KERNEL CHANGE-POINT DETECTION UNDER DEPENDENCE FOR UNSUPERVISED TEXT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised text segmentation is important because boundary labels are expensive, subjective, and often fail to transfer across domains or granularity choices. Text segmentation can be viewed as detecting distributional changes in sequences of pretrained representations, yet existing kernel change-point detection (KCPD) theory largely assumes independent observations. To provide a principled analysis beyond the usual i.i.d. setting, we study penalized KCPD under m-dependence, a stylized finite-memory model of short-range dependence, and establish an oracle inequality for the population penalized risk together with a conservative localization guarantee. Building on this perspective, we introduce EMBED-KCPD, a training-free method that represents text units with pretrained sentence embeddings and estimates boundaries using penalized KCPD. To complement the theory empirically, we introduce an LLM-based simulation framework that generates synthetic documents with known boundaries and controlled short-range dependence. Across standard text segmentation benchmarks, EMBED-KCPD is competitive with and often outperforms strong unsupervised baselines. We further show that its segments are useful beyond intrinsic boundary-based evaluation, serving as effective chunks for evidence retrieval in retrieval-augmented generation (RAG) systems.

1 INTRODUCTION

Segmenting a document into coherent topical units is a core subroutine in many NLP and IR systems. Reliable boundaries improve retrieval, summarization, question answering, discourse analysis, and downstream modeling (Prince & Labadié, 2007; Shtekh et al., 2018; Llopis et al., 2002; Cho et al., 2022). This is especially relevant in retrieval-augmented generation (RAG) pipelines, where documents must be split into chunks before indexing, and the quality of these chunks can directly affect retrieval quality. Despite this importance, text segmentation is often a poor fit for standard supervised learning. The "correct" boundary locations depend on the downstream task, the desired granularity, and the annotation protocol, which can vary substantially across corpora. Labels are therefore expensive to obtain, difficult to standardize, and may not transfer cleanly across domains. This makes *unsupervised* segmentation particularly valuable in practice: a method that can be deployed without training labels and remains robust across datasets is often more useful than a narrowly optimized supervised model.

Change-point detection (CPD) provides a natural statistical framework for unsupervised segmentation, since boundaries correspond to indices where the underlying data distribution changes. Classical offline CPD methods come with strong guarantees, but these often rely on restrictive assumptions such as Gaussianity, independence, or homoscedasticity (Basseville & Nikiforov, 1993; Bai & Perron, 2003; Killick et al., 2012), which can be brittle for high-dimensional text representations. Kernel change-point detection (KCPD) relaxes much of this structure by comparing distributions through RKHS embeddings, enabling the detection of rich distributional shifts without explicit density estimation (Harchaoui & Cappe, 2007; Arlot et al., 2019). This makes KCPD a natural fit for embedding-based segmentation, where modern sentence encoders can capture semantic changes even when lexical cues are weak. However, applying KCPD to text also raises a theoretical challenge: most existing analyses focus on independent observations (Garreau & Arlot, 2018), whereas language exhibits clear short-range dependence because adjacent units share context, discourse structure, and lexical overlap. This motivates studying dependence-aware guarantees under simplified models of local dependence.

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

To address this gap, we provide, to our knowledge, the first theoretical analysis for dependent text-like sequences. We study KCPD under m-**dependence**, a tractable finite-memory model that serves as a stylized proxy for the short-range contextual dependence observed in language. Although natural language is not literally m-dependent, this abstraction lets us move beyond the usual i.i.d. setting and derive dependence-aware guarantees under explicit assumptions. In particular, we prove an oracle inequality for the population penalized risk and a conservative localization guarantee showing that each true change point is recovered within a window of width δ_T under spacing and signal conditions.

Building on this perspective, we introduce EMBED-KCPD and provide, to our knowledge, the first systematic study of kernel change-point detection for unsupervised text segmentation using pretrained sentence representations. Given a sequence of text units $X_1, \ldots, X_T$, we compute embeddings $Y_t = f(X_t) \in \mathbb{R}^d$ using a fixed encoder f and estimate boundaries by minimizing a penalized KCPD objective. Unlike approaches based on local lexical cohesion or pairwise embedding similarity, this formulation views semantic boundaries as distributional changes in representation space. The penalized objective can be optimized exactly with dynamic programming and efficiently with pruning via PELT after kernel precomputation. The pipeline is training-free and encoder-agnostic, decoupling representation learning from statistical segmentation.

We evaluate EMBED-KCPD along three complementary dimensions, spanning controlled statistical behavior, benchmark performance, and downstream utility. First, we introduce an LLM-based simulation framework that generates synthetic documents with known change points and controlled short-range dependence, providing a stylized testbed for studying the method under sequential dependence. The resulting trends qualitatively corroborate the theoretical predictions, with segmentation accuracy improving as the sequence length grows under the penalty scaling suggested by the analysis. Second, we evaluate EMBED-KCPD across standard text segmentation benchmarks and multiple modern encoders, where it is competitive with and often outperforms established unsupervised baselines. Third, we assess whether the resulting segments are useful beyond intrinsic boundary metrics by using them as chunks for retrieval-oriented evaluation in RAG. Across these settings, the results support the theoretical analysis while demonstrating that EMBED-KCPD is robust across representation choices, effective for unsupervised segmentation, and practically useful as a training-free chunking strategy for retrieval systems.

Contributions. Our main contributions are: (i) a dependence-aware analysis of penalized KCPD under m-**dependence**, a stylized finite-memory model of short-range dependence. To our knowledge, this is the first theoretical study of KCPD under m-dependence; we establish an oracle inequality for the population penalized risk together with a conservative change-point localization guarantee; (ii) EMBED-KCPD, a modular, training-free framework for applying KCPD to unsupervised text segmentation using pretrained sentence representations. To our knowledge, this is the first systematic study of KCPD for this task; (iii) an LLM-based simulation framework with known change points and controlled short-range dependence, providing a stylized testbed in which the observed scaling behavior qualitatively corroborates the theoretical predictions; and (iv) an extensive empirical evaluation across diverse text segmentation benchmarks, multiple modern encoders, and retrieval-oriented chunking for RAG, showing that EMBED-KCPD is robust across representation choices, competitive with and often stronger than established unsupervised baselines, and practically useful as a training-free chunking strategy for retrieval systems.

2 RELATED WORK

Change-point detection methods. Classical algorithms include Binary Segmentation (Scott & Knott, 1974), dynamic programming (Bai & Perron, 2003), and the Pruned Exact Linear Time (PELT) method (Killick et al., 2012), which offer consistency guarantees under parametric cost functions. Nonparametric approaches relax such assumptions using rank or divergence measures (Aminikhanghahi & Cook, 2017), while kernel methods embed data into reproducing kernel Hilbert spaces (Harchaoui et al., 2008). Recent work explores online and streaming algorithms for real-time detection (Ferrari et al., 2023; Hushchyn et al., 2020), ensemble and statistical inference methods for more reliable boundaries (Duy et al., 2020; Shiraishi et al., 2024), deep kernel learning for adaptive representations (Chang et al., 2019), and unsupervised deep frameworks (Truong et al., 2020).

Theoretical results on CPD beyond independence. Beyond independence, CPD under dependence has been studied mainly for parametric or low-dimensional settings: CUSUM/MOSUM with mixing and long-run variance or self-normalization (Csörgö & Horváth, 1997; Aue & Horváth, 2013; Horváth & Rice, 2014), econometric structural-break tests with robust covariances (Andrews, 1993; Bai & Perron, 1998), variance change via ICSS (Inclán & Tiao, 1994), and penalized-contrast methods for dependent series (Lavielle & Moulines, 2000; Lavielle, 2005), with extensions to high-dimensional mean shifts (Cho & Fryzlewicz, 2014; Wang & Samworth, 2017). To our knowledge, we provide the first theoretical results for non-parametric kernel CPD under m-dependence.

Text segmentation methods. Early methods like TextTiling (Hearst, 1994) exploit lexical cohesion, while later probabilistic approaches, including pLSA-based segmentation (Brants et al., 2002), dynamic programming over TF–IDF similarity (Fragkou et al., 2004), BayesSeg (Eisenstein & Barzilay, 2008), and LDA-based extensions (Riedl & Biemann, 2012; Du et al., 2013), model topical transitions via latent distributions. Recent techniques incorporate coherence-aware segmentation, semantic or embedding signals (Glavaš et al., 2016; Solbiati et al., 2021; Maraj et al., 2024; Yu et al., 2023; Gklezakos et al., 2024); mainly tailored to specific applications, rather than general-purpose text segmentation. In parallel, supervised methods frame segmentation as boundary classification, from attention-based BiLSTMs (Badjatiya et al., 2018) and hierarchical BiLSTMs (Koshorek et al., 2018), to Transformer variants using cross-segment attention (Lukasik et al., 2020) and multi-level Transformer designs (Somasundaran et al., 2020). On the contrary, our approach is fully unsupervised text segmentation.

Chunking for RAG systems. Early RAG systems typically follow passage-based dense retrieval pipelines, using fixed passages or short windows as the unit of indexing and retrieval (Karpukhin et al., 2020; Lewis et al., 2020; Izacard & Grave, 2021). Later work showed that retrieval granularity itself can materially affect performance, motivating finer-grained and semantically informed segmentation strategies (Chen et al., 2024). In parallel, the chunking literature includes classic unsupervised text segmentation methods, as well as more recent semantic, LLM-guided, hierarchical, and contextualized approaches (Hearst, 1997; Qu et al., 2025; Duarte et al., 2024; Sarthi et al., 2024; Günther et al., 2024). However, recent evaluations suggest that chunking quality is highly task-dependent and that expensive chunkers do not reliably outperform simpler baselines across settings (Qu et al., 2025; Bennani & Moslonka, 2026; Zhou et al., 2026; Shaukat et al., 2026). Our scope is not on LLM-based chunkers, which incur additional indexing cost, or structure-based methods, which depend on document markup that is not consistently available across corpora. Instead, we introduce EMBED-KCPD as a broadly applicable, unsupervised, LLM-free semantic chunking method.

3 EMBED-KCPD

3.1 KERNEL CHANGE POINT DETECTION

Let $Y_1, \cdots, Y_T \in \mathbb{R}^d$ be an observed sequence. A segmentation of $1, \cdots T$ into $K+1$ contiguous blocks is determined by change points $\boldsymbol{\tau}_K = (\tau_0, \tau_1, \ldots, \tau_K, \tau_{K+1})$ with $0 = \tau_0 < \tau_1 < \cdots < \tau_K < \tau_{K+1} = T$. We assume there exist true change points $\boldsymbol{\tau}_K$ such that the distribution of (Y_t) is piecewise stationary across the $K+1$ blocks. The task is to recover K and the locations $\tau_1, \ldots, \tau_K$.

Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive definite kernel with RKHS $\mathcal{H}$. The mapping function ϕ: $\mathbb{R}^d \to \mathcal{H}$ is implicitly defined by $\phi(y_t) = k(y_t, \cdot) \in \mathcal{H}$. For distributions P, Q, the squared maximum mean discrepancy is $\text{MMD}^2(P, Q) = \|\mu_P - \mu_Q\|^2_{\mathcal{H}}$. For data $Y_s, \ldots, Y_e$, define the empirical block cost

$$\widehat{C}(s, e) = \sum_{t=s}^{e} k(Y_t, Y_t) - \frac{1}{e-s+1} \sum_{i=s}^{e} \sum_{j=s}^{e} k(Y_i, Y_j),$$

with expectation $C(s, e) = \mathbb{E}[\widehat{C}(s, e)]$. Intuitively, $\widehat{C}(s, e)$ measures within-block dispersion in $\mathcal{H}$.

Penalized segmentation criterion. For a candidate segmentation $\boldsymbol{\tau}'_{K'}$, its cost is

$$L(\boldsymbol{\tau}'_{K'}) = \sum_{k=1}^{K'+1} \widehat{C}(\tau'_{k-1} + 1, \tau'_k) + \beta_T K',$$

where β_T penalizes over-segmentation. Let $\mathcal{P}_T := \Big\{\boldsymbol{\tau}'_{K'} : 0 = \tau_0 < \tau_1 < \cdots < \tau_{K'} < \tau_{K'+1} = T\Big\}$ denote the set of all possible partitions of $\{1, \ldots, T\}$. The KCPD estimator is

$$\widehat{\boldsymbol{\tau}}_{\widehat{K}} = \arg \min_{\boldsymbol{\tau}'_{K'} \in \mathcal{P}_T} L(\boldsymbol{\tau}'_{K'}).$$

L can be minimized exactly with the pruned exact linear time (PELT) algorithm, which under mild conditions has computational cost linear in the document length T.

3.2 Instantiation of KCPD for Text Segmentation

We now instantiate EMBED-KCPD as a general KCPD framework for text segmentation. The observed sequence $X_1, \ldots, X_T$ consists of contiguous text units (sentences, paragraphs, or dialogue turns). Each X_t is mapped to a normalized vector representation $Y_t = f(X_t) \in \mathbb{R}^d$, where f is a sentence-embedding model.

In this setting, segment boundaries correspond to topic or discourse changes that induce distributional shifts in the embedding sequence $Y_1, \ldots, Y_T$. KCPD is well suited to this problem because it detects general distributional changes in embedding space rather than only mean shifts, remains fully unsupervised, and can be combined directly with modern pretrained encoders. This yields a simple and modular pipeline: represent each text unit with a sentence embedding, then estimate boundaries by minimizing the penalized KCPD objective in Section 3.1.

3.3 KCPD Under m-Dependence

We now study the statistical behavior of $\widehat{\boldsymbol{\tau}}_{\widehat{K}}$ under dependent observations. Our goal is not to model natural language literally, but to analyze KCPD under a tractable finite-memory abstraction that captures short-range dependence between nearby text units. Specifically, we work under m-dependence, where observations more than m time steps apart are independent. This provides a stylized but useful proxy for the local contextual dependence that arises in sequential text.

The following assumptions formalize the setting.

Assumption 3.1 (m-dependence + within-block stationarity)**.** The sequence $(Y_t)_{t=1}^T$ is m-*dependent*: whenever $I, J \subseteq \{1, \ldots, T\}$ are finite index sets with $\min_{i \in I,\, j \in J} |i - j| > m$, the collections $\{Y_i : i \in I\}$ and $\{Y_j : j \in J\}$ are independent. Moreover, for each $k = 1, \ldots, K+1$, there exists a strictly stationary process with marginal distribution P_k such that the block $(Y_t)_{\tau_{k-1} < t \le \tau_k}$ is a finite stretch from that process.

Assumption 3.2 (kernel)**.** The kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is bounded and characteristic: $0 \le k(x, y) \le M < \infty$. Let $\mathcal{H}$ denote the associated RKHS.

Assumption 3.3 (detectability)**.** Let μ_{P_k} be the RKHS mean embedding of block k and define $\Delta_k^2 := \|\mu_{P_k} - \mu_{P_{k+1}}\|_{\mathcal{H}}^2 > 0$. Set $\Delta_\star^2 := \min_k \Delta_k^2 > 0$.

Assumption 3.4 (minimum spacing)**.** Let $\ell_T := \min_{1 \le k \le K+1}(\tau_k - \tau_{k-1})$ denote the minimum block length. Assume there exists a sequence $\delta_T \to \infty$ such that $\delta_T \le \ell_T/3$ for sufficiently large T.

Assumption 3.5 (penalty)**.** The penalty β_T satisfies $\beta_T \ \ge \ 16M\sqrt{2(8m+5)T \log T} + 2M(1 + 6m),\ \beta_T = O(\sqrt{T \log T})$.

For later use, define $\lambda_T \ := \ 4\sqrt{2}\, M\sqrt{(8m+5)\log T}$. For the localization analysis, we restrict attention to admissible partitions whose segments all have length at least δ_T. Define

$$\mathcal{P}_T(\delta_T) := \Big\{\boldsymbol{\tau}'_{K'} : 0 = \tau_0 < \tau_1 < \cdots < \tau_{K'} < \tau_{K'+1} = T,\ \tau_k - \tau_{k-1} \ge \delta_T,\ \forall k = 1, \ldots, K'+1\Big\}.$$

Assumption 3.6 (admissible segmentation scale)**.** For the localization analysis, the estimator $\widehat{\boldsymbol{\tau}}_{\widehat{K}}$ is defined as the minimizer of the penalized cost over $\mathcal{P}_T(\delta_T)$, where δ_T is the sequence from Assumption 3.4.

Assumption 3.7 (signal dominance at scale δ_T)**.** There exists a constant $C_0 > 6$ such that for all sufficiently large T, $\delta_T \Delta_\star^2 \ \ge \ C_0\Big(\beta_T + \lambda_T\sqrt{T} + \overline{B}_T\Big)$, where $\overline{B}_T = (4m+2)M + \frac{(2m^2+2m)M}{\delta_T}$.

