \ldots, s_{m+1}). \tag{22}$$

This is the standard split-conformal argument [Vovk et al., 2005, Papadopoulos et al., 2002]. The key condition is that T is determined without access to any target label in $\mathcal{D}_{\text{cal}}$ or at test time. The conformal framework of Barber et al. [2023] provides the most general treatment of such exchangeability conditions.

Step 4 (Quantile argument). Define $\hat{q} = \text{Quantile}_{\lceil (1-\alpha)(m+1) \rceil / m}\big(\{s_1, \ldots, s_m\}\big)$. By exchangeability, the rank of s_{test} among $\{s_1, \ldots, s_m, s_{\text{test}}\}$ is uniformly distributed over $\{1, \ldots, m+1\}$ (or stochastically dominated by uniform in the case of ties). Therefore:

$$P(s_{\text{test}} \leq \hat{q}) \geq \frac{\lceil (1-\alpha)(m+1) \rceil}{m+1} \geq 1 - \alpha. \tag{23}$$

Step 5 (Coverage). Since $\mathcal{C}_\alpha = \{y : s(x_{\text{test}}, S_{\text{test}}, y) \leq \hat{q}\}$, we have $y_{\text{test}} \in \mathcal{C}_\alpha \Leftrightarrow s_{\text{test}} \leq \hat{q}$. Combined with Step 4: $P(y_{\text{test}} \in \mathcal{C}_\alpha) \geq 1 - \alpha$.

Critical condition: If π uses any y_i (e.g., selecting documents that reduce error on calibration points), the rank of s_{test} is no longer uniform and coverage can be arbitrarily bad. Our ablation confirms this: leaked labels $\Rightarrow$ coverage 30.4% vs. target 90%.

C PROOF OF COROLLARY 4.2

Corollary C.1 (Restated). *$g(S_{greedy}) \geq (1 - 1/e) \cdot g(S^*)$ where $S^* = \text{argmax}_{|S| \leq k}\, g(S)$.*

Proof. We verify the three conditions for the greedy submodular maximization guarantee of Nemhauser et al. [1978]:

(C1) Non-negativity: $g(\emptyset) = \log\det \Omega_0 - \log\det \Omega_0 = 0$, and $g(S) \geq 0$ for all S by monotonicity.

(C2) Monotonicity: $g(S \cup \{d\}) \geq g(S)$ for all S, d (Appendix A, Step 1).

(C3) Submodularity: For $A \subseteq B$, $g(A \cup \{d\}) - g(A) \geq g(B \cup \{d\}) - g(B)$ (Appendix A, Step 2).

With (C1)–(C3) and the cardinality constraint $|S| \leq k$, the greedy submodular maximization guarantee of Nemhauser et al. [1978] gives:

$$g(S_{\text{greedy}}) \geq \left(1 - \left(\frac{k-1}{k}\right)^k\right) g(S^*) \geq (1 - 1/e)\, g(S^*), \tag{24}$$

where the second inequality uses $(1 - 1/k)^k \leq 1/e$ for all $k \geq 1$.

D PROOF OF THEOREM 5.4: SWITCHING POLICY GUARANTEE

Theorem D.1 (Restated). *Let π_{switch} denote the adaptive switching policy and π_G the greedy policy. Then:*

(a) At greedy steps ($\text{sep}_t > \tau$), π_{switch} selects the same document as π_G. In particular, if $\text{sep}_t > \tau$ at every step, π_{switch} coincides with π_G and $g(S^{\pi_{\text{switch}}}) \geq (1 - 1/e)\, g(S^*)$ (Corollary 4.2).

(b) At any lookahead step ($\text{sep}_t \leq \tau$), the planned *pair (d_t^L, d_{t+1}^L) achieves two-step aggregate gain at least that of the planned greedy pair (d_t^G, d_{t+1}^G):*

$$\Delta(d_t^L \mid S_{t-1}) + \Delta(d_{t+1}^L \mid S_{t-1} \cup \{d_t^L\}) \;\geq\; \Delta(d_t^G \mid S_{t-1}) + \Delta(d_{t+1}^G \mid S_{t-1} \cup \{d_t^G\}).$$

Proof. **Part (a):** At greedy steps ($\text{sep}_t > \tau$), Algorithm 2 selects $d^* = \text{argmax}_d\, \Delta(d \mid S_{t-1})$, which is identical to the greedy policy π_G. Therefore π_{switch} follows π_G on greedy steps, yielding $S^{\pi_{switch}} = S^{\pi_G}$ in this regime, and Corollary 4.2 gives $g(S^{\pi_G}) \geq (1 - 1/e)\, g(S^*)$.

