Budget-Aware Routing for Long Clinical Text

Anonymous ACL submission

Abstract

A growing consideration for large language models is the token cost per query and the cost of deployment. In clinical settings, the input is long, heterogeneous, and redundant, while the downstream task is often short and high stakes. This paper studies a concrete primitive for long context use: select a subset of document units under a strict token budget so that an off-the-shelf generator can operate within fixed cost and latency. We formulate budgeted context selection as a knapsack-constrained subset selection problem with two design axes: *unitization* (how to segment documents) and *selection* (which units to keep). For selection, we present RCD, a monotone submodular objective coupling relevance, coverage, and diversity. For unitization, we evaluate sentence-level, section-based, window-based, and cluster-based strategies. We introduce a routing heuristic that chooses among methods based on budget regime. Across MIMIC discharge notes, Cochrane abstracts, and L-Eval, we find that optimal strategy depends on evaluation paradigm: for extractive evaluation, positional heuristics dominate at low budgets; for LLM-based generation, diversity-aware selection (MMR) provides consistent gains. Unitization matters less than selector choice–cluster-based grouping hurts performance, while other strategies perform similarly. We observe that ROUGE saturates for LLM-generated summaries and that BERTScore better captures quality differences between methods.

1 Introduction

Long-context capability is now a headline feature of large language models, yet clinical deployment is constrained by simple arithmetic. Each additional input token increases inference cost and latency, and clinical systems invoke models repeatedly across high-volume workflows. Discharge notes, radiology reports, and evidence syntheses are long because they are templated, redundant, and stitched from multiple sources. For example, MIMIC-IV discharge notes have a median length in the low thousands of tokens (Johnson et al., 2023), and long-document benchmarks such as L-Eval routinely exceed ten thousand tokens per instance (An et al., 2024; Bai et al., 2023).

Cost scales linearly with the token budget and with operational volume. The American Hospital Association reports that United States community hospitals account for tens of millions of admissions annually (American Hospital Association, 2025). Even a single large hospital generates on the order of ten thousand long notes per year (Rosenbloom et al., 2010). Let p_{in} and p_{out} denote the price per million input and output tokens, respectively. Processing a document with input budget B and output length S incurs

$$\text{Cost}(B, S) = \frac{p_{\text{in}} B + p_{\text{out}} S}{10^6}, \qquad (1)$$

with current prices reported by API providers (OpenAI, 2025). A reduction of even a few hundred tokens per call can translate into substantial annual savings at a hospital scale, and can reduce end-to-end latency for time-sensitive decision support.

These constraints motivate a systems perspective where we treat context construction as a budget-constrained optimization problem that sits between raw text and an off-the-shelf generator. Classical submodular maximization under knapsack constraints provides a principled mathematical foundation (Nemhauser et al., 1978; Sviridenko, 2004; Krause and Golovin, 2014), but our focus differs from the canonical setting in three ways. First, the constraint is measured in tokens, which couples directly to monetary and latency budgets. Second, the objective must behave predictably across budgets, since practitioners often operate under heterogeneous constraints across users, devices, and clinical services. Third, the system must choose among multiple selectors and representations, rather than

commit to a single surrogate objective. This motivates the routing layer studied in this paper.

1.1 Our contributions

We adopt a system-first view in which the context builder is treated as an explicit module with a measurable input budget, compute footprint, and an auditable output context. The module receives a long document D and a budget B, and emits a shorter context C_B that is passed to a downstream generator. This decoupling facilitates model-agnostic analysis of unitization and selection and permits comparisons across budget regimes.

Our contributions are as follows. (i) We formalize budgeted context construction with explicit **unitization** and **selection** stages, and we evaluate sentence-, window-, section-, and cluster-based unitization strategies. (ii) We introduce **RCD** (Relevance-Coverage-Diversity), a monotone submodular objective that combines relevance, facility-location coverage (Lin and Bilmes, 2011), and log-determinant diversity (Kulesza and Taskar, 2012) under a token knapsack constraint. (iii) We implement a unified selector suite including lead baselines (See et al., 2017), shuffled controls, sliding-window selection, hierarchical expansion, graph-based semantic clustering, maximal marginal relevance (MMR) (Carbonell and Goldstein, 1998), and RCD. (iv) We propose a lightweight **budget-aware router** that selects an algorithm as a function of budget and document statistics, and we evaluate it against oracle upper and lower bounds. (v) We report an experimental study across MIMIC discharge notes, Cochrane abstracts, and L-Eval long–document summarization tasks, including both extractive evaluation and end-to-end LLM generation.

2 Related Work

A recent wave of benchmarks evaluates model behavior as input length grows. L-Eval collects long documents with human-written questions and summarization tasks (An et al., 2024). LongBench evaluates long context understanding across multiple tasks (Bai et al., 2023). RULER and needle-in-a-haystack evaluations isolate retrieval behavior (Hsieh et al., 2024; Kamradt, 2023). Our work is orthogonal to these efforts: we change the context that reaches the model rather than the model itself. A pragmatic response to a long context is to compress prompts or select salient content. LLMLingua compresses prompts while preserving answer quality (Jiang et al., 2023). Retrieval-augmented generation selects passages from external corpora (Lewis et al., 2020). Our setting differs in that we select from a single long document under a strict token budget rather than retrieving from a large corpus. Submodular functions formalize diminishing returns and have a long history in summarization (Lin and Bilmes, 2011). Maximal marginal relevance trades relevance against redundancy using a similarity kernel (Carbonell and Goldstein, 1998). Determinantal point processes provide diversity through log-determinant objectives (Kulesza and Taskar, 2012). We build directly on these ideas, adding explicit token budgets and a routing layer that adapts to the budget regime.

3 Methodology

3.1 Framework overview

Given a long document D and a token budget B, our pipeline has four stages. (1) **Unitization**: We segment D into sentence-level candidates and estimate costs c_i. (2) **Representation**: We encode each unit with lexical or semantic features and compute relevance signals. (3) **Routing**: We map (D, B) to a selector regime, using a lightweight heuristic that trades off expected marginal utility against computational cost. (4) **Selection**: We choose a budget-feasible subset with one of several algorithms, including lead, shuffled, greedy MMR, hierarchical selection, graph-based semantic clustering, and our RCD objective. Figure 1 summarizes the end-to-end workflow and the associated evaluation loop on MIMIC, Cochrane, and L-Eval.