Assumption 3.8 (Detectability on mixed intervals)**.** There exist constants $c_0 > 0$, $C_m \geq 0$, and T_0 such that $3c_0C_0 > 2$, and for all $T \geq T_0$, for every pair of consecutive change points (τ_k, τ_{k+1}) and every interval $[s, e]$ with $s \leq \tau_k < \tau_{k+1} \leq e$ and $e - s + 1 \geq 2\delta_T$, one has

$$\max_{t \in \mathcal{T}_{k,s,e}} \Big\{ C(s,e) - C(s,t) - C(t+1,e) \Big\} \;\geq\; c_0\, g_k\, \Delta_\star^2 - C_m, \qquad g_k := \tau_{k+1} - \tau_k,$$

where $\mathcal{T}_{k,s,e} := \big\{ t \in [\tau_k, \tau_{k+1} - 1] : t - s + 1 \geq \delta_T,\ e - t \geq \delta_T \big\}$ is the set of admissible split points inside $[\tau_k, \tau_{k+1}]$ for which both subsegments $[s, t]$ and $[t + 1, e]$ have length at least δ_T.

Assumptions 3.2 to 3.5 are standard or familiar in kernel change-point analysis, while Assumption 3.1 extends the usual setup to short-range dependence. Assumption 3.2 ensures the kernel cost is well behaved and that distributional changes are detectable through RKHS mean embeddings. Assumption 3.3 is a separation condition ensuring consecutive blocks differ nontrivially. Assumption 3.4 introduces an admissible localization scale δ_T and requires the true segments to be long enough to accommodate splits at that scale. Assumption 3.5 calibrates the penalty to prevent over-segmentation.

Assumptions 3.6, 3.7, and 3.8 are stronger and used only for the structural and localization guarantees. Assumption 3.6 imposes a minimum admissible scale for the analysis. Assumption 3.7 requires the accumulated signal at that scale to dominate the penalty and stochastic error. Assumption 3.8 is an additional mixed-interval detectability condition that rules out cancellation around true change points and ensures that an admissible split yields a population gain.

In the text setting, these assumptions should be interpreted as stylized rather than literal. Short-range dependence is natural because adjacent text units share context, discourse structure, and lexical overlap, while piecewise stationarity captures the idea that local segments are topically coherent. The spacing and detectability conditions are stronger worst-case assumptions used to obtain identifiable and localizable change points. Accordingly, we view the theory as principled support for EMBED-KCPD under a simplified dependence model, while the empirical sections evaluate its behavior on realistic text distributions.

3.4 THEORETICAL RESULTS

The first step is to control how well the empirical cost approximates the population cost, uniformly over all segments. A Bernstein type bound for each fixed segment is provided by Proposition A.1 in the appendix. A union bound over all segments yields:

Lemma 3.9 (uniform deviation over all segments)**.** *Let Assumptions 3.1 and 3.2 hold. Let* $\mathcal{E}_T := \Big\{ \forall 1 \leq s \leq e \leq T : \; |\widehat{C}(s,e) - C(s,e)| \leq \lambda_T \sqrt{e - s + 1} \Big\}$. *Then, for all integers* $T \geq 3$, $\Pr(\mathcal{E}_T) \geq 1 - T^{-1}$.

Informally, the lemma ensures that, with high probability, the empirical cost is a good approximation of the corresponding population cost for every segment in the sequence, simultaneously.

As a direct consequence of this concentration and the penalty choice in Assumption 3.5, we obtain a simple structural property on truly homogeneous regions.

Proposition 3.10 (stability on homogeneous segments)**.** *Let Assumptions 3.1, 3.2, and 3.5 hold. Then, with probability at least* $1 - T^{-1}$*, the following holds simultaneously for every segment* $[s, e]$ *that does not contain a true change point (that is,* $\tau_{k-1} < s \leq e < \tau_k$ *for some* k*) and every split point* t *with* $s \leq t < e$*:*

$$\widehat{C}(s,e) \;<\; \widehat{C}(s,t) \;+\; \widehat{C}(t{+}1,e) \;+\; \beta_T.$$

In simple terms, in a region where the distribution does not change, inserting an extra change point does not improve the penalized empirical objective, so the procedure has no incentive to create spurious splits inside stationary blocks.

For our first main result, we compare the population performance of the estimated segmentation to that of the best segmentation with the same penalty. This result only requires Assumption 3.1 and 3.2.

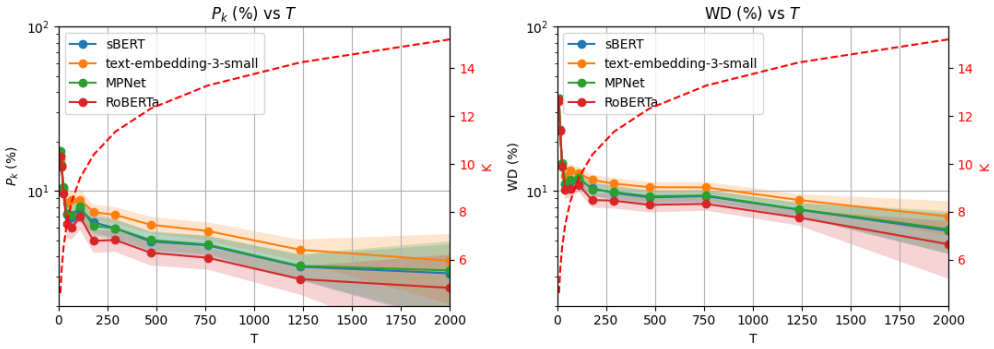


Figure 1: Segmentation accuracies versus sequence length T for EMBED-KCPD applied to synthetically generated short-range dependent text data with GPT-4.1 and $m = 20$. Curves compare four embedding methods (sBERT, MPNet, text-embedding-3-small, RoBERTa). Dashed red line shows the growth of the number of change points $K \approx 2 \log T$.

Theorem 3.11 (oracle inequality)**.** *Assume that Assumptions 3.1 and 3.2 hold. With probability at least* $1 - T^{-1}$,

$$\sum_{k=1}^{\widehat{K}+1} C(\widehat{\tau}_{k-1} + 1, \widehat{\tau}_k) + \beta_T \widehat{K} \;\le\; \inf_{\boldsymbol{\tau}'_{K'}} \Big\{ \sum_{k=1}^{K'+1} C(\tau'_{k-1} + 1, \tau'_k) + \beta_T K' \Big\} + 2\lambda_T T. \tag{1}$$

This result shows that, in terms of the ideal population criterion, our estimator performs almost as well as the best segmentation that could be chosen with full knowledge of the true block distributions, up to a controlled statistical error term arising from Lemma 3.9.

To understand individual change points, we use the stronger assumptions 3.3–3.8. A key structural consequence, proved via Lemma A.3 and Lemma A.4 in the appendix, is that the estimator does not merge multiple true changes into a single segment. *In simple terms, this means that each estimated segment can hide at most one true change point; the procedure does not lump several true changes together into a single segment.*

Combined with a strict improvement property for mixed segments (Lemma A.6 in the appendix) and the uniform deviation event $\mathcal{E}_T$, this leads to the localization guarantee. The following localization result is stated for the penalized estimator minimized over the admissible class $\mathcal{P}_T(\delta_T)$.

Theorem 3.12 (localization rate)**.** *Let Assumptions 3.1–3.8 hold. Let* δ_T *be the admissible scale from Assumptions 3.4 and 3.6. Then as* $T \to \infty$,

$$\Pr\Big(\widehat{K} \ge 1 \text{ and } \forall\, 1 \le k \le K : \min_{1 \le j \le \widehat{K}} |\widehat{\tau}_j - \tau_k| \le \delta_T\Big) \;\longrightarrow\; 1 \tag{2}$$

In particular,

$$\max_{1 \le k \le K} \min_{1 \le j \le \widehat{K}} |\widehat{\tau}_j - \tau_k| = O_p(\delta_T).$$

This result implies that, under our spacing and signal assumptions, every true change point is matched by an estimated one within a window of length δ_T, with probability tending to one. The worst-case localization error is therefore of order δ_T. Under Assumption 3.4, this window is at most one third of the minimum block length for all sufficiently large T, so the estimator localizes each change point to a controlled neighborhood inside its surrounding segment.

3.5 SIMULATION STUDY WITH CONTROLLED SHORT-RANGE DEPENDENCE

To study EMBED-KCPD under controlled sequential structure, we design a simulation based on synthetic sequences generated by GPT-4.1. The goal is not to reproduce the theoretical setting exactly, but to provide a stylized testbed with known boundaries and explicit short-range dependence.

We first generate five topic-specific documents (*soccer*, *coffee*, *AI*, *travel*, *dogs*), each with 500 sentences. Within each document, sentences are produced sequentially by prompting GPT4.1 to add

Table 1: Performance of Baselines and EMBED-KCPD in Wikipedia, Elements and arXiv Dataset. The bolded P_k or WD values denote values where EMBED-KCPD surpassed all unsupervised baselines. The last 3 rows serve only as a reference on supervised methods. x_y denotes mean x with standard deviation y. $*$ indicates values reported from the original papers.

Methods	Wiki-300		Wiki-50		Elements		arXiv	
	$P_k \downarrow$	WD $\downarrow$	$P_k \downarrow$	WD $\downarrow$	$P_k \downarrow$	WD $\downarrow$	$P_k \downarrow$	WD $\downarrow$
Unsupervised Methods								
EMBED-KCPD								
sBERT	$\mathbf{34.2}_{12.6}$	$\mathbf{37.3}_{13.7}$	$47.2_{17.4}$	$51.9_{21.0}$	$\mathbf{33.3}_{15.9}$	$44.0_{16.6}$	$\mathbf{11.2}_{9.5}$	$\mathbf{11.8}_{10.1}$
MPNet	$\mathbf{34.0}_{11.9}$	$\mathbf{35.1}_{11.3}$	$44.8_{17.3}$	$49.3_{20.1}$	$\mathbf{32.9}_{16.2}$	$43.3_{16.2}$	$\mathbf{14.7}_{11.1}$	$\mathbf{15.7}_{12.0}$
text-embedding-3-small	$\mathbf{33.8}_{12.6}$	$\mathbf{34.7}_{12.0}$	$43.9_{17.0}$	$48.5_{20.5}$	$\mathbf{32.1}_{16.3}$	$43.0_{17.4}$	$\mathbf{11.3}_{10.6}$	$\mathbf{11.7}_{11.0}$
RoBERTa	$\mathbf{33.1}_{12.9}$	$\mathbf{34.3}_{12.1}$	$44.5_{17.2}$	$49.2_{20.4}$	$\mathbf{33.8}_{16.5}$	$45.4_{16.8}$	$\mathbf{10.7}_{9.4}$	$\mathbf{11.7}_{10.1}$
Baselines								
Coherence	50.2^*	53.4^*	$53.5_{12.3}$	$71.1_{18.4}$	$42.4_{18.1}$	$54.7_{16.6}$	$43.0_{8.9}$	$45.4_{9.3}$
GraphSeg	$50.7_{11.4}$	$54.8_{12.8}$	$50.2_{17.1}$	$50.9_{18.7}$	$52.9_{19.3}$	$42.3_{16.3}$	$29.0_{11.7}$	$29.1_{12.1}$
TextTiling	$60.3_{9.1}$	$66.3_{11.2}$	$47.6_{11.8}$	$48.3_{11.8}$	$49.6_{18.3}$	$50.4_{20.5}$	$47.9_{9.1}$	$40.1_{7.7}$
TextTiling (MPNet)	$38.1_{12.4}$	$46.0_{14.3}$	$38.9_{14.5}$	$44.6_{15.4}$	$60.8_{19.3}$	$60.8_{19.3}$	$27.1_{7.2}$	$39.9_{7.9}$
TextTiling (sBERT)	$41.1_{14.6}$	$53.8_{18.9}$	$40.7_{13.8}$	$49.8_{19.0}$	$60.8_{19.0}$	$60.9_{18.9}$	$34.8_{8.3}$	$73.7_{10.5}$
Supervised Methods								
NTS	*34.4**	*31.5**	–	–	–	–	–	–
CATS	–	–	*16.5**	–	*18.4**	–	–	–
TextSeg	–	–	*18.2**	–	*41.6**	–	–	–

one sentence at a time, conditioning on the previous $m \in \{10, 20, 30\}$ sentences and the document topic. This induces short-range dependence and yields coherent topic-specific sequences.

For sequence lengths $T \leq 2000$, we set the number of change points to $K = \lceil 2 \log T \rceil$, randomize change-point locations, and assemble each sequence by concatenating segments drawn from a random selection of the five documents, with the constraint that consecutive segments have different topics. We generate 100 replicates for each (T, K). See Appendix D.5 for details. We then estimate change points with EMBED-KCPD using four sentence-embedding variants and a penalty of the form $\beta_T = C\sqrt{T \log T}$, following the asymptotic scaling suggested by the theory.

Evaluation metrics. Following prior work, we evaluate segmentation quality using P_k (Beeferman et al., 1999) and WindowDiff (WD) (Pevzner & Hearst, 2002). P_k measures whether nearby sentences are incorrectly assigned to the same or different segments, while WD compares the number of predicted and true boundaries within each window. Lower values indicate better segmentation. For both metrics, the window size is set to half the average true segment length.

Results. Figures 2 and 3 in Appendix C.2 summarize results on P_k as C and m vary. The value $C = 0.1$ gives the most stable performance as T increases, in line with the $\sqrt{T \log T}$ penalty scaling suggested by the theory. Results are also relatively insensitive to the choice of m, which is encouraging since the effective dependence range is typically unknown in practice. Full results for $C = 0.1$ and $m = 20$ are shown in Fig. 1. Empirically, both P_k and WD decrease as T grows with K scaling as above, indicating that EMBED-KCPD becomes more accurate on these controlled synthetic sequences as more data are available.

These simulations should therefore be viewed as a controlled stress test under stylized short-range dependence, rather than as a direct verification of the theoretical assumptions.

4 TEXT SEGMENTATION EVALUATION

Datasets. We evaluate our methods on several widely used datasets for text segmentation. Choi's dataset (Choi, 2000), consisting of 700 synthetic documents, serves as the benchmark for segmentation performance, although highly synthetic and left to the Appendix C.1. Wiki-300, introduced by (Badjatiya et al., 2018), contains 300 documents. We also include two smaller datasets: Wiki-50 introduced by Koshorek et al. (2018) and Elements (Chen et al., 2009) about 118 chemical elements. In addition, we construct a new dataset of 50 documents by randomly selecting some recent abstracts from arXiv papers and concatenating them to form one document, to add a clean dataset unknown

Table 2: Retrieval performance across three QA benchmarks at k=5. EMBED-KCPD choosing C with the elbow method. For all other baselines, the best result among all settings is reported; **+ov** denotes segment-level overlap; *Random* is Random-matched (same chunk count as KCPD Base, random boundaries). Bold = best per column. †NQ has one gold passage per question so Hit = Rec = Cov. Confidence intervals available in Appendix Table 8.

Method	HotpotQA				QASPER				NQ†			
	Hit	Rec	Cov	MRR	Hit	Rec	Cov	MRR	Hit	Rec	Cov	MRR
Baselines												
Fixed	0.60	0.32	0.09	0.41	0.41	0.24	0.13	0.25	0.32	0.32	0.32	0.22
Overlap	0.20	0.09	0.00	0.12	0.48	0.29	0.16	0.28	0.36	0.36	0.36	0.23
Sent. Budget	0.43	0.20	0.02	0.29	0.45	0.27	0.16	0.29	0.36	0.36	0.36	0.23
TextTiling	0.08	0.03	0.00	0.05	0.43	0.24	0.13	0.25	0.32	0.32	0.32	0.21
TextTiling +ov	0.01	0.00	0.00	0.00	0.27	0.16	0.08	0.18	0.24	0.24	0.24	0.14
SemanticDrop	0.62	0.33	0.08	0.43	0.41	0.24	0.14	0.26	0.27	0.27	0.27	0.18
SemanticDrop +ov	0.19	0.08	0.00	0.13	0.44	0.26	0.14	0.28	0.28	0.28	0.28	0.19
EMBED-KCPD												
Base	**0.66**	**0.35**	**0.09**	**0.48**	0.44	0.26	0.13	0.29	0.32	0.32	0.32	0.21
+ov	0.07	0.03	0.00	0.04	**0.50**	**0.32**	**0.19**	**0.32**	**0.40**	**0.40**	**0.40**	**0.28**
Random	0.56	0.28	0.05	0.38	0.38	0.23	0.13	0.22	0.23	0.23	0.23	0.15

to all baseline methods. A summary of dataset statistics is presented in Table D.1. The detailed procedure for constructing the arXiv dataset is in Appendix D.6.