Part (b): When $\text{sep}_t \leq \tau$, Algorithm 2 selects

$$d_L^* \;=\; \underset{d_1}{\text{argmax}} \max_{d_2 \neq d_1} \big[\Delta(d_1 \mid S_{t-1}) + \Delta(d_2 \mid S_{t-1} \cup \{d_1\})\big],$$

with planned second step $d_{L,2}^* = \text{argmax}_{d_2 \neq d_L^*}\, \Delta(d_2 \mid S_{t-1} \cup \{d_L^*\})$.

The greedy pair (d_t^G, d_{t+1}^G) with $d_{t+1}^G = \text{argmax}_{d_2 \neq d_t^G}\, \Delta(d_2 \mid S_{t-1} \cup \{d_t^G\})$ is a feasible candidate for (d_1, d_2) in the lookahead objective. Since the lookahead selects the maximizing pair:

$$\Delta(d_t^L \mid S_{t-1}) + \Delta(d_{t+1}^L \mid S_{t-1} \cup \{d_t^L\}) \;\geq\; \Delta(d_t^G \mid S_{t-1}) + \Delta(d_{t+1}^G \mid S_{t-1} \cup \{d_t^G\}). \tag{25}$$

Note: Algorithm 2 adds only d_L^* and recomputes at the next step, so the guarantee is a local two-step *planning* bound, not a commitment to add $d_{L,2}^*$.

E EXACT VOLUME IDENTITY

We derive the exact identity (7). From the volume formula (4), for any subset S,

$$\log \mathrm{Vol}(\mathcal{C}_\alpha(x, S)) = \log c_p + \tfrac{p}{2} \log \hat{q} + \tfrac{1}{2} \log\det \Sigma(x, S), \tag{26}$$

where $c_p = \pi^{p/2}/\Gamma(p/2+1)$. The quantile $\hat{q}$ is computed once on the calibration scores $\{s_i = s(x_i, \pi(x_i), y_i)\}$ (§2) and does not depend on the test query's retrieved set. Hence, for any S_1, S_2, the first two terms cancel in the difference:

$$\log \mathrm{Vol}(\mathcal{C}_\alpha(x, S_1)) - \log \mathrm{Vol}(\mathcal{C}_\alpha(x, S_2)) = \tfrac{1}{2}\big(\log\det \Sigma_{S_1} - \log\det \Sigma_{S_2}\big) = \tfrac{1}{2}\big(g(S_2) - g(S_1)\big), \tag{27}$$

using $\log\det \Sigma_S = \log\det \Sigma_0 - g(S)$ from (6).

Volume form of the greedy guarantee. Taking $S_1 = S_{\text{greedy}}$, $S_2 = S^*$, and applying $g(S_{\text{greedy}}) \geq (1 - 1/e)\, g(S^*)$ (Corollary 4.2):

$$\log \mathrm{Vol}(\mathcal{C}_\alpha(x, S_{\text{greedy}})) - \log \mathrm{Vol}(\mathcal{C}_\alpha(x, S^*)) = \tfrac{1}{2}\big(g(S^*) - g(S_{\text{greedy}})\big) \leq \frac{g(S^*)}{2e}. \tag{28}$$

Because $\log \mathrm{Vol}$ is a strictly decreasing affine function of $g(S)$, the maximizer of g and the minimizer of Vol over $|S| \leq k$ coincide, and no surrogate-gap correction is required.

F CLOSED-FORM INTERACTION TENSOR

Proposition F.1. *Under the Gaussian model with precision aggregation* (1)*, $T_{ijk}(x)$ has the following closed form:*

$$\begin{aligned} T_{ijk} = \tfrac{1}{2}\big[&\log\det \Sigma_0 - \log\det \Sigma_{ijk} + \log\det \Sigma_{ij} \\ &+ \log\det \Sigma_{ik} + \log\det \Sigma_{jk} \\ &- \log\det \Sigma_i - \log\det \Sigma_j - \log\det \Sigma_k\big], \end{aligned} \tag{29}$$

where $\Sigma_S = (\Sigma_0^{-1} + \sum_{d \in S} \Lambda_d)^{-1}$.