3.2 Problem Formulation

Let a document D be partitioned into units $u_1, \ldots, u_n$ by a unitization function, where each unit u_i incurs a token cost c_i. Given a budget B, the goal is to select a subset $S \subseteq \{1, \ldots, n\}$ such that $\sum_{i \in S} c_i \leq B$ while maximizing downstream utility. We assume access to embeddings $\phi(u_i) \in \mathbb{R}^d$ and define a similarity kernel $k(i, j) = \langle \phi(u_i), \phi(u_j) \rangle$. A selection policy maps $(u_{1:n}, B)$ to a subset S. The selected units are concatenated in document order and either evaluated directly in an extractive setting or passed to a large language model (LLM) for abstractive summarization. Algorithm 1 summarizes the complete system.

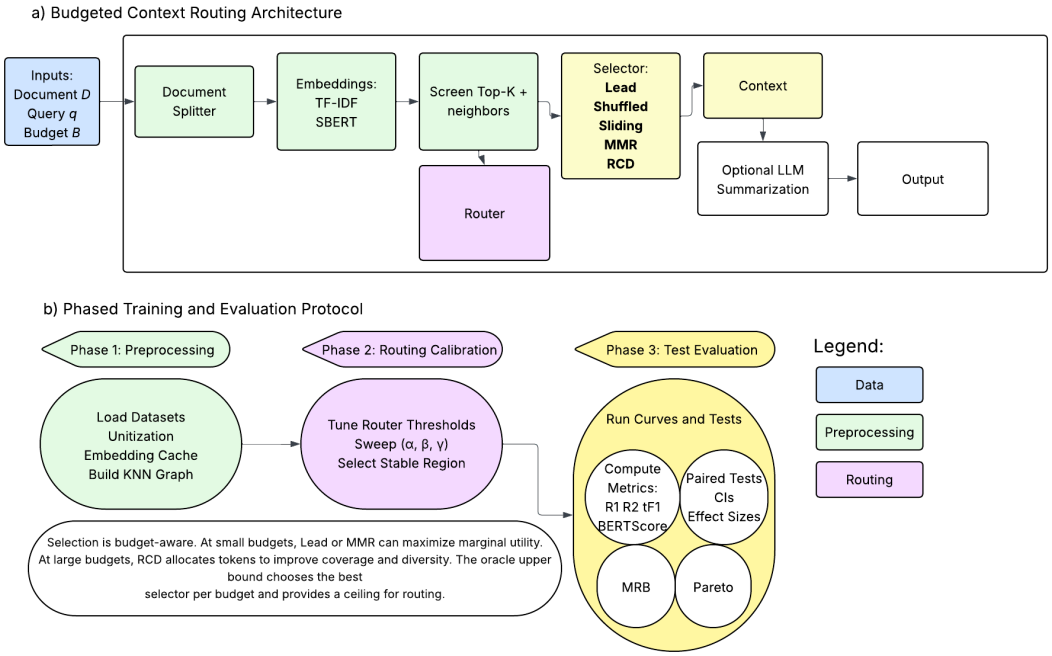


Figure 1: End-to-end architecture of the budgeted context construction framework. We begin with sentence-level unitization and feature extraction, route each (document, budget) pair to a selector regime, construct a budget-feasible context, and evaluate with ROUGE and token-level F1. The bottom phase bar reflects the experimental protocol: preprocessing, routing calibration, and held-out evaluation across MIMIC, Cochrane, and L-Eval.

Algorithm 1 Budgeted context selection

Require: Document D, budget B, unitization $\mathcal{U}$, selector m
Ensure: Selected context y
1: Build units $u_{1:n} \leftarrow \mathcal{U}(D)$
2: Compute costs c_i and embeddings $\phi(u_i)$
3: Compute relevance r_i and similarity kernel k
4: Select subset $S \leftarrow m(u_{1:n}, c_{1:n}, r_{1:n}, K, B)$
5: Concatenate context $y \leftarrow \text{concat}(\{u_i : i \in S\})$
6: **return** y

3.3 Unitization Strategies

Unitization determines the granularity and structure of selectable units. We explore four strategies that create different unit types before selection.

Sentence-Unit. Splits at sentence boundaries, treating each sentence as an independent unit. This provides maximum flexibility but may fragment coherent content.

Section-Unit. Parses section headers (e.g., `CHIEF COMPLAINT` or `HOSPITAL COURSE`) and assigns sentences to their corresponding sections enabling section-aware scoring.

Window-Unit. Creates overlapping chunks (base 50 words) with size varying by local content density.

Cluster-Unit. Groups semantically related sentences using a similarity graph with cosine similarity and proximity decay; connected components become units.

3.4 Selection Algorithms

Given units from any unitization strategy, selection determines which units to keep under the budget.

Baselines The simplest baseline is **Lead Selection**, which takes units from the beginning of the document until the budget is exhausted. It requires minimal computation beyond measuring unit lengths and serves as a strong baseline when editorial conventions prioritize important content. **Shuffled selection** randomizes unit order before applying lead selection. This strategy allows us to quantify the contribution of positional information by comparing its performance with that of standard lead selection. **Sliding selection** finds the contiguous sequence of units with the highest total relevance under the budget. It preserves local coherence and performs well when salient information is clustered in the document. **Hierarchical selection** identifies high-relevance anchor sentences and progressively adds adjacent context until the budget is filled. This strategy balances readability with relevance. **GraphCluster selection** builds a sentence similarity graph and traverses high-relevance con-

Study	Long-context eval	Budgeted context	Budget-aware routing	Multi-domain	Selector family	Objective
L-Eval (An et al., 2024)	✓	✗	✗	✓	–	–
LongBench (Bai et al., 2023)	✓	✗	✗	✓	–	–
RULER (Hsieh et al., 2024)	✓	✗	✗	✗	–	–
MMR selection (1998)	✗	✓	✗	✗	MMR	submodular variants
Facility (Lin and Bilmes, 2011)	✗	✓	✗	✗	submodular	facility location
This work	✓	✓	✓	✓	lead, MMR, hier, GSC, RCD	knapsack + submodular

Table 1: Positioning relative to prior work. Existing long-context benchmarks evaluate models at fixed or implicit context lengths, while classical selectors optimize within a budget but lack benchmark-driven, budget-aware routing across methods. We unify these threads by treating context construction as a budgeted operations research problem and evaluating the same selector family across clinical, scientific, and benchmark corpora.

nected components, selecting representative sentences from each component within the budget. This strategy encourages topical coverage by allocating budget across semantic clusters.