Experimental details. For each dataset, we apply EMBED-KCPD with RBF kernel, using four modern sentence embeddings for text segmentation: sBERT (Reimers & Gurevych, 2019), MPNet (Song et al., 2020), text-embedding-3-small (OpenAI, 2025), and RoBERTa (Liu et al., 2019). As unsupervised baselines, we include TextTiling (Hearst, 1994), GraphSeg (Glavaš et al., 2016), and Coherence (Maraj et al., 2024), and compare their performance with EMBED-KCPD across all datasets. We also compare with a modern version of TextTiling using modern embeddings, following saeedabc (2025) tuning configuration for sBERT and MPNet embeddings. For the comparison of Choi's dataset Choi (2000), we further compare other unsupervised methods, (Choi, 2000; Brants et al., 2002; Fragkou et al., 2004; Misra et al., 2009). See Appendix D.2 for more implementation details. For the Wikipedia-based datasets, we additionally include supervised approaches reported in prior work: NTS (Badjatiya et al., 2018), CATS (Somasundaran et al., 2020), TextSeg (Koshorek et al., 2018). We use $\beta_T = C\sqrt{T \log T}$. We select C using an unsupervised elbow method (see Appendix D.3 for details). Figure 8 in Appendix indicates that performance remains stable across a range of C values.

4.1 RESULTS

Table 1 summarizes results on more realistic datasets: Wiki-300, Wiki-50, Elements, and arXiv. We include supervised methods for reference.

Comparing EMBED-KCPD to unsupervised baselines. EMBED-KCPD variants outperform all baselines across most datasets and evaluation metrics. As shown in Table 1, EMBED-KCPD achieves lower P_k and WD in nearly all settings, with few exceptions. Importantly, even when TextTiling is augmented with sentence embeddings, EMBED-KCPD typically achieves superior performance, indicating that its gains are not solely attributable to the use of embeddings. These results demonstrate the effectiveness of EMBED-KCPD as an unsupervised method.

Comparing embeddings. EMBED-KCPD with text-embedding-3-small yields the lowest P_k and WD on Wiki-50, while RoBERTa achieves the lowest score on the remaining three datasets. Overall, performance differences across embeddings are modest, underscoring the robustness of EMBED-KCPD to embedding choices.

Comparing EMBED-KCPD with supervised methods. On Wiki-300, EMBED-KCPD achieves lower P_k than Badjatiya et al. (2018) across all embeddings, with WD approaching the supervised baseline. On Elements, EMBED-KCPD attains lower P_k than Koshorek et al. (2018) for most embeddings, with the exception of text-embedding-3-small, where performance remains close. These

findings suggest that EMBED-KCPD, despite being unsupervised, achieves performance comparable to strong supervised methods.

5 RAG-BASED CHUNKING EVALUATION

Standard segmentation metrics such as P_k and WindowDiff measure agreement with annotated boundaries, but they do not directly test downstream utility. We therefore include a retrieval-based evaluation to test whether unsupervised text segmentation produces more useful chunks for evidence retrieval than standard chunking mechanisms.

Datasets. We evaluate chunking quality for retrieval-augmented question answering on three QA benchmarks spanning short- and long-document retrieval settings. HotpotQA (Yang et al., 2018) is evaluated in the distractor setting, where ten paragraphs, including two gold and eight distractors, are concatenated into a single document per question, yielding documents of roughly 500 to 1,500 tokens; gold evidence is annotated at the sentence level. We sample 400 validation questions. QASPER (Dasigi et al., 2021) pairs each question with a full scientific paper of roughly 5,000 to 10,000 tokens, flattened into a single document, with gold evidence given as an annotated paragraph string. We use 300 test questions. Natural Questions (NQ) (Kwiatkowski et al., 2019) pairs each question with a full Wikipedia article of roughly 5,000 to 30,000 tokens, with the annotated long-answer paragraph serving as gold evidence. We sample 400 validation questions. All samples are drawn with a fixed random seed of 42.

Experimental details. We compare EMBED-KCPD with several chunking strategies under a common retrieval pipeline. For methods that require sentence representations, we use `text-embedding-3-small`. As baselines, we consider Fixed chunking with non-overlapping token windows of size $w \in \{50, 100, 200\}$, Overlap chunking with windows of size $2w$ and stride w, and SentBudget, which greedily packs spaCy sentences up to a $2w$ token budget. We also include TextTiling and SemanticDrop, which places a boundary whenever the cosine similarity between consecutive sentence embeddings falls below $\mu - K\sigma$, for $K \in \{0.25, 0.5, 1.0\}$. For EMBED-KCPD, SemanticDrop, and TextTiling, we also evaluate overlapping variants (+ov) by augmenting each predicted segment with the immediately following segment. We choose C for EMBED-KCPD using the elbow method. To isolate the effect of semantically informed boundary placement from the number of chunks, we include a Random-matched ablation that places the same number of breakpoints as EMBED-KCPD uniformly at random over sentence positions. See Appendix E for details. For retrieval, chunks and queries are embedded with `text-embedding-3-small`. A retrieved chunk is counted as matching a gold evidence unit when their token-level Jaccard overlap is at least 0.5. We report retrieval quality at depth $k = 5$ using Hit@5, Recall@5, Coverage@5, and MRR@5.

Results. Table 2 reports the EMBED-KCPD result obtained by choosing C with the elbow method (Appendix E). For all other benchmarks, the best result among all settings is reported. Results show that EMBED-KCPD is the best performing unsupervised chunking method across all three QA benchmarks, thus the comparison is conservative for our method (See full results in Appendix Tables 5-7). On HotpotQA, EMBED-KCPD achieves the top score surpassing the strongest baseline, SemanticDrop. On QASPER and NQ, EMBED-KCPD+ov delivers the best overall results. Overall, the results indicate that EMBED-KCPD provides more retrieval-effective segmentations than both fixed-length and semantic alternatives, and benefits further from overlap augmentation on longer, more heterogeneous documents. A further practical advantage of EMBED-KCPD over traditional length-based chunking methods is that these approaches require the chunk size to be chosen ad hoc, or tuned in a supervised manner, which is undesirable in a fully unsupervised setting. By contrast, C for EMBED-KCPD can be selected with a standard elbow method, as described in Appendix E.

6 CONCLUSION

We introduced EMBED-KCPD, a simple and training-free method for unsupervised text segmentation, and demonstrated across synthetic and real benchmarks that it is a strong practical baseline. We showed that its segments are useful beyond intrinsic evaluation by serving as effective chunks for retrieval-augmented question answering. Complementing these empirical results, we analyzed KCPD under m-dependence and proved an oracle inequality together with a conservative localization guarantee, providing principled support under a stylized finite-memory dependence model.

AI USE STATEMENT

In this work, we used generative AI tools to generate the synthetic text data used in the controlled short-range dependence experiments, as described in Section 3.5. Specifically, GPT-4.1 was used to generate topic-specific synthetic documents following the prompting procedure described in the paper. We have not used generative AI tools for other tasks requiring disclosure, and the remaining required-disclosure tasks are not applicable to this work.

Additionally, we used generative AI tools for language editing, including rephrasing, polishing, and grammar checking. We reviewed all AI-assisted content and verified that the generated synthetic data and edited text were consistent with the experimental setup and intended meaning. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The assumptions and main theoretical results are presented in Section 3.3 and Section 3.4, with complete proofs provided in the appendix. The construction of the synthetic short-range dependent data is described in Section 3.5, with further details in the appendix. Experimental settings, datasets, embeddings, baselines, and penalty selection for text segmentation are described in Section 4 and Appendix D. The RAG evaluation is described in Section 5 and Appendix E. The code is included in the supplementary materials and will be released publicly on GitHub upon acceptance.

REFERENCES

Samaneh Aminikhanghahi and Diane J. Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367, 2017. ISSN 0219-3116. doi: 10.1007/s10115-016-0987-z. URL https://doi.org/10.1007/s10115-016-0987-z.

Donald W. K. Andrews. Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(4):821–856, 1993. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2951764.

Sylvain Arlot, Alain Celisse, and Zaid Harchaoui. A kernel multiple change-point algorithm via model selection. *Journal of Machine Learning Research*, 20(162):1–56, 2019. URL http://jmlr.org/papers/v20/16-155.html.

Alexander Aue and Lajos Horváth. Structural breaks in time series. *Journal of Time Series Analysis*, 34(1):1–16, 2013. doi: https://doi.org/10.1111/j.1467-9892.2012.00819.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9892.2012.00819.x.

Pinkesh Badjatiya, Litton J. Kurisinkel, Manish Gupta, and Vasudeva Varma. Attention-based neural text segmentation. In Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury (eds.), *Advances in Information Retrieval*, pp. 180–193, Cham, 2018. Springer International Publishing. ISBN 978-3-319-76941-7.

Jushan Bai and Pierre Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78, 1998. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2998540.

Jushan Bai and Pierre Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22, 2003. doi: https://doi.org/10.1002/jae.659. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.659.

Michèle Basseville and Igor Nikiforov. *Detection of Abrupt Changes: Theory and Application*, volume 15. 04 1993. ISBN 0-13-126780-9.

Doug Beeferman, Adam Berger, and John Lafferty. Statistical models for text segmentation. *Machine Learning*, 34(1):177–210, 1999. ISSN 1573-0565. doi: 10.1023/A:1007506220214. URL https://doi.org/10.1023/A:1007506220214.

Sofia Bennani and Charles Moslonka. A systematic analysis of chunking strategies for reliable question answering. *arXiv preprint arXiv:2601.14123*, 2026. doi: 10.48550/arXiv.2601.14123. URL https://arxiv.org/abs/2601.14123. Latest public version checked: v1.

Thorsten Brants, Francine Chen, and Ioannis Tsochantaridis. Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM '02, pp. 211–218, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581134924. doi: 10.1145/584792.584829. URL https://doi.org/10.1145/584792.584829.

Wei-Cheng Chang, Chun-Liang Li, Yiming Yang, and Barnabás Póczos. Kernel change-point detection with auxiliary deep generative models. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1GbfhRqF7.

Harr Chen, S.R.K. Branavan, Regina Barzilay, and David R. Karger. Global models of document structure using latent permutations. In Mari Ostendorf, Michael Collins, Shri Narayanan, Douglas W. Oard, and Lucy Vanderwende (eds.), *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 371–379, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL https://aclanthology.org/N09-1042/.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. Dense x retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15159–15177, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.845. URL https://aclanthology.org/2024.emnlp-main.845/.

Haeran Cho and Piotr Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(2):475–507, 07 2014. ISSN 1369-7412. doi: 10.1111/rssb.12079. URL https://doi.org/10.1111/rssb.12079.

Sangwoo Cho, Kaiqiang Song, Xiaoyang Wang, Fei Liu, and Dong Yu. Toward unifying text segmentation and long document summarization. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 106–118, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.8. URL https://aclanthology.org/2022.emnlp-main.8/.

Freddy Y. Y. Choi. Advances in domain independent linear text segmentation. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 2000. URL https://aclanthology.org/A00-2004/.

M. Csörgö and L. Horváth. *Limit Theorems in Change-Point Analysis*. Wiley Series in Probability and Statistics. Wiley, 1997. ISBN 9780471955221. URL https://books.google.ca/books?id=iyXvAAAAMAAJ.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365/.

Lan Du, Wray Buntine, and Mark Johnson. Topic segmentation with a structured topic model. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff (eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 190–200, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1019/.

André V. Duarte, João DS Marques, Miguel Graça, Miguel Freire, Lei Li, and Arlindo L. Oliveira. Lumberchunker: Long-form narrative document segmentation. In *Findings of the Association*

for Computational Linguistics: EMNLP 2024, pp. 6473–6486, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.377. URL https://aclanthology.org/2024.findings-emnlp.377/.

Vo Nguyen Le Duy, Hiroki Toda, Ryota Sugiyama, and Ichiro Takeuchi. Computing valid p-value for optimal changepoint by selective inference using dynamic programming. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 11356–11367. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/82b04cd5aa016d979fe048f3ddf0e8d3-Paper.pdf.

Jacob Eisenstein and Regina Barzilay. Bayesian unsupervised topic segmentation. In Mirella Lapata and Hwee Tou Ng (eds.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 334–343, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. URL https://aclanthology.org/D08-1035/.

André Ferrari, Cédric Richard, Anthony Bourrier, and Ikram Bouchikhi. Online change-point detection with kernels. *Pattern Recognition*, 133:109022, 2023. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2022.109022. URL https://www.sciencedirect.com/science/article/pii/S0031320322005027.

Paraskevi Fragkou, Vassilis Petridis, and Alexandros Kehagias. A dynamic programming algorithm for linear text segmentation. *Journal of Intelligent Information Systems*, 23(2):179–197, 2004. doi: 10.1023/B:JIIS.0000039534.65423.00. URL https://doi.org/10.1023/B:JIIS.0000039534.65423.00.

Damien Garreau and Sylvain Arlot. Consistent change-point detection with kernels. *Electronic Journal of Statistics*, 12(2):4440 – 4486, 2018. doi: 10.1214/18-EJS1513. URL https://doi.org/10.1214/18-EJS1513.

Dimitrios C. Gklezakos, Timothy Misiak, and Diamond Bishop. Treeseg: Hierarchical topic segmentation of large transcripts, 2024. URL https://arxiv.org/abs/2407.12028.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. Unsupervised text segmentation using semantic relatedness graphs. In Claire Gardent, Raffaella Bernardi, and Ivan Titov (eds.), *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pp. 125–130, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-2016. URL https://aclanthology.org/S16-2016/.

Michael Günther, Isabelle Mohr, Daniel James Williams, Bo Wang, and Han Xiao. Late chunking: Contextual chunk embeddings using long-context embedding models. *arXiv preprint arXiv:2409.04701*, 2024. doi: 10.48550/arXiv.2409.04701. URL https://arxiv.org/abs/2409.04701. Latest public version: v3, revised 2025-07-07.

Zaid Harchaoui and Olivier Cappe. Retrospective mutiple change-point estimation with kernels. In *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, pp. 768–772, 2007. doi: 10.1109/SSP.2007.4301363.

Zaïd Harchaoui, Eric Moulines, and Francis Bach. Kernel change-point analysis. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (eds.), *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper_files/paper/2008/file/08b255a5d42b89b0585260b6f2360bdd-Paper.pdf.

Marti A. Hearst. Multi-paragraph segmentation expository text. In *32nd Annual Meeting of the Association for Computational Linguistics*, pp. 9–16, Las Cruces, New Mexico, USA, June 1994. Association for Computational Linguistics. doi: 10.3115/981732.981734. URL https://aclanthology.org/P94-1002/.

Marti A. Hearst. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997. URL https://aclanthology.org/J97-1003/.

Lajos Horváth and Gregory Rice. Extensions of some classical methods in change point analysis. *TEST*, 23(2):219–255, 2014. doi: 10.1007/s11749-014-0368-4.

Mikhail Hushchyn, Kenenbek Arzymatov, and Denis Derkach. Online neural networks for change-point detection, 2020. URL https://arxiv.org/abs/2010.01388.

Carla Inclán and George C. Tiao. Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89(427):913–923, 1994. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2290916.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74/.

Svante Janson. Large deviations for sums of partly dependent random variables. *Random Structures & Algorithms*, 24(3):234–248, 2004. doi: https://doi.org/10.1002/rsa.20008. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.20008.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012. doi: 10.1080/01621459.2012.737745. URL https://doi.org/10.1080/01621459.2012.737745.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. Text segmentation as a supervised learning task. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 469–473, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2075. URL https://aclanthology.org/N18-2075/.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal Processing*, 85(8):1501–1510, 2005. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2005.01.012. URL https://www.sciencedirect.com/science/article/pii/S0165168405000381.

Marc Lavielle and Eric Moulines. Least-squares estimation of an unknown number of shifts in a time series. *Journal of Time Series Analysis*, 21(1):33–59, 2000. doi: https://doi.org/10.1111/1467-9892.00172. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9892.00172.

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 Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

Fernando Llopis, Antonio Ferrández Rodríguez, and José Luis Vicedo González. Text segmentation for efficient information retrieval. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, pp. 373–380, Berlin, Heidelberg, 2002. Springer-Verlag. ISBN 3540432191.

Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Simões. Text segmentation by cross segment attention. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4707–4716, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.380. URL https://aclanthology.org/2020.emnlp-main.380/.