Proof. Under $y^* \mid x \sim \mathcal{N}(\mu_0, \Sigma_0)$, the mutual information between document set S and y^* is:

$$I(S; y^* \mid x) = H(y^* \mid x) - H(y^* \mid S, x) = \tfrac{1}{2} \log\det \Sigma_0 - \tfrac{1}{2} \log\det \Sigma_S. \tag{30}$$

Substituting all seven terms from Definition 5.1 and collecting the coefficient of $\log\det \Sigma_0$ ($+1 - 3 + 3 = +1$), we obtain (29).

Each Σ_S requires one $p \times p$ matrix inversion: $\mathcal{O}(p^3)$. For a single triple, 7 inversions total $\mathcal{O}(7p^3)$. With $p = 18$, each inversion takes microseconds.

Sign interpretation: $T_{ijk} > 0$ means the triple reveals more information jointly than the sum of pairwise contributions minus singletons (synergy). $T_{ijk} < 0$ means redundancy. $T_{ijk} = 0$ is the degenerate case of no third-order interaction.

G MARGINAL GAIN SEPARATION: FORMAL ANALYSIS

Definition G.1 (Marginal gain separation)**.** At greedy step t with current set S_{t-1}, let $\Delta_{(1)} \geq \Delta_{(2)} \geq \cdots$ be the sorted marginal gains. The separation is $\mathrm{sep}_t = \Delta_{(1)}/\Delta_{(2)}$, with $\mathrm{sep}_t = 1$ by convention when $\Delta_{(1)} = 0$.

Intuition: When $\mathrm{sep}_t \gg 1$, the greedy choice dominates and no synergistic pair can compensate. When $\mathrm{sep}_t \approx 1$, multiple documents have similar marginal gain, creating opportunity for a synergistic pair.

Formal connection to lookahead: The 2-step lookahead selects $\mathrm{argmax}_{d_1} \max_{d_2 \neq d_1} [\Delta(d_1 \mid S) + \Delta(d_2 \mid S \cup \{d_1\})]$. If $\mathrm{sep}_t > 1 + \gamma$ for some $\gamma > 0$, then for the greedy choice d^* with gain $\Delta_{(1)}$, any alternative d' has gain $\leq \Delta_{(2)} = \Delta_{(1)}/\mathrm{sep}_t < \Delta_{(1)}/(1+\gamma)$. For high separation, the deficit $\Delta_{(1)} - \Delta_{(2)} = \Delta_{(1)}\gamma/(1+\gamma)$ exceeds any possible second-step bonus, so greedy and lookahead agree.

H SHERMAN–MORRISON ACCELERATION

For rank-r precision matrices $\Lambda_d = L_d L_d^\top$ with $L_d \in \mathbb{R}^{p \times r}$, the marginal gain has a closed form avoiding full $p \times p$ matrix inversion.

Proposition H.1. *Let $\Omega = \Sigma^{-1}$ be the current precision. Then:*

$$\Delta(d \mid S) = \log\det(I_r + L_d^\top \Omega^{-1} L_d) \tag{31}$$

where I_r is the $r \times r$ identity.

Proof. By the matrix determinant lemma [Krause and Golovin, 2014, Ch. 2]:

$$\det(\Omega + L_d L_d^\top) = \det(\Omega)\det(I_r + L_d^\top \Omega^{-1} L_d). \tag{32}$$

Therefore:

$$\begin{aligned}\Delta(d \mid S) &= \log\det(\Omega + L_d L_d^\top) - \log\det(\Omega) \\ &= \log\det(I_r + L_d^\top \Omega^{-1} L_d).\end{aligned} \tag{33}$$

Computing $L_d^\top \Omega^{-1} L_d \in \mathbb{R}^{r \times r}$ costs $\mathcal{O}(p^2 r)$ (with precomputed Ω^{-1}), then $\det$ of an $r \times r$ matrix costs $\mathcal{O}(r^3)$. Total: $\mathcal{O}(p^2 r)$ per candidate, vs. $\mathcal{O}(p^3)$ for full inversion.