Maximal Marginal Relevance (MMR) MMR selects units iteratively, balancing relevance against redundancy (Carbonell and Goldstein, 1998). At each step, MMR scores each candidate unit i as

$$\mathrm{MMR}(i \mid S_t) = \lambda r_i - (1-\lambda) \max_{j \in S_t} k(i,j), \quad (2)$$

where r_i measures relevance to the query or document centroid, and $\lambda \in [0,1]$ controls the relevance–redundancy tradeoff. The redundancy term is updated at each iteration based on the currently selected set S_t. This enables adaptive diversification: once a topic is covered, additional units on the same topic are penalized.

Relevance, Coverage, Diversity (RCD) RCD defines an objective that explicitly separates three criteria. The relevance term $R(S) = \sum_{i \in S} r_i$ encourages selecting units aligned with the query. Facility location coverage $C(S) = \sum_{i=1}^{n} \max_{j \in S} k(i,j)$ encourages selecting a set that represents the entire document. Log-determinant diversity $D(S) = \log\det(I + \eta K_S)$ where K_S is the kernel submatrix indexed by S, provides a principled diversity model. The combined objective is:

$$F(S) = \alpha R(S) + \beta C(S) + \gamma D(S). \quad (3)$$

If $k(i,j) \geq 0$ and K is positive semidefinite, then F is monotone submodular. This property enables approximation guarantees under knapsack constraints via lazy greedy selection.

3.5 Routing by Budget Regime

The optimal selection strategy depends on the budget and evaluation paradigm. At small budgets, positional heuristics perform well because important content is front-loaded, and extractive metrics reward lexical overlap. At moderate budgets, redundancy becomes limiting, making diversity-aware selection advantageous. At large budgets, coverage dominates, as the challenge shifts to representing diverse topics. We formalize this intuition as a **budget-aware routing policy** that maps budget B to a selection method. Two document statistics inform the routing decision. The **front-loading index** measures the fraction of total relevance captured by units fitting in the budget:

$$\phi = \frac{\sum_{i=1}^{k} r_i}{\sum_{i=1}^{n} r_i}, \quad (4)$$

where k is the number of units fitting in B. The **redundancy index** measures local repetition:

$$\rho = \frac{1}{n-1} \sum_{i=1}^{n-1} k(i, i+1). \quad (5)$$

A simple heuristic then chooses the algorithm based on budget:

$$\pi(B) = \begin{cases} \textsc{Lead} & \text{if } B \leq B_1, \\ \text{MMR} & \text{if } B_1 < B \leq B_2, \\ \text{RCD} & \text{otherwise.} \end{cases} \quad (6)$$

Thresholds B_1 and B_2 are tuned on the validation set to maximize the evaluation metric (here, mean ROUGE-1) across the budget sweep.

4 Experimental Setup

We evaluate on three datasets spanning clinical writing, scientific abstracts, and long document benchmarks (Table 2). **MIMIC** contains discharge notes

Dataset	Input tokens		Target tokens	
	median	mean	median	mean
MIMIC	2175	2272	456	567
Cochrane	542	600	149	166
L-Eval	13069	16027	324	310

Table 2: Dataset scale and length statistics. Token counts are either provided by the dataset or estimated from word counts using an empirical tokens-per-word ratio.

paired with Brief Hospital Course summaries written by physicians (Johnson et al., 2023). **Cochrane** provides abstract-to-conclusion pairs from systematic reviews, representing clinical evidence synthesis. **L-Eval** includes summarization tasks across government reports, meetings, news, and patents (An et al., 2024).

We evaluate across token budgets $B \in \{256, 512, 1024, 2048, 4096, 8192, 16384\}$. Selection methods use sentence-level units for extractive evaluation, while unitization experiments consider Sentence-, Section-, Window-, and Cluster-Unit crossed with Lead, MMR, and our proposed RCD method.

We report ROUGE-1 and ROUGE-2 F1, which measures unigram and bigram overlap between generated text and reference summary. For end-to-end LLM evaluation, we also report BERTScore F1, which measures semantic similarity using contextualized embeddings from DeBERTa-xlarge-MNLI with baseline rescaling(Zhang et al., 2020). Scores near zero indicate chance-level similarity and scores of 0.10–0.15 indicate moderate semantic alignment typical of abstractive summaries. Full metric definitions are provided in Appendix D.

5 Results

5.1 Extractive Selection Results

Table A1 reports ROUGE-1 scores for sentence-level selection methods across datasets and budgets.

On MIMIC, Lead performs competitively at 256–512 tokens due to front-loaded content in the discharge notes. With larger budgets, diversity-aware selectors such as MMR and RCD help reduce redundancy, though ROUGE-1 gains are modest. At very large budgets, all methods converge as the entire document fits within the budget. Cochrane abstracts are short, causing selection methods to converge quickly. At 256–512 tokens, RCD typically performs best, showing that it prioritizes broader coverage over front-loading.

For L-Eval, where key content is span-specific, the Sliding selector is particularly effective for medium and large budgets. At the smallest budget, non-redundant selection methods remain competitive.

Figures 2 and 3 show ROUGE-1 and ROUGE-2 F1 across budgets. On MIMIC, Lead is strongest at low budgets due to front-loaded content, but MMR dominates at moderate budgets (1024) tokens by filtering redundancy. On Cochrane, short documents cause rapid convergence across methods. On L-Eval, Sliding selection consistently outperforms other methods by identifying concentrated spans of relevant content, while positional heuristics (Lead) provide little advantage over random selection. ROUGE-2 shows similar patterns with greater method separation.

5.2 Pareto View and Budget Efficiency

Figure 4 shows the best achievable ROUGE-1 at each budget over evaluated policies. This upper envelope defines a Pareto frontier between cost and quality, and represents the natural target for a routing policy.

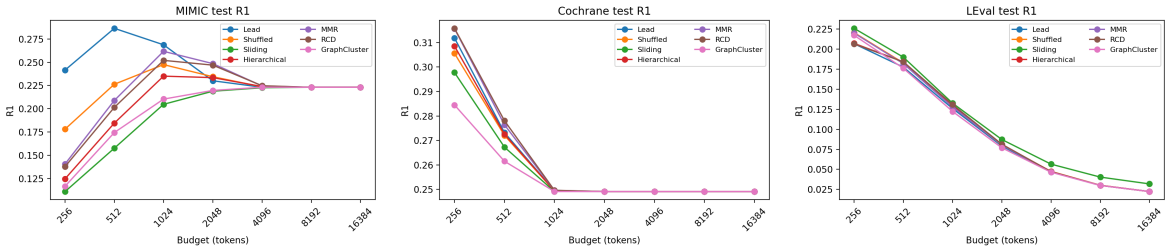
5.3 End-to-End Results with LLM Generation

Tables A2 and A3 report ROUGE-1 and BERTScore F1 on MIMIC when the selected context is passed to GPT-4o for abstractive summary generation. We evaluate all combinations of unitization strategies and selection algorithms.