Amit Maraj, Miguel Vargas Martin, and Masoud Makrehchi. Words that stick: Using keyword cohesion to improve text segmentation. In Libby Barak and Malihe Alikhani (eds.), *Proceedings of the 28th Conference on Computational Natural Language Learning*, pp. 1–9, Miami, FL, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.conll-1.1. URL https://aclanthology.org/2024.conll-1.1/.

Hemant Misra, François Yvon, Joemon M. Jose, and Olivier Cappe. Text segmentation via topic modeling: an analytical study. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pp. 1553–1556, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585123. doi: 10.1145/1645953.1646170. URL https://doi.org/10.1145/1645953.1646170.

OpenAI. *OpenAI Platform Documentation*, 2025. URL https://platform.openai.com/docs/overview.

Lev Pevzner and Marti A. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, 2002. doi: 10.1162/089120102317341756. URL https://aclanthology.org/J02-1002/.

Violaine Prince and Alexandre Labadié. Text segmentation based on document understanding for information retrieval. In Zoubida Kedad, Nadira Lammari, Elisabeth Métais, Farid Meziane, and Yacine Rezgui (eds.), *Natural Language Processing and Information Systems*, pp. 295–304, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-73351-5.

Renyi Qu, Ruixuan Tu, and Forrest Sheng Bao. Is semantic chunking worth the computational cost? In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2155–2177, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.114. URL https://aclanthology.org/2025.findings-naacl.114/.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Martin Riedl and Chris Biemann. TopicTiling: A text segmentation algorithm based on LDA. In Jackie C. K. Cheung, Jun Hatori, Carlos Henriquez, and Ann Irvine (eds.), *Proceedings of ACL 2012 Student Research Workshop*, pp. 37–42, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/W12-3307/.

saeedabc. Extended texttiling using llm embeddings for text segmentation. https://github.com/saeedabc/llm-text-tiling, 2025. [Software].

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. RAPTOR: recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, 2024. URL https://openreview.net/forum?id=GN921JHCRw.

A. J. Scott and M. Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3):507–512, 1974. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2529204.

Muhammad Arslan Shaukat, Muntasir Adnan, and Carlos C. N. Kuhn. A systematic investigation of document chunking strategies and embedding sensitivity. *arXiv preprint arXiv:2603.06976*, 2026. doi: 10.48550/arXiv.2603.06976. URL https://arxiv.org/abs/2603.06976. Latest public version checked: v1.

Tomohiro Shiraishi, Daiki Miwa, Vo Nguyen Le Duy, and Ichiro Takeuchi. Selective inference for change point detection by recurrent neural network. *Neural Computation*, 37(1):160–192, 12 2024. ISSN 0899-7667. doi: 10.1162/neco_a_01724. URL https://doi.org/10.1162/neco_a_01724.

Gennady Shtekh, Polina Kazakova, Nikita Nikitinsky, and Nikolay Skachkov. Applying topic segmentation to document-level information retrieval. In *Proceedings of the 14th Central and Eastern European Software Engineering Conference Russia*, CEE-SECR '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450361767. doi: 10.1145/3290621.3290630. URL https://doi.org/10.1145/3290621.3290630.

Alessandro Solbiati, Kevin Heffernan, Georgios Damaskinos, Shivani Poddar, Shubham Modi, and Jacques Cali. Unsupervised topic segmentation of meetings with bert embeddings, 2021. URL https://arxiv.org/abs/2106.12978.

Swapna Somasundaran et al. Two-level transformer and auxiliary coherence modeling for improved text segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7797–7804, 2020.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2019.107299. URL https://www.sciencedirect.com/science/article/pii/S0165168419303494.

Tengyao Wang and Richard J. Samworth. High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):57–83, 08 2017. ISSN 1369-7412. doi: 10.1111/rssb.12243. URL https://doi.org/10.1111/rssb.12243.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Hai Yu, Chong Deng, Qinglin Zhang, Jiaqing Liu, Qian Chen, and Wen Wang. Improving long document topic segmentation models with enhanced coherence modeling. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5592–5605, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.341. URL https://aclanthology.org/2023.emnlp-main.341/.

Yongjie Zhou, Shuai Wang, Bevan Koopman, and Guido Zuccon. Beyond chunk-then-embed: A comprehensive taxonomy and evaluation of document chunking strategies for information retrieval. *arXiv preprint arXiv:2602.16974*, 2026. doi: 10.48550/arXiv.2602.16974. URL https://arxiv.org/abs/2602.16974. Latest public version checked: v1.

A PROOFS

A.1 AUXILIARY RESULTS FOR LEMMA 3.9

Proposition A.1 (m-dependent concentration for segment cost)**.** *Fix integers $1 \le s \le e \le T$ and set $n = e - s + 1$. Under Assumptions 3.1–3.2, for every $x > 0$,*

$$\Pr\big(|\widehat{C}(s,e) - C(s,e)| > x\big) \;\le\; 4\exp\Big(-\frac{x^2}{8(8m+5)M^2\,n}\Big).$$

Proof. Write

$$\widehat{C}(s,e) - C(s,e) = \underbrace{\sum_{t=s}^{e}\big(k(Y_t,Y_t) - \mathbb{E}[k(Y_t,Y_t)]\big)}_{=:A} - \underbrace{\frac{1}{n}\sum_{i=s}^{e}\sum_{j=s}^{e}\big(k(Y_i,Y_j) - \mathbb{E}[k(Y_i,Y_j)]\big)}_{=:B}.$$

Since $0 \le k \le M$, each centered summand is bounded in absolute value by M.

We use Janson's inequality for sums with a dependency graph (Thm. 2.1 Janson (2004)). If $\{X_v\}_{v\in V}$ are centered, $|X_v| \le b$, and $G = (V, E)$ is a dependency graph with chromatic number $\chi(G)$, then for any $t > 0$,

$$\Pr\Big(\Big|\sum_{v\in V} X_v\Big| > t\Big) \;\le\; 2\exp\Big(-\frac{t^2}{2\,\chi(G)\,|V|\,b^2}\Big). \tag{3}$$

For A, take $V_A = \{s, \dots, e\}$ and connect t, t' when $|t - t'| \le m$. This is a valid dependency graph by m-dependence (Assumption 3.1): variables further than m apart are independent. The graph is properly colored by $t \bmod (m+1)$, hence $\chi(G_A) \le m+1$ and $|V_A| = n$. Applying (3) with $b = M$ and threshold $t = x/2$ gives

$$\Pr\big(|A| > x/2\big) \;\le\; 2\exp\Big(-\frac{x^2}{8(m+1)nM^2}\Big). \tag{4}$$

Write $B = \frac{1}{n}S$ with

$$S := \sum_{i=s}^{e}\sum_{j=s}^{e} Z_{ij}, \qquad Z_{ij} := k(Y_i,Y_j) - \mathbb{E}[k(Y_i,Y_j)].$$

We consider ordered pairs (i, j) so that $|V_B| = n^2$. Define a dependency graph G_B on $V_B = \{(i,j) : s \le i, j \le e\}$ by connecting (i, j) and (i', j') iff

$$\min\{|i - i'|,\, |i - j'|,\, |j - i'|,\, |j - j'|\} \;\le\; m.$$

Each Z_{ij} is a function of (Y_i, Y_j). If two disjoint vertex sets $U, W \subseteq V_B$ have no edges between them, then the index sets of Y's underlying U and W are pairwise more than m apart in time, hence independent by m-dependence; therefore $\{Z_u : u \in U\}$ and $\{Z_w : w \in W\}$ are independent, as required.

Fix (i, j). Let $T_{ij} := \{k : |k - i| \le m \text{ or } |k - j| \le m\}$; then $|T_{ij}| \le (2m+1) + (2m+1) = 4m + 2$. Any neighbor (i', j') must satisfy $i' \in T_{ij}$ or $j' \in T_{ij}$. Thus the number of neighbors is at most

$$n\,|T_{ij}| \;+\; n\,|T_{ij}| \;\le\; 2n(4m+2) \;=\; (8m+4)\,n,$$

so $\Delta(G_B) \le (8m+4)n$ and hence $\chi(G_B) \le \Delta(G_B) + 1 \le (8m+4)n + 1 \le (8m+5)n$ for $n \ge 1$. Applying (3) to S with $b = M$, $|V_B| = n^2$, $\chi(G_B) \le (8m+5)n$, and threshold $t = nx/2$ yields

$$\Pr\big(|B| > x/2\big) \;=\; \Pr\big(|S| > nx/2\big) \;\le\; 2\exp\Big(-\frac{x^2}{8(8m+5)nM^2}\Big). \tag{5}$$

If $|\widehat{C}(s,e) - C(s,e)| = |A - B| > x$, then $|A| > x/2$ or $|B| > x/2$. Hence, by (4)–(5),

$$\Pr\big(|\widehat{C}(s,e) - C(s,e)| > x\big) \;\le\; 2\exp\Big(-\frac{x^2}{8(m+1)nM^2}\Big) + 2\exp\Big(-\frac{x^2}{8(8m+5)nM^2}\Big) \;\le\; 4\exp\Big(-\frac{x^2}{8(8m+5)nM^2}\Big),$$

where the last inequality uses $(8m+5) \ge (m+1)$ for all $m \ge 0$. This completes the proof. □

A.2 Proof of Lemma 3.9

Fix $[s, e]$ with length $n = e - s + 1$. By Proposition A.1, with $x = \lambda_T \sqrt{n}$,

$$\Pr\big(|\widehat{C}(s,e) - C(s,e)| > x\big) \;\le\; 4\exp\Big(-\frac{x^2}{8(8m+5)M^2 n}\Big) \;=\; 4\exp(-4\log T) \;=\; 4T^{-4}.$$

There are $\frac{T(T+1)}{2}$ segments, so by a union bound,

$$\Pr(\mathcal{E}_T^{\mathrm{c}}) \;\le\; \frac{T(T+1)}{2} \cdot 4T^{-4} \;=\; \frac{2(T+1)}{T^3} \;\le\; T^{-1} \qquad \text{for all } T \ge 3,$$

since $T^2 - 2T - 2 \ge 0$ for $T \ge 3$. Hence $\Pr(\mathcal{E}_T) \ge 1 - T^{-1}$.

A.3 Proof of Proposition 3.10

Fix a clean segment $[s, e]$ (i.e., $\tau_{k-1} < s \le e < \tau_k$) and $t \in \{s, \dots, e-1\}$. Write

$$\Delta\widehat{C}(a,b) := \widehat{C}(a,b) - C(a,b), \qquad \Delta_C := C(s,t) + C(t{+}1,e) - C(s,e).$$

We aim to lower bound

$$\Big[\widehat{C}(s,t) + \widehat{C}(t{+}1,e) - \widehat{C}(s,e)\Big] + \beta_T = \underbrace{\Delta_C}_{\text{expectation}} + \underbrace{\big(\Delta\widehat{C}(s,t) + \Delta\widehat{C}(t{+}1,e) - \Delta\widehat{C}(s,e)\big)}_{\text{deviation}} + \beta_T.$$

On the event $\mathcal{E}_T$ of Lemma 3.9 (which holds with probability $\ge 1 - T^{-1}$), for all $1 \le a \le b \le T$,

$$|\Delta\widehat{C}(a,b)| \le \lambda_T \sqrt{b - a + 1}, \qquad \lambda_T := 4\sqrt{2}\, M \sqrt{(8m+5)\log T}.$$

Hence, for any $s \le t < e$,

$$\begin{aligned}
\Delta\widehat{C}(s,t) + \Delta\widehat{C}(t{+}1,e) - \Delta\widehat{C}(s,e) &\ge -\big(|\Delta\widehat{C}(s,t)| + |\Delta\widehat{C}(t{+}1,e)| + |\Delta\widehat{C}(s,e)|\big) \\
&\ge -\lambda_T\big(\sqrt{t-s+1} + \sqrt{e-t} + \sqrt{e-s+1}\big) \\
&\ge -3\lambda_T\sqrt{T}.
\end{aligned}$$

Because $[s, e]$ lies within a single stationary block (Assumption 3.1), $C(s, e)$ depends only on the length $n := e - s + 1$. Denote $C(n) := C(s, e)$. Set $n_1 := t - s + 1$, $n_2 := e - t$, so $n = n_1 + n_2$. For a stationary segment of length n,

$$C(n) = (n-1)c_0 - \frac{2}{n}\sum_{l=1}^{n-1}(n-l)c_l, \quad \text{where} \quad c_l := \mathbb{E}\big[k(Y_1, Y_{1+l})\big]. \tag{6}$$

Under m-dependence, Y_1 and Y_{1+l} are independent for $l > m$, hence by bilinearity of the RKHS inner product (no "characteristic" property needed),

$$c_l = \mathbb{E}\,\langle \phi(Y_1), \phi(Y_{1+l})\rangle_{\mathcal{H}} = \langle \mathbb{E}\,\phi(Y_1), \mathbb{E}\,\phi(Y_{1+l})\rangle_{\mathcal{H}} = \|\mu_P\|_{\mathcal{H}}^2 =: c_\infty \qquad (l > m).$$

Define $\delta_l := c_l - c_\infty$; then $\delta_l = 0$ for $l > m$ and, since $|k| \le M$ (Assumption 3.2), we have $|c_l| \le M$, $|c_\infty| \le M$, thus $|\delta_l| \le 2M$. Plugging $c_l = c_\infty + \delta_l$ into (6) and using $\sum_{l=1}^{n-1}(n-l) = \frac{n(n-1)}{2}$ yields

$$C(n) = (n-1)(c_0 - c_\infty) - 2\sum_{l=1}^{\min(n-1,m)}\Big(1 - \frac{l}{n}\Big)\delta_l.$$

Let $V_P := c_0 - c_\infty$ and $S(k) := \sum_{l=1}^{\min(k-1,m)}(1 - l/k)\,\delta_l$. Then

$$\Delta_C = C(n_1) + C(n_2) - C(n_1+n_2) = -V_P - 2\big(S(n_1) + S(n_2) - S(n)\big), \quad n = n_1 + n_2.$$

Since $|\delta_l| \le 2M$ and $(1 - l/k) \in [0, 1]$, we have $|S(k)| \le \sum_{l=1}^{m}|\delta_l| \le 2mM$ for all $k \ge 1$. Also $|V_P| = |c_0 - c_\infty| \le 2M$. Therefore

$$|\Delta_C| \le |V_P| + 2\big(|S(n_1)| + |S(n_2)| + |S(n)|\big) \le 2M + 2(3 \cdot 2mM) = 2M(1 + 6m) =: C_K.$$

On $\mathcal{E}_T$,

$$\big[\widehat{C}(s,t) + \widehat{C}(t{+}1,e) - \widehat{C}(s,e)\big] + \beta_T \;\geq\; -C_K - 3\lambda_T\sqrt{T} + \beta_T.$$

By Assumption 3.5,

$$\beta_T \;\geq\; 16M\sqrt{2(8m+5)T\log T} \;+\; 2M(1+6m) \;=\; 4\lambda_T\sqrt{T} + C_K,$$

so the RHS is at least $\lambda_T\sqrt{T} > 0$. Hence

$$\widehat{C}(s,e) \;<\; \widehat{C}(s,t) + \widehat{C}(t{+}1,e) + \beta_T.$$

Since $\mathcal{E}_T$ holds with probability $\geq 1 - T^{-1}$ and all bounds above are uniform in $[s,e]$ and t, the result holds simultaneously for all clean segments and all splits with that probability.

A.4 Proof of Theorem 3.11

Define the empirical penalized criterion

$$L(\boldsymbol{\tau}'_{K'}) := \sum_{k=1}^{K'+1} \widehat{C}(\tau'_{k-1}+1, \tau'_k) + \beta_T K'$$

and the corresponding population penalized criterion

$$L^\star(\boldsymbol{\tau}'_{K'}) := \sum_{k=1}^{K'+1} C(\tau'_{k-1}+1, \tau'_k) + \beta_T K'.$$

We work on the event $\mathcal{E}_T$, which holds with probability at least $1 - T^{-1}$.