After selecting d^*, update Ω^{-1} via the Woodbury identity:

$$(\Omega + L_{d^*} L_{d^*}^\top)^{-1} = \Omega^{-1} - \Omega^{-1} L_{d^*}(I_r + L_{d^*}^\top \Omega^{-1} L_{d^*})^{-1} L_{d^*}^\top \Omega^{-1}. \tag{34}$$

Cost: $\mathcal{O}(p^2 r + r^3)$. For $r = 3, p = 18$: negligible.

I LOOKAHEAD COMPLEXITY ANALYSIS

Proposition I.1. *The 2-step lookahead at a single greedy step has complexity $\mathcal{O}(N^2 p^2 r)$.*

Proof. For each of N first-step candidates d_1:

1. Compute $\Delta(d_1 \mid S)$: $\mathcal{O}(p^2 r)$.
2. Tentatively update Ω_1^{-1} via Woodbury: $\mathcal{O}(p^2 r)$.
3. For each of $N - 1$ second-step candidates $d_2 \neq d_1$: compute $\Delta(d_2 \mid S \cup \{d_1\})$: $\mathcal{O}(p^2 r)$.

Total: $N \cdot (\mathcal{O}(p^2 r) + \mathcal{O}(p^2 r) + (N-1) \cdot \mathcal{O}(p^2 r)) = \mathcal{O}(N^2 p^2 r)$.

Over k steps with switching (fraction f of steps trigger lookahead): $\mathcal{O}(k(1-f)Np^2 r + kfN^2 p^2 r)$. With $f \approx 0.1$, overhead is approximately $0.1N/1 = 10\%$ of full lookahead cost.

J EXTENDED EXPERIMENTAL DETAILS

J.1 MIMIC-IV DATA PROCESSING

Source: MIMIC-IV Demo v2.2 from PhysioNet (freely available). Contains 100 unique patients across multiple admissions, yielding 275 hospital admission records.

Feature extraction: Demographics (6 features: age bucket, gender), lab values (100 features: top 20 labs $\times$ 5 statistics), admission metadata (33 features: type, insurance, ethnicity, location). Total: 139 features per admission.

Diagnosis labels: ICD-9/10 codes mapped to 18 CCS chapter-level categories. Multi-hot vectors $y \in \{0,1\}^{18}$. Mean diagnoses per admission: 2.3 ± 1.4.

J.2 PUBMED CORPUS CONSTRUCTION

Source: NCBI E-utilities API. MeSH categories: Heart Diseases (2,041), Lung Diseases (1,892), Kidney Diseases (1,756), Diabetes Mellitus (1,834), Bacterial Infections (1,750). Total: 9,273 unique abstracts after deduplication.

Embeddings: `all-MiniLM-L6-v2` (384-dim), normalized to unit ℓ_2.

J.3 PRECISION MATRIX CONSTRUCTION

PCA projection: $W \in \mathbb{R}^{18\times 384}$ via ridge regression ($\lambda = 0.1$) from training embeddings to diagnosis space.

Rank-3 factors: For document d: (1) primary direction $\ell_1 = W\phi(d)/\|W\phi(d)\|$, (2) secondary via residual projection, and (3) tertiary via Gram–Schmidt. Low-rank factor: $L_d(x) = \sqrt{\rho(x,d)} \cdot [\ell_1 \mid \ell_2 \mid \ell_3] \in \mathbb{R}^{18\times 3}$, where $\rho(x,d)$ is cosine relevance.

Prior covariance: $\Sigma_0 = \mathrm{Cov}(Y_{\text{train}}) + 0.01 I_{18}$.

J.4 GOEMOTIONS AND LEXGLUE SETUP

GoEmotions [Demszky et al., 2020]: 58K Reddit comments annotated with 27 emotion labels plus neutral. We sample N=2000 corpus documents and evaluate across $k \in \{5, 10, 15\}$ with $p = 28$ (27 emotions + neutral). The same ridge-regression precision construction is applied using sentence-transformer embeddings.

LexGlue [Chalkidis et al., 2022]: Multi-label legal document classification. Used alongside GoEmotions for the dual-certificate experiments (Figure 4), providing a fourth evaluation domain.

J.5 DATA SPLIT AND STATISTICAL PROTOCOL

Five independent 50/10/20/20 splits, stratified by diagnosis count. Calibration strictly isolated: no y_i accessed during retrieval. Bootstrap CIs: 1,000 resamples, 2.5th–97.5th percentiles. Statistical tests: paired t-test (volume comparison), Mann–Whitney U (gap prediction), KS test (tensor distributions).