MMR achieves the highest ROUGE-1 at low budgets across most unitization strategies. At 256 tokens, Hierarchical+MMR achieves 0.267, outperforming Lead by 2.1 points. While the margin is modest, it is consistent: MMR's redundancy penalty improves context even when the LLM generates abstractive summaries.

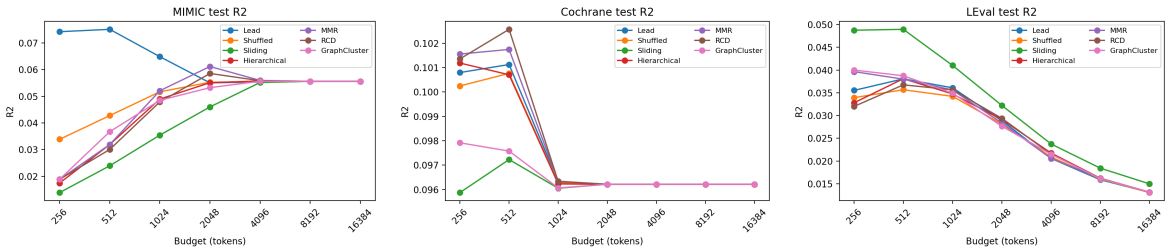
BERTScore highlights differences that ROUGE misses. While ROUGE-1 varies by only 1–3 points across methods, BERTScore shows clearer separation. MMR consistently outperforms Lead (0.111 vs 0.091 at 256 for Hierarchical), suggesting that non-redundant context helps the LLM produce more semantically appropriate summaries.

Hierarchical-Unit and Sentence-Unit produce nearly identical results across all selectors and budgets, confirming that current selectors do not exploit section metadata. Cluster-Unit underperforms



(a) MIMIC (b) Cochrane (c) L-Eval

Figure 2: ROUGE-1 F1 as a function of budget. The strongest method depends on the budget regime.



(a) MIMIC (b) Cochrane (c) L-Eval

Figure 3: ROUGE-2 F1 as a function of budget. Bigram overlap highlights redundancy effects at larger budgets.

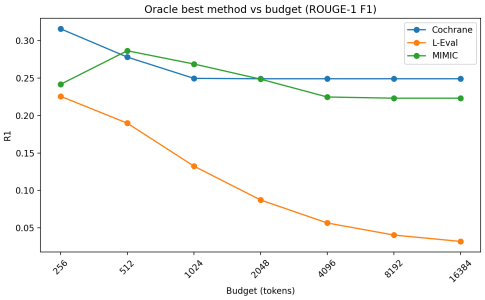


Figure 4: Pareto frontier for ROUGE-1 over budgets for each dataset, defined as the best score over evaluated policies at each budget.

	256	512	1024	2048
Lead	0.247	0.264	0.269	0.280
Shuffled	0.230	0.257	0.268	0.271
Δ	+0.017	+0.007	+0.001	+0.009

Table 3: Position dependence: Lead vs. Shuffled on MIMIC with LLM generation (N=100).

at low budgets (0.230 vs 0.247 for Lead at 256 tokens), though the gap narrows at higher budgets.

Finally, ROUGE scores plateau around 0.27–0.28 regardless of budget or method. This ceiling likely reflects the abstractive nature of LLM generation: the model paraphrases rather than copying, limiting lexical overlap with the reference.

5.4 Position Dependence

Positional signals are present but modest (Table 3). At 256 tokens, Lead outperforms Shuffled by 1.7 points (0.247 vs 0.230). This suggests that while MIMIC discharge notes front-load relevant information, the advantage is limited when an LLM processes the context. The model can identify salient information regardless of position. By 512 tokens and beyond, the performance gap is negligible.

5.5 ROUGE vs. BERTScore

ROUGE measures lexical overlap between system output and a reference summary (Lin, 2004). We report ROUGE-1 and ROUGE-2 as n-gram F1, which is appropriate when the desired summary copies domain-specific phrases and entities. However, when an LLM is used as the generator, outputs often paraphrase rather than copy, compressing the dynamic range of ROUGE across methods.

To complement ROUGE we report BERTScore F1 (Zhang et al., 2020). BERTScore aligns tokens via contextual embeddings and computes a soft precision and recall, which makes it sensitive to semantic similarity under paraphrase. We use baseline rescaling, which subtracts the expected similarity of unrelated sentence pairs, so values near zero indicate chance-level alignment and values around 0.10 indicate moderate semantic agreement for abstractive summaries.

Unitization	256	512	1024	2048
Sentence	**0.247**	**0.264**	**0.282**	0.273
Hierarchical	0.246	0.262	0.279	0.271
Sliding	0.238	0.255	0.278	**0.281**
Cluster	0.230	0.258	0.275	0.278

Table 4: Routing heuristic results on MIMIC with LLM generation. Policy: Lead for $B \leq 512$, MMR for $512 < B \leq 1024$, RCD for $B > 1024$.

On MIMIC with LLM generation, ROUGE-1 concentrates in a narrow band (roughly 0.23 to 0.28), while BERTScore separates selectors more clearly. At $B = 256$ tokens, MMR achieves the highest BERTScore across unitizations, consistent with the intuition that redundancy control improves the semantic content available to the generator even when lexical overlap changes little. Together these metrics distinguish two failure modes. A selector can achieve high ROUGE by copying frequent boilerplate. A selector can achieve high BERTScore by covering clinically salient concepts that the generator paraphrases. We therefore report both throughout.

5.6 Routing Results

Figure 5 compares the performance of the optimized routing heuristic against the oracle upper and lower bounds. The heuristic tracks the upper bound closely across all datasets, and the gap between best and worst methods is narrow (1–2 points), suggesting that budget matters more than selector choice. The degradation on MIMIC beyond 1024 tokens likely reflects signal dilution from including less relevant content.

The routing heuristic closely tracks the oracle upper bound across all three datasets, demonstrating that a simple budget-based policy can approximate per-instance method selection. Table 4 shows routing results on MIMIC with LLM generation. Using Sentence-Unit with routing, the system achieves a ROUGE-1 score of 0.282 at 1024 tokens, matching the performance of the best single-method configuration.