Step 1: deviation bound for any fixed segmentation. Fix an arbitrary segmentation $\boldsymbol{\tau}'_{K'}$. For each $k \in \{1, \ldots, K'+1\}$, let

$$n_k := \tau'_k - \tau'_{k-1} \quad \text{so that} \quad \sum_{k=1}^{K'+1} n_k = T.$$

The segment $[\tau'_{k-1}+1, \tau'_k]$ has length n_k. On $\mathcal{E}_T$, the uniform deviation bound gives

$$\big|\widehat{C}(\tau'_{k-1}+1, \tau'_k) - C(\tau'_{k-1}+1, \tau'_k)\big| \leq \lambda_T\sqrt{n_k} \quad \text{for all } k.$$

Summing this over all segments, we obtain

$$\begin{aligned}\left|\sum_{k=1}^{K'+1} \widehat{C}(\tau'_{k-1}+1, \tau'_k) - \sum_{k=1}^{K'+1} C(\tau'_{k-1}+1, \tau'_k)\right| &\leq \sum_{k=1}^{K'+1} \big|\widehat{C}(\tau'_{k-1}+1, \tau'_k) - C(\tau'_{k-1}+1, \tau'_k)\big| \\ &\leq \lambda_T \sum_{k=1}^{K'+1} \sqrt{n_k}.\end{aligned}$$

By the Cauchy–Schwarz inequality,

$$\sum_{k=1}^{K'+1} \sqrt{n_k} \leq \sqrt{(K'+1)\sum_{k=1}^{K'+1} n_k} = \sqrt{(K'+1)\,T}.$$

Hence

$$\left|\sum_{k=1}^{K'+1} \widehat{C}(\tau'_{k-1}+1, \tau'_k) - \sum_{k=1}^{K'+1} C(\tau'_{k-1}+1, \tau'_k)\right| \leq \lambda_T\sqrt{(K'+1)\,T}.$$

Since $K'+1 \leq T$ for any segmentation (there can be at most $T-1$ change points), we have the simpler bound

$$\left|\sum_{k=1}^{K'+1} \widehat{C}(\tau'_{k-1}+1, \tau'_k) - \sum_{k=1}^{K'+1} C(\tau'_{k-1}+1, \tau'_k)\right| \leq \lambda_T T. \tag{7}$$

For the penalized criteria this implies

$$\left|L(\boldsymbol{\tau}'_{K'}) - L^\star(\boldsymbol{\tau}'_{K'})\right| = \left|\sum_{k=1}^{K'+1} \widehat{C}(\tau'_{k-1}+1, \tau'_k) - \sum_{k=1}^{K'+1} C(\tau'_{k-1}+1, \tau'_k)\right| \le \lambda_T T, \tag{8}$$

since the penalty term $\beta_T K'$ is identical in both L and $L^\star$.

Step 2: comparison between the empirical minimizer and a competitor. Let $\widehat{\boldsymbol{\tau}}_{\widehat{K}}$ be any minimizer of L over all segmentations. Fix an arbitrary competitor $\boldsymbol{\tau}'_{K'}$. We derive a chain of inequalities on $\mathcal{E}_T$.

First, apply (8) with $\boldsymbol{\tau}'_{K'} = \widehat{\boldsymbol{\tau}}_{\widehat{K}}$ to obtain

$$L^\star(\widehat{\boldsymbol{\tau}}_{\widehat{K}}) = L(\widehat{\boldsymbol{\tau}}_{\widehat{K}}) - \Big[\sum_{k=1}^{\widehat{K}+1} \widehat{C}(\widehat{\tau}_{k-1}+1, \widehat{\tau}_k) - \sum_{k=1}^{\widehat{K}+1} C(\widehat{\tau}_{k-1}+1, \widehat{\tau}_k)\Big] \le L(\widehat{\boldsymbol{\tau}}_{\widehat{K}}) + \lambda_T T. \tag{9}$$

Second, by the optimality of $\widehat{\boldsymbol{\tau}}_{\widehat{K}}$ for the empirical criterion,

$$L(\widehat{\boldsymbol{\tau}}_{\widehat{K}}) \le L(\boldsymbol{\tau}'_{K'}). \tag{10}$$

Third, apply (8) with $\boldsymbol{\tau}'_{K'}$ as given to get

$$L(\boldsymbol{\tau}'_{K'}) = L^\star(\boldsymbol{\tau}'_{K'}) + \Big[\sum_{k=1}^{K'+1} \widehat{C}(\tau'_{k-1}+1, \tau'_k) - \sum_{k=1}^{K'+1} C(\tau'_{k-1}+1, \tau'_k)\Big] \le L^\star(\boldsymbol{\tau}'_{K'}) + \lambda_T T. \tag{11}$$

Combining (9), (10), and (11), we obtain

$$\begin{aligned} L^\star(\widehat{\boldsymbol{\tau}}_{\widehat{K}}) &\le L(\widehat{\boldsymbol{\tau}}_{\widehat{K}}) + \lambda_T T \\ &\le L(\boldsymbol{\tau}'_{K'}) + \lambda_T T \\ &\le L^\star(\boldsymbol{\tau}'_{K'}) + 2\lambda_T T. \end{aligned}$$

Since this holds for an arbitrary competitor $\boldsymbol{\tau}'_{K'}$, we can take the infimum over all segmentations to get

$$L^\star(\widehat{\boldsymbol{\tau}}_{\widehat{K}}) \le \inf_{\boldsymbol{\tau}'_{K'}} L^\star(\boldsymbol{\tau}'_{K'}) + 2\lambda_T T.$$

Unwrapping the definition of $L^\star$, this inequality is exactly (1):

$$\sum_{k=1}^{\widehat{K}+1} C(\widehat{\tau}_{k-1}+1, \widehat{\tau}_k) + \beta_T \widehat{K} \;\le\; \inf_{\boldsymbol{\tau}'_{K'}} \Big\{ \sum_{k=1}^{K'+1} C(\tau'_{k-1}+1, \tau'_k) + \beta_T K' \Big\} + 2\lambda_T T.$$

We have proved that the inequality holds on the event $\mathcal{E}_T$, which has probability at least $1 - T^{-1}$ by Lemma 3.9. This completes the proof.

A.5 Additional Results for Theorem 3.12

Lemma A.2 (Signal strength on a mixed segment)**.** *Let $[s, e]$ contain exactly one true change-point τ_k with $s \le \tau_k < e$. Define*

$$n_1 := \tau_k - s + 1, \qquad n_2 := e - \tau_k, \qquad n := n_1 + n_2, \qquad \rho := \frac{n_1 n_2}{n}.$$

Under Assumptions 3.1–3.3,

$$C(s,e) - C(s,\tau_k) - C(\tau_k{+}1, e) \;\ge\; \rho\,\Delta_k^2 \;-\; \Big((4m{+}2)M + \frac{(2m^2{+}2m)M}{n}\Big). \tag{12}$$

If, in addition, Assumption 3.4 holds and the segment $[s, e]$ satisfies $n_1 \ge \ell_T/2$ and $n_2 \ge \ell_T/2$, then

$$C(s,e) - C(s,\tau_k) - C(\tau_k{+}1, e) \;\ge\; \frac{\Delta_\star^2}{4}\,\ell_T \;-\; \Big((4m{+}2)M + \frac{(2m^2{+}2m)M}{\ell_T}\Big). \tag{13}$$

Proof. We prove (12) and then deduce (13).

Part 1: Proof of (12). Using $C(u,v) = \mathbb{E}[\widehat{C}(u,v)]$ and expanding the quadratic terms, the diagonal pieces cancel, and we obtain

$$C(s,e)-C(s,\tau_k)-C(\tau_k{+}1,e) = \mathbb{E}\left[\left(\tfrac{1}{n_1}-\tfrac{1}{n}\right)\sum_{i,j=s}^{\tau_k} k(Y_i,Y_j) + \left(\tfrac{1}{n_2}-\tfrac{1}{n}\right)\sum_{i,j=\tau_k+1}^{e} k(Y_i,Y_j) - \tfrac{2}{n}\sum_{i=s}^{\tau_k}\sum_{j=\tau_k+1}^{e} k(Y_i,Y_j)\right].$$

Write $\mu_{P_k} := \mathbb{E}[\phi(Y) \mid Y \sim P_k] \in \mathcal{H}$ and recall $k(x,y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{H}}$. Introduce the population (independence) proxy by replacing $\mathbb{E}\, k(Y_i,Y_j)$ with $\langle \mu_{\mathrm{dist}(i)}, \mu_{\mathrm{dist}(j)}\rangle_{\mathcal{H}}$, where $\mathrm{dist}(i) \in \{k, k{+}1\}$ indicates the block of i. This yields the *population term*

$$\frac{n_1 n_2}{n}\left(\|\mu_{P_k}\|_{\mathcal{H}}^2 + \|\mu_{P_{k+1}}\|_{\mathcal{H}}^2 - 2\langle \mu_{P_k}, \mu_{P_{k+1}}\rangle_{\mathcal{H}}\right) = \rho\, \|\mu_{P_k} - \mu_{P_{k+1}}\|_{\mathcal{H}}^2 = \rho\, \Delta_k^2,$$

and a *remainder* R capturing the m-dependence corrections.

Let $\delta_{i,j} := \mathbb{E}[k(Y_i,Y_j)] - \langle \mu_{\mathrm{dist}(i)}, \mu_{\mathrm{dist}(j)}\rangle_{\mathcal{H}}$. Then $\delta_{i,j} = 0$ whenever $|i-j| > m$ by m-dependence (Assumption 3.1); moreover, by boundedness (Assumption 3.2), $|\delta_{i,j}| \le 2M$. Writing

$$R = \frac{n_2}{n n_1} E_1 + \frac{n_1}{n n_2} E_2 - \frac{2}{n} E_{12},$$

where $E_1 := \sum_{i,j=s}^{\tau_k} \delta_{i,j}$, $E_2 := \sum_{i,j=\tau_k+1}^{e} \delta_{i,j}$, and $E_{12} := \sum_{i=s}^{\tau_k}\sum_{j=\tau_k+1}^{e} \delta_{i,j}$, we bound each piece by counting ordered pairs with $|i-j| \le m$:

$$\begin{aligned}
|E_1| &\le \big(\text{at most } n_1(2m{+}1) \text{ pairs}\big)\cdot 2M = n_1(2m{+}1)\,2M,\\
|E_2| &\le n_2(2m{+}1)\,2M,\\
|E_{12}| &\le \Big(\sum_{d=1}^{m} d\Big)\cdot 2M = \frac{m(m{+}1)}{2}\cdot 2M = m(m{+}1)\,M,
\end{aligned}$$

where the last line counts the cross-boundary pairs with offsets $d = 1, \dots, m$ once (note the algebra above contributes $-\frac{2}{n}E_{12}$, so only left-to-right ordered pairs appear). Consequently,

$$\begin{aligned}
|R| &\le \frac{n_2}{n n_1}\, n_1(2m{+}1)\,2M + \frac{n_1}{n n_2}\, n_2(2m{+}1)\,2M + \frac{2}{n}\, m(m{+}1)\,M\\
&= \frac{n_1+n_2}{n}\,(4m{+}2)M + \frac{2m^2+2m}{n} M = (4m{+}2)M + \frac{(2m^2{+}2m)M}{n}.
\end{aligned}$$

Since $C(s,e) - C(s,\tau_k) - C(\tau_k{+}1,e) = \rho\,\Delta_k^2 + R$, we obtain (12) from $R \ge -|R|$.

Part 2: Proof of (13). By Assumption 3.3, $\Delta_k^2 \ge \Delta_\star^2$. Under $n_1, n_2 \ge \ell_T/2$, the function $\rho = \frac{n_1 n_2}{n_1+n_2}$ is minimized at $n_1 = n_2 = \ell_T/2$, hence

$$\rho \;\ge\; \frac{(\ell_T/2)^2}{\ell_T} \;=\; \frac{\ell_T}{4}.$$

Also $n = n_1 + n_2 \ge \ell_T$, so

$$-\Big((4m{+}2)M + \frac{(2m^2{+}2m)M}{n}\Big) \;\ge\; -\Big((4m{+}2)M + \frac{(2m^2{+}2m)M}{\ell_T}\Big).$$

Combining with (12) yields (13). □

Lemma A.3 (Population gap on intervals covering two consecutive changes)**.** *Let Assumptions 3.3–3.8 hold. Then there exist constants $\delta_0 > 0$ and T_1 such that for all $T \ge T_1$, for every pair of consecutive true change points (τ_k, τ_{k+1}) and every interval $[s,e]$ with*

$$s \le \tau_k < \tau_{k+1} \le e \qquad \textit{and} \qquad e - s + 1 \ge 2\delta_T,$$

there exists a split point $t^\star \in \mathcal{T}_{k,s,e}$ such that

$$C(s,e) - C(s,t^\star) - C(t^\star{+}1,e) \;\ge\; \beta_T + 4\lambda_T\sqrt{T} + \delta_0\beta_T.$$

Proof. Fix T and an interval $[s, e]$ satisfying

$$s \le \tau_k < \tau_{k+1} \le e, \qquad e - s + 1 \ge 2\delta_T.$$

By Assumption 3.8, there exists a split point $t^\star \in \mathcal{T}_{k,s,e}$ such that

$$C(s, e) - C(s, t^\star) - C(t^\star + 1, e) \;\ge\; c_0\, g_k\, \Delta_\star^2 - C_m,$$

where

$$g_k := \tau_{k+1} - \tau_k.$$

Since g_k is a block length, we have

$$g_k \ge \ell_T.$$

By Assumption 3.4, for all sufficiently large T,

$$\ell_T \ge 3\delta_T,$$

and therefore

$$g_k \ge 3\delta_T.$$

Hence

$$C(s, e) - C(s, t^\star) - C(t^\star + 1, e) \;\ge\; 3c_0\, \delta_T\, \Delta_\star^2 - C_m.$$

Now Assumption 3.7 gives

$$\delta_T \Delta_\star^2 \;\ge\; C_0\big(\beta_T + \lambda_T\sqrt{T} + \overline{B}_T\big),$$

so

$$C(s, e) - C(s, t^\star) - C(t^\star + 1, e) \;\ge\; 3c_0 C_0\big(\beta_T + \lambda_T\sqrt{T} + \overline{B}_T\big) - C_m.$$

Dropping the nonnegative terms involving $\lambda_T\sqrt{T}$ and $\overline{B}_T$ yields

$$C(s, e) - C(s, t^\star) - C(t^\star + 1, e) \;\ge\; 3c_0 C_0\, \beta_T - C_m.$$

Choose any constant δ_0 such that

$$0 < \delta_0 < 3c_0 C_0 - 2.$$

This is possible because $3c_0 C_0 > 2$ by assumption. Since $\beta_T \to \infty$ by Assumption 3.5, there exists T_1' such that for all $T \ge T_1'$,

$$C_m \le (3c_0 C_0 - 2 - \delta_0)\beta_T.$$

Therefore, for all $T \ge T_1'$,

$$3c_0 C_0\, \beta_T - C_m \;\ge\; (2 + \delta_0)\beta_T.$$

It follows that for all $T \ge T_1'$,

$$C(s, e) - C(s, t^\star) - C(t^\star + 1, e) \;\ge\; (2 + \delta_0)\beta_T.$$

Finally, Assumption 3.5 implies that for all sufficiently large T,

$$4\lambda_T\sqrt{T} \le \beta_T.$$

Hence

$$\beta_T + 4\lambda_T\sqrt{T} + \delta_0\beta_T \;\le\; (2 + \delta_0)\beta_T.$$

Combining the last two displays, we obtain

$$C(s, e) - C(s, t^\star) - C(t^\star + 1, e) \;\ge\; \beta_T + 4\lambda_T\sqrt{T} + \delta_0\beta_T.$$

This proves the claim. □

Lemma A.4 (No overfull estimated segments)**.** *Let Assumptions 3.1–3.8 hold. Then, with probability at least $1 - T^{-1}$, no segment of an optimal penalized partition in $\mathcal{P}_T(\delta_T)$ contains two true change points.*

Proof. Let $\widehat{\boldsymbol{\tau}}_{\widehat{K}} \in \mathcal{P}_T(\delta_T)$ be any minimizer of the penalized cost over $\mathcal{P}_T(\delta_T)$, i.e.,

$$\widehat{\boldsymbol{\tau}}_{\widehat{K}} \in \arg\min_{\boldsymbol{\tau}'_{K'} \in \mathcal{P}_T(\delta_T)} L(\boldsymbol{\tau}'_{K'}).$$

Work on the event

$$\mathcal{E}_T := \Big\{\forall\, 1 \le s \le e \le T :\ |\widehat{C}(s,e) - C(s,e)| \le \lambda_T \sqrt{e-s+1}\Big\},$$

which satisfies $\Pr(\mathcal{E}_T) \ge 1 - T^{-1}$ by Lemma 3.9.