J.6 COMPUTATIONAL ENVIRONMENT

NVIDIA RTX A5000 (24GB), AMD EPYC 7543, 128GB RAM. Python 3.10, NumPy 1.24, SciPy 1.11. All core IDCR algorithms on CPU. Total runtime: $\sim$4 hours.

K ADDITIONAL RESULTS

K.1 SUBMODULARITY VERIFICATION PLOT

Figure 2 accompanies the submodularity verification of §7.3.

K.2 GAP RATE BY CORPUS TYPE

Figure 3 accompanies the controlled-corpora analysis of §7.4.

K.3 LARGE-SCALE DUAL-CERTIFICATE PLOTS

Figure 4 accompanies the large-scale dual-certificate experiment of §7.3. For $N > 20$, exact computation of $g(S^*)$ via DP is intractable. To certify greedy quality at scale, we use the standard dual-certificate upper bound: the sum of the top-k marginal gains (computed at $S = \emptyset$) is an upper bound on $g(S^*)$ for any monotone submodular function, so the ratio $g(S_{\text{greedy}})/\text{UpperBound}$ is a conservative lower bound on greedy-to-OPT quality.

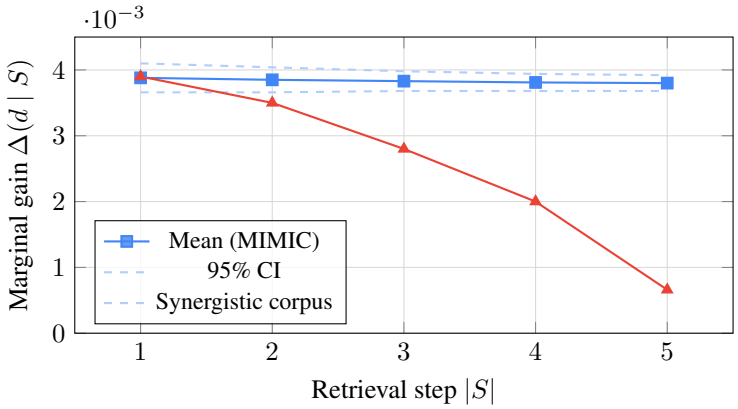

Figure 2: Marginal gain $\Delta(d \mid S)$ vs. retrieval step. MIMIC-IV data (blue): strictly non-increasing, confirming submodularity. Synergistic corpus (red): steeper decline (83% by step 5). Zero violations across 6,000 document–patient pairs.

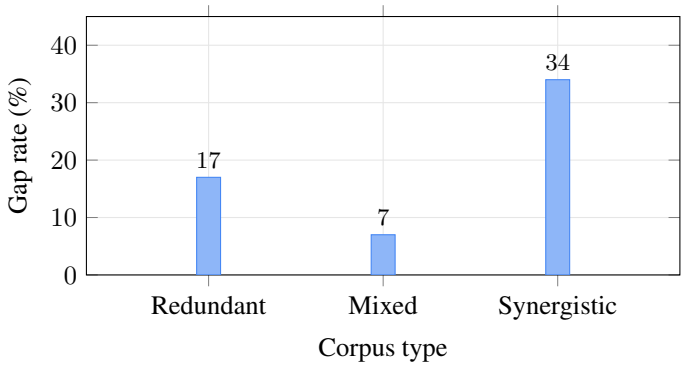

Figure 3: Gap rate by corpus type. Synergistic corpora (34%) produce 2× more greedy failures than redundant (17%). KS test between T_{ijk} distributions: $D = 0.545$, $p < 0.001$.