The heuristic assigns Lead to low budgets where positional structure is most informative, MMR to moderate budgets where redundancy filtering provides the largest gains, and RCD to high budgets where coverage becomes the limiting factor.

This routing policy provides a practical guideline: at tight budgets, prioritize non-redundant selection; at generous budgets, ensure broad coverage. While the current thresholds were tuned on extractive evaluation, the framework extends naturally to generative pipelines by recalibrating on LLM outputs.

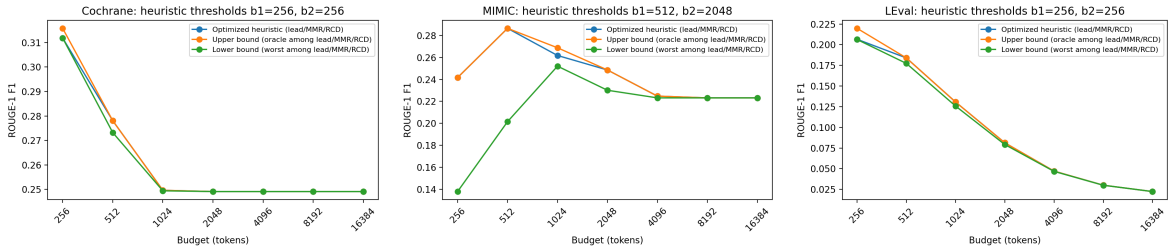
6 Discussion

The central finding of this work is that the optimal selection strategy depends on the evaluation paradigm. In the extraction evaluation (Table A1), Lead dominates at low budgets on MIMIC (0.242 vs 0.141 for MMR at 256 tokens), because discharge notes front-load relevant content and extractive metrics reward direct lexical overlap. With LLM generation (Table A2), MMR outperforms other selectors at low budgets (0.267 vs 0.246 for Lead at 256 tokens), because the LLM benefits from non-redundant input even when generating abstractive summaries. This reversal suggests different bottlenecks: extractive metrics reward lexical overlap, which positional heuristics achieve by selecting content that shares vocabulary with early-appearing reference material. LLM generation is less sensitive to exact word choice since the model can paraphrase and reorganize. In this setting, redundant input wastes context-window capacity without adding information, thereby making diversity-aware selection more valuable.

The submodular structure of RCD provides approximation guarantees under knapsack constraints, but our experiments suggest that simpler objectives often suffice. MMR, which lacks explicit coverage guarantees, performs comparably to RCD across most settings. We conjecture that coverage is important when documents contain multiple disjoint themes with low cross-similarity–a setting where MMR's local redundancy penalty cannot ensure global coverage. The sensitivity analysis (Appendix E) shows that RCD performance is robust to weight variation, but the gains over MMR remain modest.

The routing heuristic achieves near-oracle performance on extractive evaluation (Figure 5), validating the intuition that the different objectives suit different budget regimes. However, the thresholds were tuned for extractive metrics. Our LLM results suggest recalibration: the front-loading index ϕ correctly identifies when positional selection suffices for extractive tasks, but may overestimate its value for generative pipelines where the model can identify salient content regardless of position.

The 1.7-point gap between Lead and Shuffled



(a) Cochrane (b) MIMIC (c) L-Eval

Figure 5: Optimized routing heuristic performance on test, with oracle upper and lower bounds over Lead, MMR, and RCD.

at 256 tokens is smaller than expected, given editorial conventions in clinical writing. Physicians often place diagnoses and chief complaints early, yet this positional structure provides only modest benefits when our LLM generates the final summary. One interpretation is that LLMs are robust to input ordering: given relevant content anywhere in the context, the model can identify and synthesize it effectively. This robustness is valuable for deployment but complicates the design of selection policies that rely on positional heuristics.

BERTScore proves more informative than ROUGE for our task. The 2–4 point gaps in BERTScore between MMR and Lead are proportionally larger than the 1–2 point ROUGE gaps, and align better with intuitions about selection quality. For evaluating LLM-based summarization pipelines, semantic similarity metrics may be more appropriate than lexical overlap.

The failure of Cluster-Unit is instructive. We hypothesized that pre-grouping related sentences would help selectors allocate budget across topics. Instead, clustering appears to create units that are either too large for fine-grained selection or that fragment the temporal structure of clinical narratives. The admission-to-discharge arc of a hospital course may not align with semantic clusters.

Section-Unit's equivalence to Sentence-Unit reveals a gap in implementation. Parsing section structure provides metadata that could inform scoring, but our current selectors ignore this information. Exploiting section structure for scoring remains an avenue for future work.

We stress-tested the RCD objective against weight perturbations on the simplex for each dataset. Appendix Figure A1 visualizes the performance landscape at fixed budgets, and Figures A3a–A4 summarize tradeoffs and stability proxies. Two patterns are consistent across domains. First, the landscape is smooth rather than brittle. Moving weight mass from relevance toward redundancy reduction degrades ROUGE gradually, which indicates that gains are not driven by a knife-edge choice of (α, β, γ). Second, stability increases with budget. The fraction of weight settings within a small tolerance of the optimum increases as B grows, which supports the routing premise. At large budgets, multiple policies are effectively interchangeable, while at small budgets the router matters. These observations also reduce the risk of overfitting. The router is calibrated on a development split, yet the held-out curves preserve the same ordering, and the near-optimal regions in weight space overlap across datasets.

7 Conclusion

We formalized budgeted context construction as a knapsack-constrained optimization problem and introduced RCD, a monotone submodular objective combining relevance, coverage, and diversity. Across MIMIC, Cochrane, and L-Eval, we find that optimal selection strategy depends on the evaluation paradigm: Lead dominates for extractive evaluation at low budgets, while MMR provides consistent gains for LLM-based generation by filtering redundancy. Unitization matters less than selector choice, and ROUGE saturates for abstractive summaries while BERTScore better differentiates methods. A simple budget-based routing heuristic achieves near-oracle performance, providing practitioners with a practical guideline: prioritize diversity-aware selection at tight budgets when using LLMs, and ensure coverage at generous budgets. Future work includes section-aware relevance models for clinical note structure, tighter integration with learned compression operators, and robust calibration of routing policies under distribution shift.

8 Limitations

This study isolates context construction as a separate module, which clarifies the budgeted selection problem but omits end-to-end training of the generator. As a result, the reported improvements represent what can be achieved by selection alone with a fixed downstream model. Our experiments focus on summarization-style tasks with reference summaries. The conclusions may not transfer directly to settings such as long-context question answering or tool use, where the notion of relevance is conditioned on a question and where correctness can depend on a single rare fact. Similarly, our routing policy is tuned on held-out validation data within each dataset. While we include sensitivity analyses over objective weights, the router could degrade under distribution shift in note structure or writing style.