Suppose, for contradiction, that some estimated segment $[s,e]$ induced by $\widehat{\boldsymbol{\tau}}_{\widehat{K}}$ contains two consecutive true change points $\tau_k < \tau_{k+1}$ with

$$s \le \tau_k < \tau_{k+1} \le e.$$

Let $\delta_0 > 0$ be the constant from Lemma A.3. Since $[s,e]$ contains two consecutive true change points $\tau_k < \tau_{k+1}$, we have

$$e - s + 1 \ge \tau_{k+1} - \tau_k + 1 \ge \ell_T + 1 \ge 3\delta_T + 1 > 2\delta_T$$

for all sufficiently large T, by Assumption 3.4. Hence the interval $[s,e]$ satisfies the length requirement of Lemma A.3. By Lemma A.3, there exists a split point $t^\star \in \mathcal{T}_{k,s,e}$ such that

$$G := C(s,e) - C(s,t^\star) - C(t^\star+1,e) \ge \beta_T + 4\lambda_T\sqrt{T} + \delta_0\beta_T.$$

Since $t^\star \in \mathcal{T}_{k,s,e}$, both subsegments $[s,t^\star]$ and $[t^\star+1,e]$ have length at least δ_T. Hence inserting a split at $t^\star$ yields another admissible partition.

Now define

$$\widehat{G} := \widehat{C}(s,e) - \widehat{C}(s,t^\star) - \widehat{C}(t^\star+1,e).$$

On the event $\mathcal{E}_T$,

$$\begin{aligned}\widehat{G} &\ge G - \big|\widehat{C}(s,e) - C(s,e)\big| - \big|\widehat{C}(s,t^\star) - C(s,t^\star)\big| - \big|\widehat{C}(t^\star+1,e) - C(t^\star+1,e)\big| \\ &\ge G - \lambda_T\Big(\sqrt{e-s+1} + \sqrt{t^\star - s + 1} + \sqrt{e - t^\star}\Big).\end{aligned}$$

Using

$$\sqrt{e-s+1} \le \sqrt{t^\star - s + 1} + \sqrt{e - t^\star},$$

we obtain

$$\widehat{G} \ge G - 2\lambda_T\Big(\sqrt{t^\star - s + 1} + \sqrt{e - t^\star}\Big).$$

Since each segment length is at most T,

$$\sqrt{t^\star - s + 1} \le \sqrt{T}, \qquad \sqrt{e - t^\star} \le \sqrt{T},$$

and therefore

$$\widehat{G} \ge G - 4\lambda_T\sqrt{T}.$$

Combining this with the lower bound on G gives

$$\widehat{G} \ge \beta_T + \delta_0\beta_T.$$

Refine the estimated partition by inserting a split at $t^\star$ and leaving all other segments unchanged. This decreases the empirical fit term by $\widehat{G}$ and increases the penalty by exactly β_T. Thus the net change in penalized cost is

$$\Delta L = -\widehat{G} + \beta_T \le -(\beta_T + \delta_0\beta_T) + \beta_T = -\delta_0\beta_T < 0.$$

This contradicts the optimality of $\widehat{\boldsymbol{\tau}}_{\widehat{K}}$.

Hence, on the event $\mathcal{E}_T$, no estimated segment can contain two true change points. Since $\Pr(\mathcal{E}_T) \ge 1 - T^{-1}$, the result follows. □

Corollary A.5 (No estimated segment contains ≥ 2 true changes)**.** *Let Assumptions 3.1–3.8 hold. With probability at least $1 - T^{-1}$, every estimated segment of an optimal penalized partition contains at most one true changepoint.*

Proof. If an estimated segment contained ≥ 2 true changepoints, it would contain some adjacent pair (τ_k, τ_{k+1}). Apply Lemma A.3 within that segment to obtain a split $t^\star$ such that inserting $t^\star$ strictly decreases the penalized cost, exactly as in the proof of Lemma A.4. This contradicts optimality. The high-probability event is the same as in Lemma A.4. □

Lemma A.6 (Strict improvement of a mixed segment)**.** *Let Assumptions 3.1–3.3, 3.6, and 3.7 hold. Let $\mathcal{E}_T$ be the event from Lemma 3.9,*

$$\mathcal{E}_T := \Big\{\forall\, 1 \leq s \leq e \leq T :\ |\widehat{C}(s,e) - C(s,e)| \leq \lambda_T \sqrt{e-s+1}\Big\}.$$

Let $\boldsymbol{\tau}$ be an admissible partition, meaning that every segment has length at least δ_T, and suppose that one of its segments

$$E = [s, e]$$

contains exactly one true change point τ_k with

$$s \leq \tau_k < e.$$

Write

$$n_1 := \tau_k - s + 1, \qquad n_2 := e - \tau_k, \qquad n := n_1 + n_2.$$

Assume moreover that

$$n_1 \geq \delta_T \qquad \textit{and} \qquad n_2 \geq \delta_T,$$

so that splitting E at τ_k remains admissible. Then, on the event $\mathcal{E}_T$, there exists an admissible partition $\boldsymbol{\tau}'$ obtained by inserting τ_k into $\boldsymbol{\tau}$ such that

$$L(\boldsymbol{\tau}') < L(\boldsymbol{\tau})$$

for all sufficiently large T.

Proof. Work on the event $\mathcal{E}_T$ throughout.

Let $\boldsymbol{\tau}'$ be the partition obtained from $\boldsymbol{\tau}$ by splitting the segment $E = [s, e]$ at τ_k, that is, by replacing $[s, e]$ with $[s, \tau_k]$ and $[\tau_k + 1, e]$ and leaving all other segments unchanged. Because

$$n_1 = \tau_k - s + 1 \geq \delta_T \qquad \text{and} \qquad n_2 = e - \tau_k \geq \delta_T,$$

the partition $\boldsymbol{\tau}'$ is admissible under Assumption 3.6. If $\boldsymbol{\tau}$ has K change points, then $\boldsymbol{\tau}'$ has $K + 1$, so the penalty increases by β_T.

Define

$$\Delta L := L(\boldsymbol{\tau}') - L(\boldsymbol{\tau}).$$

Since only the segment E is modified, we have

$$\Delta L = \beta_T + \widehat{C}(s, \tau_k) + \widehat{C}(\tau_k + 1, e) - \widehat{C}(s, e).$$

Decompose this as

$$\Delta L = \beta_T + \Delta C_{\text{pop}} + \Delta_{\text{noise}},$$

where

$$\Delta C_{\text{pop}} := C(s, \tau_k) + C(\tau_k + 1, e) - C(s, e),$$

and

$$\Delta_{\text{noise}} := \big(\widehat{C}(s, \tau_k) - C(s, \tau_k)\big) + \big(\widehat{C}(\tau_k + 1, e) - C(\tau_k + 1, e)\big) - \big(\widehat{C}(s, e) - C(s, e)\big).$$

We first bound the population term. Since $[s, e]$ contains exactly one true change point τ_k, Lemma A.2, inequality (12), yields

$$C(s,e) - C(s,\tau_k) - C(\tau_k+1,e) \geq \rho\, \Delta_k^2 - \Big((4m+2)M + \frac{(2m^2+2m)M}{n}\Big),$$

where

$$\rho := \frac{n_1 n_2}{n}.$$

Rearranging and using Assumption 3.3, which gives $\Delta_k^2 \geq \Delta_\star^2$, we obtain

$$\Delta C_{\text{pop}} \leq -\rho\,\Delta_\star^2 + \Big((4m+2)M + \frac{(2m^2+2m)M}{n}\Big).$$

Without loss of generality, assume $n_1 \leq n_2$. Then

$$\rho = \frac{n_1 n_2}{n_1+n_2} = n_1 \frac{n_2}{n_1+n_2} \geq \frac{n_1}{2} \geq \frac{\delta_T}{2}.$$

Also, since E is a segment of an admissible partition, its length satisfies $n = e - s + 1 \geq \delta_T$. Therefore

$$(4m+2)M + \frac{(2m^2+2m)M}{n} \leq (4m+2)M + \frac{(2m^2+2m)M}{\delta_T} = \overline{B}_T.$$

Hence

$$\Delta C_{\text{pop}} \leq -\frac{\delta_T}{2}\Delta_\star^2 + \overline{B}_T.$$

Next we control the empirical fluctuation term. On the event $\mathcal{E}_T$,

$$|\widehat{C}(s,\tau_k) - C(s,\tau_k)| \leq \lambda_T\sqrt{n_1},$$

$$|\widehat{C}(\tau_k+1,e) - C(\tau_k+1,e)| \leq \lambda_T\sqrt{n_2},$$

and

$$|\widehat{C}(s,e) - C(s,e)| \leq \lambda_T\sqrt{n}.$$

Thus

$$|\Delta_{\text{noise}}| \leq \lambda_T(\sqrt{n_1} + \sqrt{n_2} + \sqrt{n}).$$

Using

$$\sqrt{n_1} + \sqrt{n_2} \leq \sqrt{2(n_1+n_2)} = \sqrt{2n},$$

we get

$$\sqrt{n_1} + \sqrt{n_2} + \sqrt{n} \leq (\sqrt{2}+1)\sqrt{n} < 3\sqrt{n} \leq 3\sqrt{T},$$

and therefore

$$|\Delta_{\text{noise}}| \leq 3\lambda_T\sqrt{T}.$$

Combining the two bounds,

$$\Delta L \leq \beta_T - \frac{\delta_T}{2}\Delta_\star^2 + \overline{B}_T + 3\lambda_T\sqrt{T}.$$

Now apply Assumption 3.7:

$$\delta_T\Delta_\star^2 \geq C_0\big(\beta_T + \lambda_T\sqrt{T} + \overline{B}_T\big).$$

Dividing by 2 gives

$$\frac{\delta_T}{2}\Delta_\star^2 \geq \frac{C_0}{2}\big(\beta_T + \lambda_T\sqrt{T} + \overline{B}_T\big).$$

Since $C_0 > 6$, we have $\frac{C_0}{2} > 3$, and hence for all sufficiently large T,

$$\frac{\delta_T}{2}\Delta_\star^2 > \beta_T + 3\lambda_T\sqrt{T} + \overline{B}_T.$$

Therefore, for all sufficiently large T,

$$\Delta L < 0.$$

Thus

$$L(\boldsymbol{\tau}') < L(\boldsymbol{\tau}),$$

which proves that inserting the true change point τ_k strictly improves the penalized objective. □

A.6 Proof of Theorem 3.12

Recall the high-probability event $\mathcal{E}_T$ from Lemma 3.9:

$$\mathcal{E}_T := \Big\{\forall\, 1 \le s \le e \le T :\ |\widehat{C}(s,e) - C(s,e)| \le \lambda_T \sqrt{e-s+1}\Big\},$$

where $\lambda_T = 4\sqrt{2}\, M \sqrt{(8m+5)\log T}$, and $\Pr(\mathcal{E}_T) \ge 1 - T^{-1}$ for all $T \ge 3$.

Let $\mathcal{N}_T$ be the high-probability event from Lemma A.4. That lemma states that, under Assumptions 3.1–3.8, for all T large enough,

$$\Pr(\mathcal{N}_T) \ge 1 - T^{-1},$$

and on $\mathcal{N}_T$, no segment of an optimal penalized partition in $\mathcal{P}_T(\delta_T)$ contains two true change points.

Define

$$\Omega_T := \mathcal{E}_T \cap \mathcal{N}_T.$$

Then $\Pr(\Omega_T) \ge 1 - 2T^{-1} \to 1$ as $T \to \infty$.

We will show that on Ω_T and for all sufficiently large T,

$$\forall\, 1 \le k \le K :\ \min_{1 \le j \le \widehat{K}} |\widehat{\tau}_j - \tau_k| \le \delta_T.$$

Since $K \ge 1$, this immediately implies $\widehat{K} \ge 1$, and hence yields (2).

So fix T large and suppose Ω_T holds. Let $\widehat{\boldsymbol{\tau}}_{\widehat{K}} \in \mathcal{P}_T(\delta_T)$ denote the optimal penalized partition.

We now argue by contradiction. Suppose that there exists at least one true change point that is not localized within δ_T. Equivalently, with the convention $\min_{\varnothing} := +\infty$, assume there exists

$$k^\star \in \{1, \dots, K\} \quad \text{such that} \quad \min_{1 \le j \le \widehat{K}} |\widehat{\tau}_j - \tau_{k^\star}| > \delta_T.$$

Fix such an index $k^\star$.

Let E be the segment of the estimated partition $\widehat{\boldsymbol{\tau}}_{\widehat{K}}$ that contains $\tau_{k^\star}$. Concretely, there exists $r \in \{1, \dots, \widehat{K}+1\}$ such that

$$E = [s, e] = [\widehat{\tau}_{r-1} + 1, \widehat{\tau}_r],$$

with the convention $\widehat{\tau}_0 = 0$, $\widehat{\tau}_{\widehat{K}+1} = T$, and

$$\widehat{\tau}_{r-1} < \tau_{k^\star} \le \widehat{\tau}_r.$$

By Lemma A.4, on $\mathcal{N}_T$ no optimal penalized segment contains two true change points. Since E contains $\tau_{k^\star}$, it must therefore contain *exactly one* true change point, namely $\tau_{k^\star}$. Thus E is a mixed segment with exactly one true change.

We next verify that splitting E at $\tau_{k^\star}$ is admissible, namely that both sides of $\tau_{k^\star}$ inside E have length at least δ_T. Define

$$n_1 := \tau_{k^\star} - s + 1, \qquad n_2 := e - \tau_{k^\star}.$$

For the left side, there are two possibilities.

First, if $r \ge 2$, then $\widehat{\tau}_{r-1}$ is an estimated change point. Since $\tau_{k^\star}$ is at distance strictly greater than δ_T from every estimated change point, we have

$$n_1 = \tau_{k^\star} - \widehat{\tau}_{r-1} > \delta_T.$$

Second, if $r = 1$, then $E = [1, \widehat{\tau}_1]$ when $\widehat{K} \ge 1$, and $E = [1, T]$ when $\widehat{K} = 0$. In either case, because E contains exactly one true change point, it must be the first true change point, so $k^\star = 1$. Indeed, if $k^\star > 1$, then the previous true change point $\tau_{k^\star - 1}$ would also lie in E, contradicting the fact that E contains exactly one true change. Therefore

$$n_1 = \tau_1 - 1 + 1 = \tau_1 \ge \ell_T \ge 3\delta_T > \delta_T,$$

for all sufficiently large T, by Assumption 3.4.

For the right side, again there are two possibilities.

First, if $r \le \widehat{K}$, then $\widehat{\tau}_r$ is an estimated change point. Hence

$$n_2 = \widehat{\tau}_r - \tau_{k^\star} > \delta_T,$$

because $\tau_{k^\star}$ is at distance strictly greater than δ_T from every estimated change point.

Second, if $r = \widehat{K} + 1$, then $E = [\widehat{\tau}_{\widehat{K}} + 1, T]$ when $\widehat{K} \ge 1$, and again $E = [1, T]$ when $\widehat{K} = 0$. Since E contains exactly one true change point, it must be the last true change point, so $k^\star = K$. Indeed, if $k^\star < K$, then the next true change point $\tau_{k^\star+1}$ would also lie in E, contradicting the fact that E contains exactly one true change. Therefore

$$n_2 = T - \tau_K \ge \ell_T \ge 3\delta_T > \delta_T,$$

for all sufficiently large T, again by Assumption 3.4.

Thus, for all sufficiently large T,

$$n_1 \ge \delta_T, \qquad n_2 \ge \delta_T. \tag{14}$$

Note also that E itself is admissible by construction, since $\widehat{\boldsymbol{\tau}}_{\widehat{K}} \in \mathcal{P}_T(\delta_T)$ implies

$$e - s + 1 = \widehat{\tau}_r - \widehat{\tau}_{r-1} \ge \delta_T.$$

Therefore we are exactly in the setting of Lemma A.6: the partition $\boldsymbol{\tau} := \widehat{\boldsymbol{\tau}}_{\widehat{K}}$ belongs to $\mathcal{P}_T(\delta_T)$, the segment $E = [s, e]$ is one of its segments and contains exactly one true change point $\tau_{k^\star}$, and the split at $\tau_{k^\star}$ is admissible by (14).

Lemma A.6 states that on the event $\mathcal{E}_T$ and under Assumptions 3.1–3.3, 3.6, and 3.7, for all sufficiently large T, there exists a new admissible partition $\boldsymbol{\tau}'$ obtained from $\boldsymbol{\tau}$ by inserting the true change point $\tau_{k^\star}$ into the segment E such that

$$L(\boldsymbol{\tau}') < L(\boldsymbol{\tau}).$$

Since we are on $\Omega_T \subseteq \mathcal{E}_T$, it follows that for all sufficiently large T, when we take $\boldsymbol{\tau} = \widehat{\boldsymbol{\tau}}_{\widehat{K}}$, there exists an admissible partition $\boldsymbol{\tau}'$ such that

$$L(\boldsymbol{\tau}') < L(\widehat{\boldsymbol{\tau}}_{\widehat{K}}).$$

But this contradicts the definition of $\widehat{\boldsymbol{\tau}}_{\widehat{K}}$ as an optimal minimizer of $L(\cdot)$ over $\mathcal{P}_T(\delta_T)$.