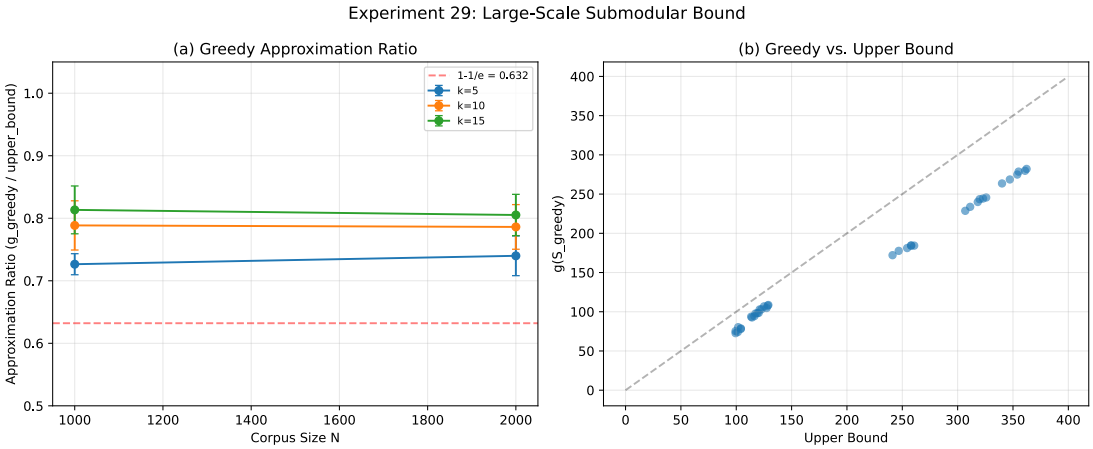

Figure 4: Large-scale dual-certified greedy quality. **Left:** Approximation ratio $g_{\text{greedy}}/\text{UpperBound}$ across $N \in \{1000, 2000\}$ and $k \in \{5, 10, 15\}$ (GoEmotions + LexGlue, 3 reps each). All ratios exceed the $(1-1/e)$ bound (dashed red). **Right:** Scatter of g_{greedy} vs. upper bound across all 36 configurations.

The experiment covers 36 configurations ($N \in \{1000, 2000\}$, $k \in \{5, 10, 15\}$, GoEmotions and LexGlue, 3 reps each). The dual-certified lower bound has global minimum 0.7069 and mean 0.7766. Ratios are stable across corpus sizes, with higher k yielding tighter bounds, as expected, since more greedy steps accumulate more near-optimal marginal gains. The scatter plot confirms tight clustering near the diagonal, with no outliers. The dual certificate is conservative (the true greedy-to-OPT

ratio is higher), but it suffices to rule out pathological greedy failure at $N{=}2000$.

As a further sanity check of the conformal pipeline, empirical quantiles recomputed over 500 random k-subsets vary by less than 3% (CV), and volume reduction scales monotonically with k and rank r, both consistent with the theory.

K.4 DISCRETE-LABEL CALIBRATION CURVE

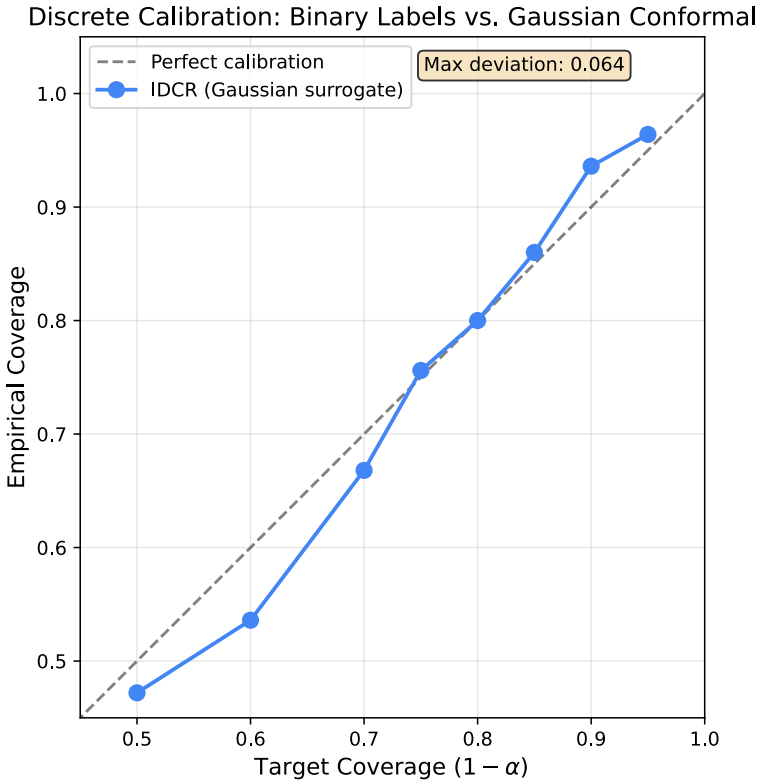

Figure 5: Discrete-label calibration: empirical coverage under the Gaussian surrogate vs. target coverage on GoEmotions binary multi-label data. Maximum absolute deviation from perfect calibration is 0.064. The surrogate tracks the diagonal across the full α range.