Finally, evaluation relies on automatic metrics. ROUGE and BERTScore provide complementary signals, but neither directly measures factual correctness, clinical safety, or decision impact. Human evaluation and task-based clinical studies are needed to assess whether improvements in overlap and semantic similarity translate into safer and more useful summaries.

9 Ethical Considerations

Budgeting tokens in clinical NLP is not only an economic decision. A budgeted context is a lossy view of the record, and the loss is structured by the selection policy. This raises three obligations.

Any sentence-level selector should return provenance identifiers. When an LLM consumes C_B, users should be able to trace each generated claim to the originating sentence and inspect what was excluded under the current budget. In our pipeline, provenance is available by construction.

A fixed budget can under-serve notes that are long, atypical, or clinically complex. In practice, institutions should monitor performance as a function of document length, service, and patient acuity, and escalate budgets for high risk encounters. The router in Section 3.5 is compatible with such policies because it can be constrained to select conservative policies at small budgets and to fall back to Lead when uncertainty is high.

If budgets are imposed for cost reasons, the resulting quality degradation must not be systematically borne by particular patient groups or clinical services. We recommend routine stratified reporting and periodic recalibration of routing parameters on representative hospital data.

Finally, our experiments evaluate informativeness using overlap metrics. Overlap is not a sufficient safety criterion for clinical summarization. A deployment should include factuality checks, abstention for low confidence cases, and human review for consequential decisions.

References

American Hospital Association. 2025. Fast facts on U.S. hospitals. https://www.aha.org/statistics/fast-facts-us-hospitals. Accessed 2025-12-30.

Chen An and 1 others. 2024. L-Eval: Evaluating long-context models with real-world tasks. *arXiv preprint arXiv:2403.16347*.

Yushi Bai and 1 others. 2023. LongBench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Jaime Carbonell and Jade Goldstein. 1998. 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.

Cheng Hsieh and 1 others. 2024. RULER: Scaling benchmarks for long context language models. *arXiv preprint arXiv:2404.06654*.

Dong Jiang, Chenghao Liu, Leyang Cui, and 1 others. 2023. LLMLingua: Compressing prompts for accelerated inference of large language models. *arXiv preprint arXiv:2310.05736*.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1–13.

Greg Kamradt. 2023. Needle in a haystack evaluations for long context models. https://github.com/gkamradt/LLMTest_NeedleInAHaystack. Accessed 2025-12-30.

Andreas Krause and Daniel Golovin. 2014. Submodular function maximization. *Tractability*, pages 71–104.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Deniz Kucuk, Mike Lewis, Wen-tau Yih, and Sebastian Riedel. 2020. Retrieval-augmented generation for

knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

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

OpenAI. 2025. Openai API pricing. `https://openai.com/api/pricing/`. Accessed 2025-12-30.

S. Trent Rosenbloom, William W. Stead, Joshua C. Denny, Dario Giuse, Nancy M. Lorenzi, Steven H. Brown, and Kevin B. Johnson. 2010. Generating clinical notes for electronic health record systems. *Applied Clinical Informatics*, 1(3):232–243.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Maxim Sviridenko. 2004. A note on maximizing a submodular set function subject to a knapsack constraint. In *Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1020–1021.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Extended results

	256	512	1024	2048
MIMIC				
Lead	**0.242**	**0.246**	0.249	0.230
Shuffled	0.178	0.226	0.247	0.234
Sliding	0.111	0.158	0.205	0.219
Hierarchical	0.125	0.185	0.235	0.233
MMR	0.141	0.209	**0.262**	**0.249**
RCD	0.138	0.201	0.252	0.247
GraphCluster	0.117	0.174	0.210	0.220
Cochrane				
Lead	0.312	0.273	0.249	**0.249**
Shuffled	0.306	0.272	0.249	**0.249**
Sliding	0.298	0.267	0.249	**0.249**
Hierarchical	0.309	0.273	0.249	**0.249**
MMR	**0.316**	0.276	**0.250**	**0.249**
RCD	**0.316**	**0.278**	**0.250**	**0.249**
GraphCluster	0.284	0.261	0.249	**0.249**
L-Eval				
Lead	0.207	0.177	0.126	0.079
Shuffled	0.220	0.183	0.128	0.080
Sliding	**0.226**	**0.190**	**0.132**	**0.087**
Hierarchical	0.207	0.184	0.130	0.081
MMR	0.220	0.182	0.128	0.081
RCD	0.206	0.184	0.131	0.081
GraphCluster	0.217	0.176	0.122	0.077

Table A1: Mean ROUGE-1 F1 across sentence-level selection policies.

Unitization	Selector	256	512	1024	2048
Sentence	Lead	.247	.264	.269	.280
	Shuffled	.230	.257	.268	.271
	MMR	.265	.277	**.282**	.279
	Facility	.247	.267	.274	.273
Hierarchical	Lead	.246	.262	.272	**.281**
	MMR	**.267**	**.279**	.279	.276
	Facility	.249	.267	.271	.271
Sliding	Lead	.238	.255	.263	.274
	MMR	.254	.276	.278	.277
	Facility	.237	.269	.276	**.281**
Cluster	Lead	.230	.258	.273	.277
	MMR	.258	.277	.275	.280
	Facility	.240	.271	.275	.278

Table A2: ROUGE-1 F1 on MIMIC with LLM generation.

Unitization	Selector	256	512	1024	2048
Sentence	Lead	.092	.101	.112	**.121**
	Shuffled	.074	.101	.109	.110
	MMR	**.112**	.117	**.120**	**.119**
	Facility	.097	.110	.116	.113
Hierarchical	Lead	.091	.099	.113	.118
	MMR	.111	**.118**	.119	.115
	Facility	.098	.112	.116	.115
Sliding	Lead	.086	.094	.101	.115
	MMR	.104	.114	.115	.117
	Facility	.089	.110	.112	.115
Cluster	Lead	.083	.098	.112	.116
	MMR	.104	**.118**	.115	.116
	Facility	.096	.110	.115	.117

Table A3: BERTScore F1 on MIMIC with LLM generation. Scores are rescaled with baseline correction. Values represent semantic similarity above a random baseline.