Therefore, our assumption was false. Hence, on Ω_T and for all sufficiently large T,

$$\forall\, 1 \le k \le K : \ \min_{1 \le j \le \widehat{K}} |\widehat{\tau}_j - \tau_k| \le \delta_T.$$

Since $\Pr(\Omega_T) \to 1$ as $T \to \infty$, we obtain (2).

Finally, the $O_p(\delta_T)$ bound on the maximal localization error follows directly from (2): for any $\varepsilon > 0$ there exists T_0 such that for all $T \ge T_0$,

$$\Pr\Big(\max_{1 \le k \le K} \min_{1 \le j \le \widehat{K}} |\widehat{\tau}_j - \tau_k| \le \delta_T \Big) \ge 1 - \varepsilon,$$

which is exactly

$$\max_{1 \le k \le K} \min_{1 \le j \le \widehat{K}} |\widehat{\tau}_j - \tau_k| = O_p(\delta_T).$$

This completes the proof.

B Computational complexity of KCPD

The computational complexity of kernel change point detection (KCPD) combined with the PELT algorithm is well understood (Arlot et al., 2019). Let n denote the number of observations and c the cost of evaluating the kernel function. In the exact kernel setting, forming the $n \times n$ Gram matrix requires $O(n^2 c)$ time and $O(n^2)$ memory. Given the Gram matrix, segment costs can be evaluated in $O(1)$ time via cumulative sums, and PELT achieves linear expected time under its standard pruning assumptions (Killick et al., 2012) (with a quadratic worst case). Consequently, in the exact-kernel setting the overall time and memory are typically dominated by Gram-matrix precomputation, i.e., $O(n^2 c)$ time and $O(n^2)$ memory.

Table 3: Performance of Baselines and EMBED-KCPD in Choi's Dataset. The bolded P_k or WD values denote the best performance for each dataset comparing EMBED-KCPD with all baselines. x_y denotes mean x with standard deviation y. $*$ marks values reported in original papers.

Methods	**3-5**		**6-8**		**9-11**		**3-11**	
	$P_k \downarrow$	WD $\downarrow$	$P_k \downarrow$	WD $\downarrow$	$P_k \downarrow$	WD $\downarrow$	$P_k \downarrow$	WD $\downarrow$
Unsupervised Methods								
EMBED-KCPD								
sBERT	$5.4_{5.1}$	$5.4_{5.1}$	$6.7_{5.4}$	$6.7_{5.5}$	$7.6_{6.6}$	$7.6_{6.6}$	$9.2_{7.0}$	$9.5_{7.1}$
MPNet	$4.4_{5.1}$	$4.4_{5.1}$	$5.1_{5.2}$	$5.1_{5.2}$	$6.3_{6.6}$	$6.3_{6.6}$	$7.7_{6.1}$	$8.0_{6.3}$
text-embedding-3-small	$3.9_{4.4}$	$3.9_{4.4}$	$4.6_{5.3}$	$4.7_{5.4}$	$5.6_{5.2}$	$7.3_{6.4}$	$7.6_{6.5}$	$5.3_{5.3}$
RoBERTa	$4.3_{5.0}$	$4.3_{5.0}$	$4.9_{5.0}$	$5.0_{5.0}$	$5.7_{5.5}$	$5.7_{5.5}$	$8.0_{6.2}$	$8.3_{6.3}$
Baselines								
Coherence	4.4^*	6.2^*	3.1^*	3.3^*	2.5^*	2.6^*	4.0^*	4.4^*
GraphSeg	5.6^*	8.7^*	7.2^*	9.4^*	6.6^*	9.6^*	7.2^*	9.0^*
TextTiling	44^*	–	43^*	–	48^*	–	46^*	–
TextTiling (MPNet)	$44.6_{5.6}$	$86.3_{9.6}$	$37.6_{6.4}$	$76.7_{10.1}$	$31.1_{5.4}$	$70.1_{8.8}$	$31.7_{6.6}$	$71.5_{9.6}$
TextTiling (sBERT)	$50.0_{3.5}$	$96.9_{3.6}$	$45.3_{5.1}$	$91.7_{4.6}$	$40.3_{4.3}$	$86.6_{5.4}$	$41.2_{5.9}$	$86.8_{6.6}$
Choi (Choi, 2000)	12.0^*	–	9.0^*	–	9.0^*	–	12.0^*	–
Brants et al. (2002)	7.4^*	–	8.0^*	–	6.8^*	–	19.7^*	–
Fragkou et al. (2004)	5.5^*	–	3.0^*	–	1.3^*	–	7.0^*	–
Misra et al. (2009)	23.0^*	–	15.8^*	–	14.4^*	–	16.1^*	–

C ADDITIONAL EXPERIMENTAL RESULTS

C.1 RESULTS ON CHOI'S DATASET

Table 3 reports performance on the synthetic Choi benchmark. EMBED-KCPD exhibits stable and consistent performance across embeddings. While Coherence achieves the best scores on the 3–11 and 9–11 groups, EMBED-KCPD delivers competitive results on a dataset that is widely used in the literature, despite being highly artificial relative to real-world text segmentation tasks.

C.2 ABLATIONS

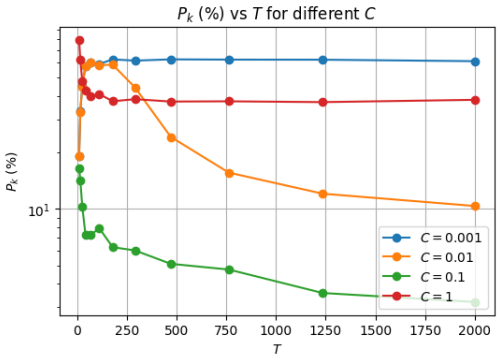


Figure 2: P_k error (%) versus sequence length T for EMBED-KCPD applied to synthetically generated short-range dependent text data with GPT-4.1, $m = 20$, for multiple values of C and sBERT embeddings.

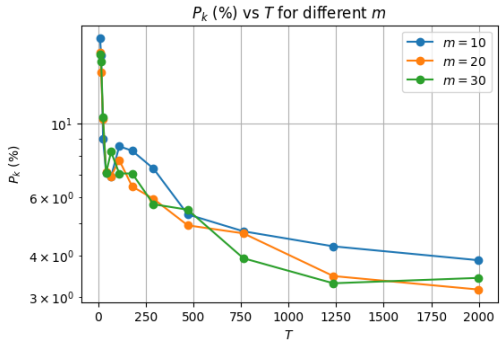


Figure 3: P_k error (%) versus sequence length T for EMBED-KCPD applied to synthetically generated short-range dependent text data with GPT-4.1, $C = 0.1$, for multiple values of m (number of sentences in LLM generation) and sBERT embeddings.

Figures 2 and 3 indicate the effect of varying C and m on P_k.

D Details for Text Segmentation experiments

D.1 Statistics of Dataset

Here is the summary of all datasets we used in the experiments. Table 4 presents the summary statistics for each dataset: total number of documents, number of segments per document, number of sentences per segment.

Table 4: Statistics of Datasets in Our Experiments.

Dataset	Documents	Segments per Document	Sentences per Segment
Choi (3-5)	100	10	4.0
Choi (6-8)	100	10	7.0
Choi (9-11)	100	10	9.9
Choi (3-11)	400	10	7.0
Wiki-300	300	7.6	26.0
Wiki-50	50	8.2	7.5
Elements	118	7.7	2.9
arXiv	50	9.5	7.1

D.2 Implementation details

EMBED-KCPD is implemented with the ruptures library (Truong et al., 2020), using its kernel-based change-point implementation. We use ruptures' median heuristic to set the bandwidth for the RBF Kernel.

We compute text-embedding-3-small sentence representations using the OpenAI API. For the LLM-based experiment, we use GPT-4.1 via the same API; the total API cost for running all experiments is below $20. All other embedding backbones are computed locally with the sentence-transformers library using the corresponding pretrained models.

For all baseline methods, we use the fine tuned hyperparameters from the original papers or from widely used public implementations.

All code and implementation is available as supplementary materials.

D.3 Optimal C via Elbow Method

For each dataset, we randomly sample 6 documents and, for each document, run EMBED-KCPD over a small logarithmically spaced C in the range $[10^{-2}, 10^{0}]$. The elbow point of the curve relating the number of detected change points to C is selected per document, and the final C is set to the average of these six values (see Fig. 4 for 6 documents from Wiki-300 datasets). Across datasets, the resulting elbow locations are highly consistent (see Figures 5-7). Based on the results, we chose $C = 0.06$ for all datasets.

D.4 Sensitivity of C on P_k and WD

To validate the robustness of our method with respect to C around the identified sweet spots, $C = 0.088$ for the KCPD kernel and $C = 0.06$ for RBF kernel, we conduct a sensitivity analysis on arXiv and Elements datasets. As shown in Figure 8, we vary C within the range $[0.08, 0.10]$ for KCPD and $[0.05, 0.07]$ for RBF. Across these intervals, both the P_k and WD metrics remain stable, indicating that performance is not sensitive to small perturbations of C near the optimal region.

D.5 Finite-range dependent Data Generation

Sentences are generated sequentially using GPT-4.1 with the fixed prompt: *Give me one more sentence to naturally continue the text specific on [Topic]. Do not add any preamble just answer with one sentence. [Input Sentences].*

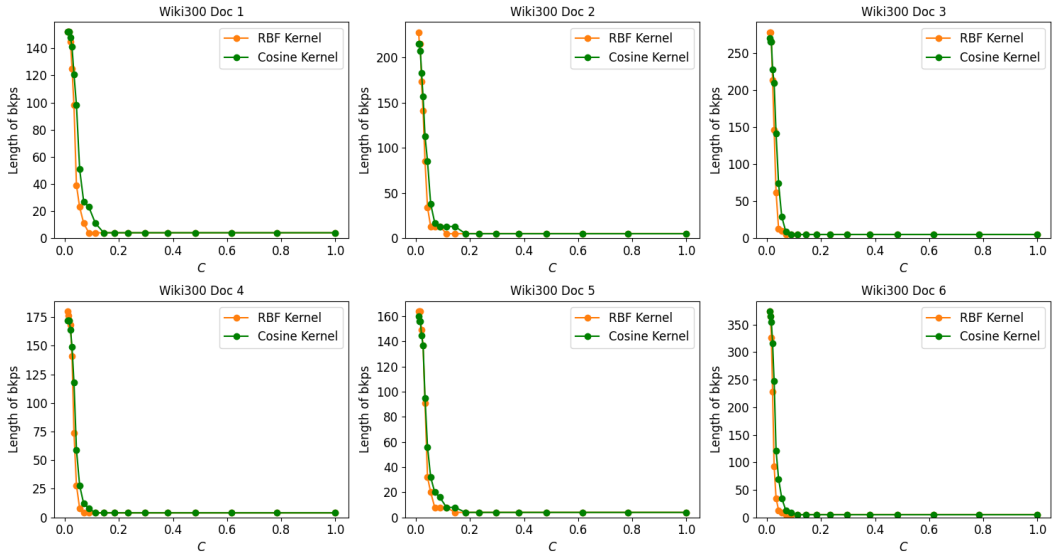


Figure 4: Sensitivity of the number of detected segments to the hyperparameter C on Wiki-300.

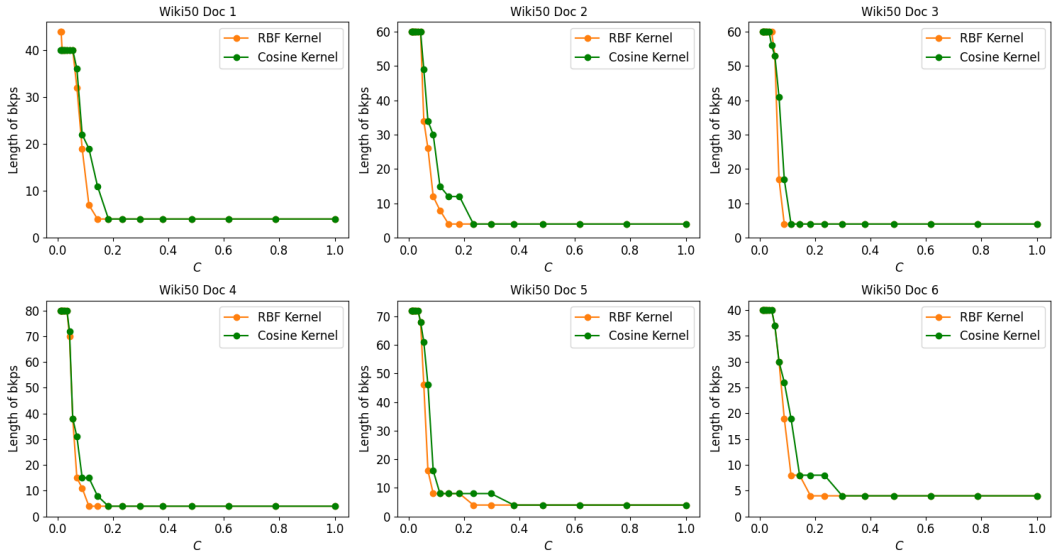


Figure 5: Sensitivity of the number of detected segments to the hyperparameter C on Wiki-50.

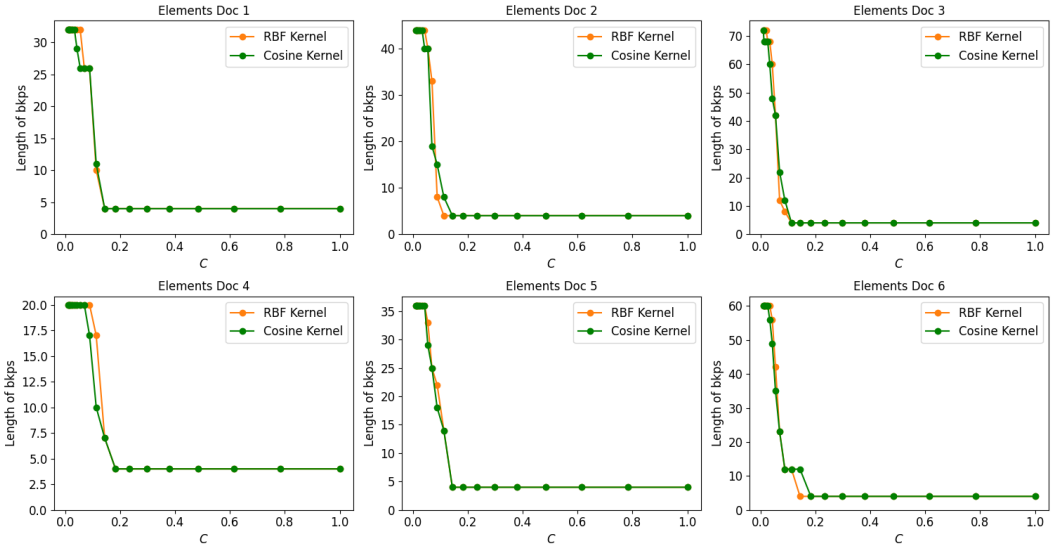


Figure 6: Sensitivity of the number of detected segments to the hyperparameter C on Elements.

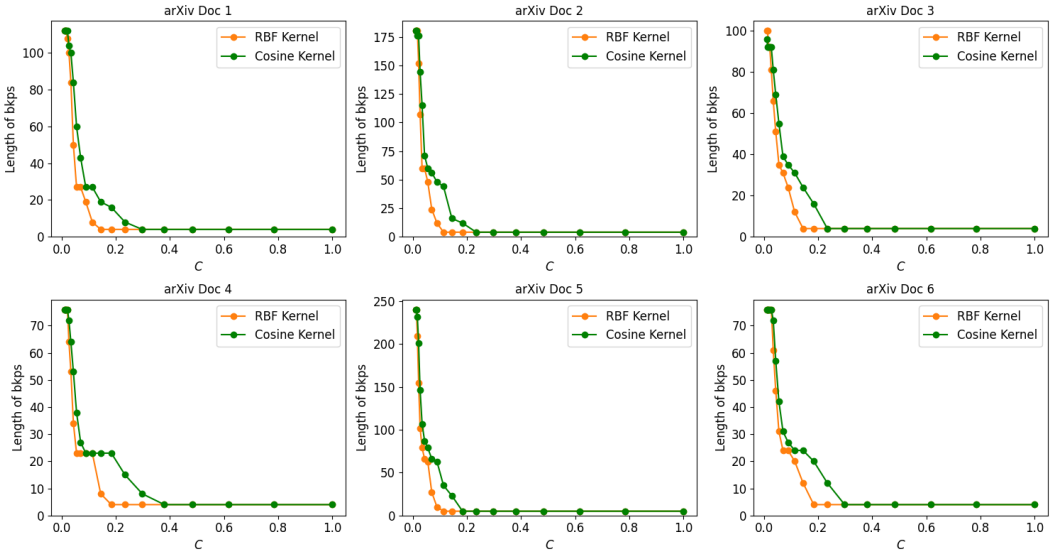


Figure 7: Sensitivity of the number of detected segments to the hyperparameter C on arXiv.