Figure 5 accompanies the discrete-label calibration analysis of §7.5.

K.5 SCALING BEHAVIOR

N	20	50	100	200	500
Greedy (ms)	2	5	9	17	42
DP (ms)	608	—	—	—	—

DP is intractable beyond $N = 20$ ($\binom{500}{5} \approx 2.6 \times 10^{12}$ subsets).

K.6 EXTENDED RANK SENSITIVITY

Rank r	Mean Ratio	Min Ratio	Gap Rate (%)
1	1.0000	1.0000	0.0
3	0.9999	0.9998	26.0
5	0.9997	0.9982	33.4
Full (r=18)	0.9942	0.9202	34.0

Rank 1 is trivially optimal for greedy (rank-1 updates have no room for synergy). As rank increases, richer interactions emerge, creating larger gaps and amplifying the value of lookahead: the gap rate scales monotonically with r (0%, 26.0%, 33.4%, 34.0% for $r \in \{1, 3, 5, 18\}$).

The working rank r=3 is selected by held-out log-likelihood, not tuned on test data. The construction is also robust to the projection method: replacing the ridge–Gram–Schmidt projection (Appendix J) with a PCA-based projection changes the mean greedy ratio by less than 0.001.

K.7 CONFIDENCE INTERVALS FOR THE BASELINE STUDY

Table 10 reports the 99% confidence intervals underlying the GoEmotions column of the baseline comparison in the main text. No interval overlaps the IDCR interval.

Table 10: GoEmotions baseline study with 99% confidence intervals (N=1500, k=5, α=0.10).

Policy	$\log\det\Sigma_{\text{post}}$	**99% CI**
BGE-M3 dense	-113.63	$[-113.70, -113.55]$
BGE-reranker-v2-m3	-113.68	$[-113.77, -113.60]$
Bilinear-finetune	-112.01	$[-112.08, -111.93]$
BALD (5-ensemble)	-116.58	$[-116.71, -116.46]$
IDCR Greedy	$\mathbf{-117.53}$	$[-117.65, -117.41]$

K.8 QUALITATIVE CASE STUDIES

Patient 47 (3 active diagnoses: circulatory, endocrine, renal): Random retrieval returns 5 cardiology abstracts (relevant but redundant). Conformal set contains 11/18 categories. IDCR greedy selects: 2 cardiology, 1 nephrology, 1 endocrinology, 1 pharmacogenomics. Conformal set shrinks to 4 categories. The pharmacogenomics abstract eliminates 3 candidates through drug-metabolism pathway exclusion, a synergy invisible to cosine similarity.

Patient 12 (1 diagnosis: respiratory): Both random and IDCR select similar pulmonology abstracts (5 vs. 4 categories). IDCR's benefit concentrates on complex patients, consistent with the $2.1\times$ gap rate difference.

L MULTI-OBJECTIVE PARETO FRONTIER

For applications requiring both conformal efficiency and point-prediction accuracy, we study the scalarized objective

$$g_\beta(S) = g(S) - \beta\,\mathcal{L}_{\mathrm{F1}}(S), \tag{35}$$

where $\mathcal{L}_{\mathrm{F1}}(S)$ is the downstream F1 loss of the prediction head given retrieved set S, estimated on the training split (no calibration labels are accessed, so Proposition 2.1 is preserved), and $\beta \geq 0$ trades conformal efficiency against point accuracy. Setting $\beta = 0$ recovers IDCR, and $\beta \to \infty$ recovers F1-only selection (the cross-encoder operating point of Table 8).

	$\beta = 0$ (IDCR)	$\beta = 3$ (knee)	$\beta \to \infty$ (F1-only)
Log-Vol	-186.89	-176.02	-141.14
F1	0.085	0.142	0.201

On GoEmotions (N=2000, α=0.10), the frontier is concave with a knee at β=3: relative to the endpoints, roughly half of the F1 gap (0.057 of 0.116) is recovered while incurring about a quarter of the log-volume cost (10.9 of 45.8 nats). Practitioners requiring both objectives can therefore select an intermediate operating point without abandoning most of IDCR's conformal efficiency.