B Theoretical Properties of RCD

We recall standard definitions. A set function $f : 2^{[n]} \to \mathbb{R}$ is **submodular** if for all $A \subseteq B \subseteq [n]$ and all $e \notin B$,

$$\Delta_f(e \mid A) \geq \Delta_f(e \mid B), \tag{7}$$

where $\Delta_f(e \mid S) = f(S \cup \{e\}) - f(S)$ denotes the marginal gain. A function is **monotone** if $f(A) \leq f(B)$ whenever $A \subseteq B$.

Facility location coverage. Define $C(S) = \sum_{i=1}^{n} \max_{j \in S} k(i,j)$, with the convention that $\max_{j \in \emptyset} k(i,j) = 0$. We assume $k(i,j) \geq 0$.

Lemma 1. *C is monotone submodular.*

Proof. Fix i and define $c_i(S) = \max_{j \in S} k(i,j)$. Monotonicity is immediate since adding elements can only increase the maximum. For submodularity, let $A \subseteq B$ and $e \notin B$. Write $m_A = c_i(A)$ and $m_B = c_i(B)$. Since $A \subseteq B$, we have $m_A \leq m_B$. The marginal gains satisfy

$$\Delta_{c_i}(e \mid A) = \max\{m_A, k(i,e)\} - m_A,$$
$$\Delta_{c_i}(e \mid B) = \max\{m_B, k(i,e)\} - m_B.$$

If $k(i,e) \leq m_B$ then $\Delta_{c_i}(e \mid B) = 0$ and $\Delta_{c_i}(e \mid A) \geq 0$. If $k(i,e) > m_B$ then $\Delta_{c_i}(e \mid B) = k(i,e) - m_B$ and $\Delta_{c_i}(e \mid A) = k(i,e) - m_A \geq k(i,e) - m_B$. In both cases $\Delta_{c_i}(e \mid A) \geq \Delta_{c_i}(e \mid B)$, so c_i is submodular. Finally, C is a nonnegative sum of submodular functions, so it is submodular (Krause and Golovin, 2014). □

Log-determinant diversity. Assume $K \succeq 0$ and $\eta > 0$, and define $D(S) = \log\det(I + \eta K_S)$.

Lemma 2. *D is monotone submodular.*

Proof. Since $K \succeq 0$, there exists a feature map $\Phi \in \mathbb{R}^{n \times d}$ such that $K = \Phi\Phi^\top$. Let Φ_S denote the submatrix with rows indexed by S. By Sylvester's determinant identity,

$$\begin{aligned}\det(I + \eta K_S) &= \det(I + \eta \Phi_S \Phi_S^\top) \\ &= \det(I + \eta \Phi_S^\top \Phi_S).\end{aligned}$$

Let $M_S = I + \eta \Phi_S^\top \Phi_S \succ 0$. For $e \notin S$, the matrix determinant lemma gives

$$\det(M_{S \cup \{e\}}) = \det(M_S)\left(1 + \eta\, \phi_e^\top M_S^{-1} \phi_e\right),$$

where ϕ_e is the feature vector for item e. Taking logs yields the marginal gain

$$\Delta_D(e \mid S) = \log\Big(1 + \eta\, \phi_e^\top M_S^{-1} \phi_e\Big). \tag{8}$$

Monotonicity holds because $M_S^{-1} \succeq 0$ implies the argument of the log is at least 1. For submodularity, note that if $A \subseteq B$ then $M_A \preceq M_B$ and hence $M_A^{-1} \succeq M_B^{-1}$ in the Loewner order. Therefore $\phi_e^\top M_A^{-1} \phi_e \geq \phi_e^\top M_B^{-1} \phi_e$, and (8) implies $\Delta_D(e \mid A) \geq \Delta_D(e \mid B)$ since $\log(1 + x)$ is increasing. Thus D is submodular (Kulesza and Taskar, 2012; Krause and Golovin, 2014). □

RCD objective. Let $R(S) = \sum_{i \in S} r_i$ with $r_i \geq 0$.

Proposition 1. *If $\alpha, \beta, \gamma \geq 0$, $k(i,j) \geq 0$, and $K \succeq 0$, then*

$$F(S) = \alpha R(S) + \beta C(S) + \gamma D(S)$$

is monotone submodular.

Proof. R is modular and hence submodular. By the previous lemmas, C and D are monotone submodular. A nonnegative linear combination of submodular functions is submodular, and a nonnegative linear combination of monotone functions is monotone (Krause and Golovin, 2014). □

Optimization under a token knapsack. Maximizing a monotone submodular F under a knapsack constraint $\sum_{i \in S} c_i \leq B$ is NP-hard in general. There exist polynomial-time algorithms achieving a $(1 - 1/e)$ approximation (Sviridenko, 2004). In practice we use a lazy greedy variant that iteratively

selects the item with largest marginal gain per token cost, combined with a best-singleton check. This heuristic is widely used in large-scale summarization and often performs close to the theoretical algorithms while remaining simple to implement.

C Ensuring Positive Semidefiniteness

When $K = XX^\top$ for an embedding matrix X, positive semidefiniteness holds by construction. Otherwise, we project by symmetrization and eigenvalue clipping: compute $K \leftarrow \frac{1}{2}(K + K^\top)$, perform eigendecomposition, replace negative eigenvalues with zero, and add small diagonal jitter.

D Performance Metrics Definitions

D.1 ROUGE-1

Let y denote the generated text and y^* denote the reference summary. Then, let $U(y)$ denote the multiset of unigrams in y. Precision and recall are defined as:

$$P_1 = \frac{|U(y) \cap U(y^*)|}{|U(y)|}, \quad R_1 = \frac{|U(y) \cap U(y^*)|}{|U(y^*)|}. \tag{9}$$

and ROUGE-1 is the harmonic mean:

$$\text{ROUGE-1} = \frac{2P_1R_1}{P_1 + R_1} \tag{10}$$

D.2 ROUGE-2

Let $B(y)$ denote the multiset of consecutive word pairs in y. Precision, recall and F1 are defined analogously:

$$P_2 = \frac{|B(y) \cap B(y^*)|}{|B(y)|}, \quad R_2 = \frac{|B(y) \cap B(y^*)|}{|B(y^*)|}. \tag{11}$$

$$\text{ROUGE-2} = \frac{2P_2R_2}{P_2 + R_2}, \tag{12}$$

D.3 BERTScore

Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)$ and $\mathbf{y}^* = (\mathbf{y}^*_1, \ldots, \mathbf{y}^*_n)$ denote the contextualized token embeddings from a pretrained model (we use DeBERTa-xlarge-MNLI). Precision and recall are computed via greedy maximum cosine similarity matching:

$$P_{\text{BERT}} = \frac{1}{m}\sum_{i=1}^{m} \max_j \mathbf{y}_i^\top \mathbf{y}^*_j, \tag{13}$$