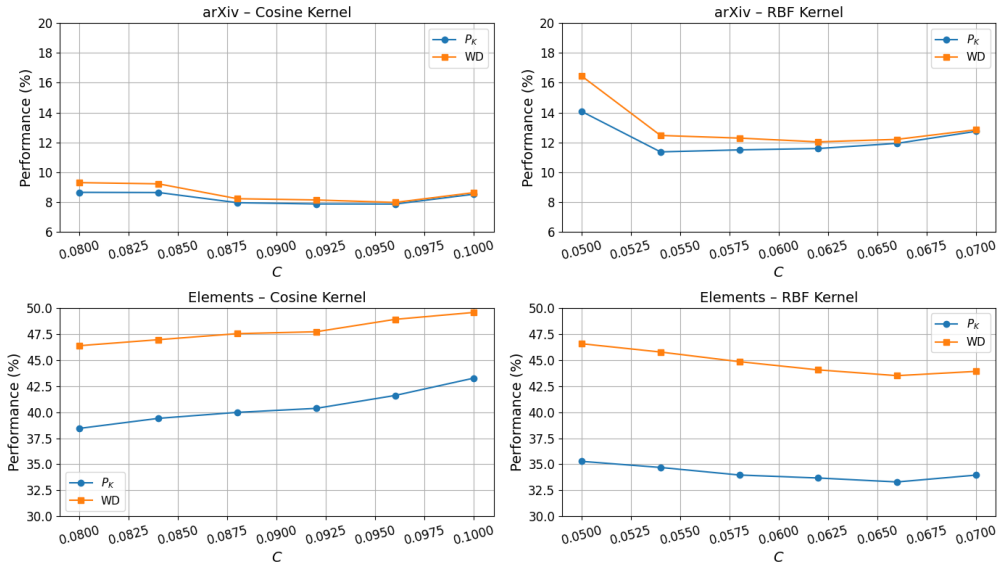


Figure 8: Sensitivity of C with cosine and RBF kernel on Elements and arXiv dataset.

For a target sequence length T, the number of change points, $K = \lceil 2 \log T \rceil$, increases slowly with T. Candidate change-point locations are sampled uniformly without replacement from $1, \ldots, T$, and converted into $K + 1$ segment lengths via successive differences. No explicit minimum segment length constraint is imposed. Instead, segment lengths are controlled implicitly: as T grows, the average segment length $T/(K + 1)$ also grows, ensuring that segments become longer asymptotically, consistent with the minimum spacing requirement in Assumption 3.4.

D.6 arXiv Dataset Generation

We construct new dataset based on the recent paper abstracts for text segmentation. The generation process is as follows:

- Select the first 1000 papers from arXiv published after August 2025.
- Randomly sample values between 5 and 20 to determine the number of unique abstracts per document.
- For each document, randomly select the corresponding number of abstracts, shuffle them to concatenate into a single text. Repeat this process.

The full list of 1000 arXiv papers used to built this dataset is included as part of the supplementary materials.

E Details for RAG-based Chunking Experiments

E.1 Datasets

We evaluate on three QA benchmarks spanning short- and long-document retrieval regimes. **HotpotQA** Yang et al. (2018) (distractor setting): ten paragraphs (two gold, eight distractors) are concatenated into one document per question (∼500–1,500 tokens); gold evidence is annotated at the sentence level. We sample 400 validation questions. **QASPER** Dasigi et al. (2021): each question refers to a full scientific paper (∼5,000–10,000 tokens) whose paragraphs are flattened into a single document; gold evidence is an annotated paragraph string. We use 300 test questions. **Natural Questions (NQ)** Kwiatkowski et al. (2019): each question is paired with a full Wikipedia article (∼5,000–30,000 tokens); gold evidence is the annotated long-answer paragraph. We sample 400 validation questions. All samples are drawn with a fixed random seed (42).

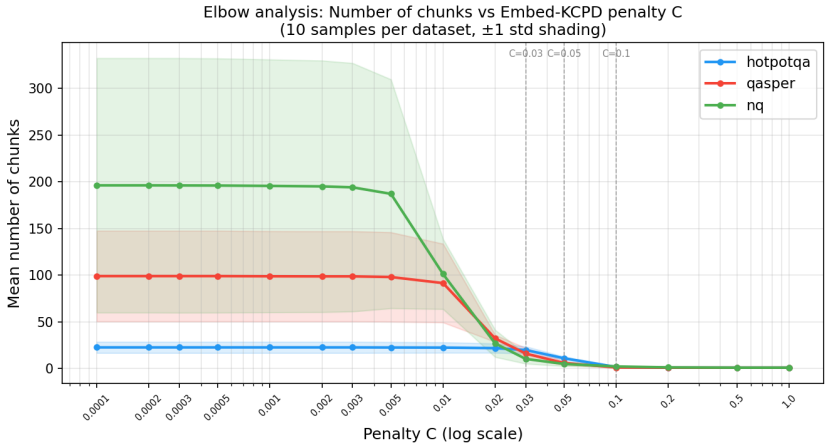


Figure 9: Elbow method analysis for RAG based experiments

E.2 Chunking Methods

We compare seven families of chunkers, all operating on the same `text-embedding-3-small` sentence embeddings (OpenAI, 1,536 dimensions) where applicable.

Fixed ($w \in \{50, 100, 200\}$): non-overlapping token windows (`tiktoken cl100k_base`).

Overlap ($2w$, stride w): sliding windows of size $2w$ with 50% overlap ($w \in \{50, 100, 200\}$).

SentBudget ($2w$): greedily packs spaCy sentences up to a $2w$-token budget ($w \in \{50, 100, 200\}$).

TextTiling: lexical cohesion segmentation (NLTK, w=20, k=10).

SemanticDrop ($K \in \{0.5, 1.0\}$): places a boundary where consecutive-sentence cosine similarity falls below $\mu - K\sigma$.

Embed-KCPD: We choose C based on the elbow method on 10 random samples (Figure 9).

E.3 Retrieval and Metrics

Chunks and queries are embedded with `text-embedding-3-small` A retrieved chunk matches a gold unit if their token-level Jaccard overlap ≥ 0.5. We report at depth k=5: **Hit@5** (at least one match retrieved), **Recall@5** (fraction of gold units covered), **Coverage@5** (all gold units covered), and **MRR@5** (mean reciprocal rank of first match).

Table 5: Full results on **HotpotQA** (400 questions, k=5). $\bar{n}$: mean chunks per question; $\bar{\ell}$: mean tokens per chunk; t: mean chunking time (s/question). Bold = best per column.

Method	Hit@5	Rec@5	Cov@5	MRR@5	$\bar{n}$	$\bar{\ell}$	t (s)
Length-based							
Fixed-50	0.600	0.317	0.085	0.413	26.5	34.3	0.001
Fixed-100	0.168	0.074	0.003	0.102	13.5	67.2	0.001
Overlap-100	0.203	0.091	0.003	0.118	17.5	68.2	0.001
Overlap-200	0.008	0.003	0.000	0.004	8.8	133.5	0.001
Sent Budget-100	0.433	0.204	0.023	0.288	15.8	57.6	0.127
Sent Budget-200	0.043	0.019	0.000	0.029	7.4	121.9	0.109
Semantic baselines							
TextTiling	0.075	0.032	0.000	0.046	8.0	113.9	0.545
TextTiling +ov	0.008	0.003	0.000	0.003	8.0	215.4	0.686
SemanticDrop K=0.25	0.623	0.328	0.083	0.430	17.8	51.9	1.336
SemanticDrop K=0.50	0.498	0.256	0.060	0.338	14.1	65.5	0.265
SemanticDrop K=1.0	0.243	0.115	0.015	0.155	8.0	116.6	0.174
SemanticDrop K=0.25 +ov	0.270	0.125	0.010	0.173	17.8	99.6	1.074
SemanticDrop K=0.50 +ov	0.188	0.083	0.003	0.126	14.1	124.7	0.230
SemanticDrop K=1.0 +ov	0.068	0.030	0.000	0.041	8.0	215.1	0.772
EMBED-KCPD							
Base	**0.658**	**0.354**	**0.093**	**0.483**	18.7	48.6	0.124
+ov	0.068	0.031	0.000	0.038	18.7	94.1	1.219
Random-matched	0.558	0.282	0.053	0.384	18.7	48.6	0.861

Table 6: Full results on **QASPER** (300 questions, k=5). See Table 5 for column definitions. Bold = best per column.

Method	Hit@5	Rec@5	Cov@5	MRR@5	$\bar{n}$	$\bar{\ell}$	t (s)
Length-based							
Fixed-50	0.220	0.094	0.037	0.124	92.2	37.3	0.003
Fixed-100	0.403	0.236	0.133	0.254	46.3	73.9	0.002
Fixed-200	0.410	0.238	0.130	0.245	23.4	145.8	0.002
Overlap-100	0.437	0.275	0.160	0.257	61.3	74.3	0.003
Overlap-200	0.483	0.291	0.163	0.281	30.7	147.2	0.002
Overlap-400	0.160	0.087	0.043	0.103	15.4	290.8	0.002
Sent Budget-100	0.410	0.253	0.147	0.252	54.8	62.2	0.418
Sent Budget-200	0.453	0.274	0.160	0.292	25.8	132.4	0.400
Sent Budget-400	0.157	0.083	0.040	0.093	12.6	270.9	0.401
Semantic baselines							
TextTiling	0.427	0.241	0.127	0.253	28.5	119.7	1.613
TextTiling +ov	0.270	0.155	0.080	0.181	28.5	234.3	1.738
SemanticDrop K=0.50	0.413	0.245	0.140	0.262	51.5	66.9	1.458
SemanticDrop K=1.0	0.333	0.196	0.110	0.226	27.7	124.3	0.491
SemanticDrop K=0.50 +ov	0.437	0.255	0.143	0.275	51.5	130.7	1.717
SemanticDrop K=1.0 +ov	0.300	0.174	0.100	0.181	27.7	238.6	0.963
EMBED-KCPD							
Base	0.443	0.259	0.133	0.294	46.4	76.1	0.434
+ov	**0.500**	**0.319**	**0.193**	**0.324**	74.7	90.1	0.575
Random-matched	0.380	0.225	0.123	0.223	54.7	62.6	0.438

Table 7: Full results on **NQ** (400 questions, k=5). Because NQ has a single gold passage per question, Hit@5, Rec@5, and Cov@5 are numerically equal (any match = full recall = full coverage). See Table 5 for other column definitions. Bold = best per column.

Method	**Hit@5**	**Rec@5**	**Cov@5**	**MRR@5**	$\bar{n}$	$\bar{\ell}$	t (s)
Length-based							
Fixed-50	0.105	0.105	0.105	0.059	206.1	39.2	0.006
Fixed-100	0.315	0.315	0.315	0.201	103.3	77.8	0.005
Fixed-200	0.323	0.323	0.323	0.219	51.9	153.7	0.016
Overlap-100	0.343	0.343	0.343	0.222	137.2	78.1	0.006
Overlap-200	0.360	0.360	0.360	0.226	68.7	155.1	0.005
Overlap-400	0.123	0.123	0.123	0.073	34.4	305.4	0.014
Sent Budget-100	0.283	0.283	0.283	0.184	117.4	69.4	0.852
Sent Budget-200	0.355	0.355	0.355	0.229	57.2	139.3	2.013
Sent Budget-400	0.145	0.145	0.145	0.089	27.9	284.5	0.848
Semantic baselines							
TextTiling	0.322	0.322	0.322	0.217	40.5	164.5	3.980
TextTiling +ov	0.241	0.241	0.241	0.148	40.5	322.9	4.091
SemanticDrop K=0.25	0.273	0.273	0.273	0.181	212.1	42.7	4.143
SemanticDrop K=0.50	0.263	0.263	0.263	0.166	170.0	53.0	8.706
SemanticDrop K=1.0	0.159	0.159	0.159	0.109	74.7	115.1	10.845
SemanticDrop K=0.25 +ov	0.275	0.275	0.275	0.186	212.1	83.5	9.719
SemanticDrop K=0.50 +ov	0.203	0.203	0.203	0.141	169.9	103.2	3.753
SemanticDrop K=1.0 +ov	0.111	0.111	0.111	0.082	74.2	219.3	0.952
EMBED-KCPD							
Base	0.320	0.320	0.320	0.205	190.3	42.8	8.124
+ov	**0.400**	**0.400**	**0.400**	**0.275**	190.3	84.6	7.426
Random-matched	0.226	0.226	0.226	0.145	87.3	116.7	8.181

Table 8: Results from Table 2 with 95% confidence intervals (normal approximation).

Method	Hit	Rec	Cov	MRR
		HotpotQA		
Baselines				
Fixed	0.60 ±.05	0.32 ±.05	0.09 ±.03	0.41 ±.05
Overlap	0.20 ±.04	0.09 ±.03	0.00 ±.00	0.12 ±.03
Sent. Budget	0.43 ±.05	0.20 ±.04	0.02 ±.01	0.29 ±.04
TextTiling	0.07 ±.03	0.03 ±.02	0.00 ±.00	0.05 ±.02
TextTiling+ov	0.01 ±.01	0.00 ±.01	0.00 ±.00	0.00 ±.01
SemanticDrop	0.62 ±.05	0.33 ±.05	0.08 ±.03	0.43 ±.05
SemanticDrop+ov	0.27 ±.04	0.13 ±.03	0.01 ±.01	0.17 ±.04
EMBED-KCPD				
Base	**0.66 ±.05**	**0.35 ±.05**	**0.09 ±.03**	**0.48 ±.05**
+ov	0.07 ±.02	0.03 ±.02	0.00 ±.00	0.04 ±.02
Rand.	0.56 ±.05	0.28 ±.04	0.05 ±.02	0.38 ±.05
		QASPER		
Baselines				
Fixed	0.41 ±.06	0.24 ±.05	0.13 ±.04	0.25 ±.05
Overlap	0.48 ±.06	0.29 ±.05	0.16 ±.04	0.28 ±.05
Sent. Budget	0.45 ±.06	0.27 ±.05	0.16 ±.04	0.29 ±.05
TextTiling	0.43 ±.06	0.24 ±.05	0.13 ±.04	0.25 ±.05
TextTiling+ov	0.27 ±.05	0.15 ±.04	0.08 ±.03	0.18 ±.04
SemanticDrop	0.41 ±.06	0.24 ±.05	0.14 ±.04	0.26 ±.05
SemanticDrop+ov	0.44 ±.06	0.25 ±.05	0.14 ±.04	0.27 ±.05
EMBED-KCPD				
Base	0.44 ±.06	0.26 ±.05	0.13 ±.04	0.29 ±.05
+ov	**0.50 ±.06**	**0.32 ±.05**	**0.19 ±.04**	**0.32 ±.05**
Rand.	0.38 ±.05	0.23 ±.05	0.13 ±.04	0.22 ±.05
		NQ†		
Baselines				
Fixed	0.32 ±.05	0.32 ±.05	0.32 ±.05	0.22 ±.04
Overlap	0.36 ±.05	0.36 ±.05	0.36 ±.05	0.23 ±.04
Sent. Budget	0.35 ±.05	0.35 ±.05	0.35 ±.05	0.23 ±.04
TextTiling	0.32 ±.05	0.32 ±.05	0.32 ±.05	0.22 ±.04
TextTiling+ov	0.24 ±.04	0.24 ±.04	0.24 ±.04	0.15 ±.03
SemanticDrop	0.27 ±.04	0.27 ±.04	0.27 ±.04	0.18 ±.04
SemanticDrop+ov	0.28 ±.04	0.28 ±.04	0.28 ±.04	0.19 ±.04
EMBED-KCPD				
Base	0.32 ±.05	0.32 ±.05	0.32 ±.05	0.20 ±.04
+ov	**0.40 ±.05**	**0.40 ±.05**	**0.40 ±.05**	**0.27 ±.04**
Rand.	0.23 ±.04	0.23 ±.04	0.23 ±.04	0.15 ±.03

F COMPUTATIONAL RESOURCES FOR EXPERIMENTS

All experiments were conducted on a standard laptop without specialized hardware. In Section 3, we query GPT-4.1 via the OpenAI API, incurring a total cost of under $5. Additionally, generating embeddings with the text-embedding-small model also cost less than $5.