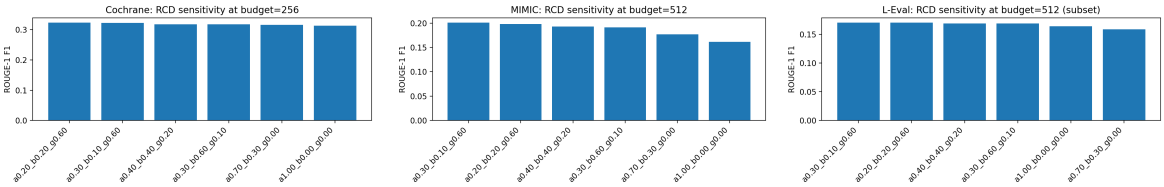
$$R_{\text{BERT}} = \frac{1}{n}\sum_{j=1}^{n} \max_i \mathbf{y}_i^\top \mathbf{y}^*_j. \tag{14}$$

BERTScore F1 is the harmonic mean:

$$\text{BERTScore} = \frac{2P_{\text{BERT}}R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}. \tag{15}$$

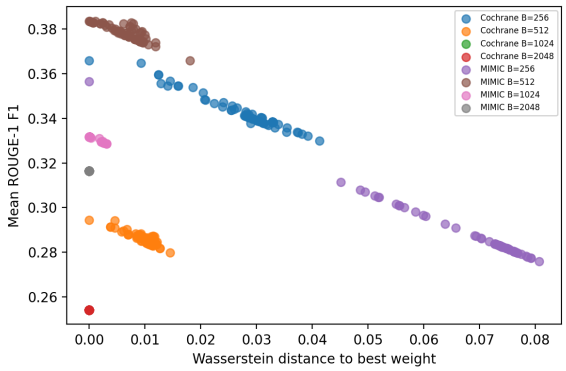
E Sensitivity Analysis

Figures A1–A4 collectively show that the RCD objective is stable across a broad range of weight configurations and budgets, rather than being tuned to a narrow parameter regime. Figure A1 demonstrates that, for a fixed budget, the mean ROUGE-1 score varies only modestly as (α, β, γ) are perturbed, with no abrupt degradation when shifting mass between relevance, coverage, and diversity. Figure A2a makes this observation quantitative by plotting mean ROUGE-1 against the Wasserstein distance from the best-performing weight vector: performance decreases smoothly and approximately monotonically as the distance increases, indicating that suboptimal weights induce gradual, not catastrophic, loss. Figure A3a, A3b further shows that the fraction of weight settings within a small tolerance of the best score grows rapidly with the budget, approaching one for larger budgets on both MIMIC and Cochrane, which implies that RCD becomes increasingly insensitive to precise weight choice as more tokens are available. Finally, the simplex visualizations in Figure A4a, A4b reveal wide plateaus of near-optimal solutions rather than isolated optima, with the empirically selected weights (highlighted in yellow) lying in the interior of these regions. Taken together, these results provide strong evidence that RCD is not overfit to a specific (α, β, γ) configuration and that its empirical gains are driven by the structure of the objective itself, not by fragile hyperparameter tuning.

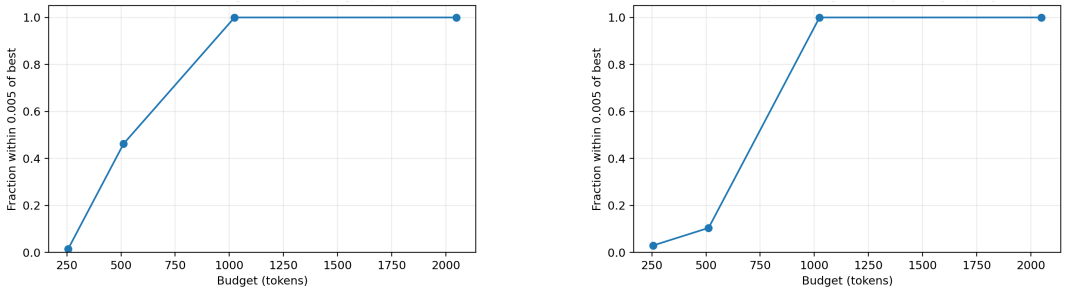


(a) Cochrane (b) MIMIC (c) L-Eval

Figure A1: Sensitivity of RCD to the objective weights α, β, γ.

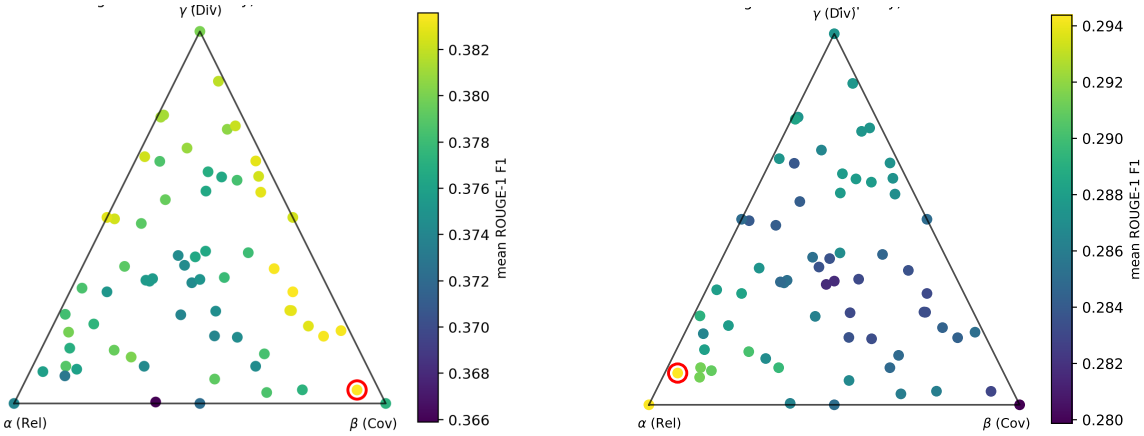


(a) Mean ROUGE-1 F1 Score vs. Wasserstein Distance from Best Weight by dataset and budget



(a) MIMIC (b) Cochrane

Figure A3: Robustness proxy plots for MIMIC and Cochrane across budgets and objectives.



(a) MIMIC (b) Cochrane

Figure A4: Weight sensitivity on MIMIC and Cochrane ($B = 512$).