UDC 004.85:004.912:519.8
DOI https://doi.org/10.32782/2663-5941/2026.2.1/03

Bilyi M.O.
https://orcid.org/0009-0004-0639-3658
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"

Krylov Ie.V.
https://orcid.org/0000-0003-4313-938X
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"

MEDIAN CHANGE POINT DETECTION: A SEMANTIC TEXT SEGMENTATION METHOD FOR INFORMATION SYSTEMS WITH VECTOR DATABASES

This article addresses the pressing issue of improving the efficiency of information systems with vector databases by advancing the methods of data preprocessing and semantic segmentation of unstructured text. The operational quality of such systems, particularly those built on the Retrieval-Augmented Generation (RAG) architecture, critically depends on how accurately the input text is divided into logically complete and coherent fragments. Existing heuristic methods (such as fixed-length chunking or paragraph-based splitting) often sever semantic connections and disrupt coreference, while semantic approaches based on encoder models are computationally prohibitive. Alternative statistical methods that utilize the perplexity (PPL) metric of Large Language Models (LLMs) to detect topic boundaries demonstrate high processing speed but suffer from a significant fundamental flaw. They are highly sensitive to local linguistic fluctuations: the occurrence of specific terminology, abbreviations, or rare proper nouns causes sharp spikes in Cross-Entropy Loss, which baseline algorithms falsely identify as semantic block boundaries, leading to excessive and unnatural text fragmentation.

To overcome this challenge, this study proposes and experimentally validates a novel semantic segmentation algorithm named Median Change Point Detection (MCPD). The proposed method radically shifts the mathematical paradigm of boundary detection. Firstly, instead of using the arithmetic mean, it employs the median of token losses within a sentence, which acts as a natural low-pass filter and allows the algorithm to ignore isolated lexical anomalies, identifying the true baseline of syntactic complexity. Secondly, the algorithm abandons unstable local cut-off thresholds in favor of global optimization using the Pruned Exact Linear Time (PELT) algorithm. PELT analyzes the variance of complexity across broad text windows and mathematically pinpoints points of sustained topic change in linear time O(N), effectively ignoring temporary bursts of linguistic "noise". Furthermore, MCPD implements a new segment formation strategy that strictly prioritizes the preservation of semantic integrity over rigid chunk length constraints.

The experimental evaluation of the algorithm was conducted on three diverse datasets: WikiSection, Qasper, and 2WikiMultihopQA, utilizing the Qwen2-1.5B-Instruct model. The research results conclusively demonstrated that the baseline PPL-based method is fundamentally unstable, requiring manual threshold tuning (ranging from 0.9 to 0.05) for each specific domain. In contrast, the developed MCPD method proved to be entirely universal and operated effectively with a single set of parameters across all datasets. On the task of identifying authorial section boundaries (WikiSection), the MCPD algorithm achieved a Boundary F1 score of 23.81%, outperforming the best possible result of the baseline method by 67%. The preservation of logical continuity in the generated fragments also ensured an increase in the quality of final answer generation (QA F1) on complex scientific papers and multi-hop reasoning tasks. The obtained results confirm that the transition to statistically grounded global optimization and the use of robust signals enables the creation of a highly efficient data preparation tool for modern information retrieval systems.

Keywords: *Retrieval-Augmented Generation, chunking, perplexity, vector databases, information systems.*

Formulation of the problem. Modern information systems, particularly those built on the Retrieval-Augmented Generation (RAG) architecture [10], utilize vector databases as an external knowledge source to overcome issues such as the generation of incorrect information ("hallucinations") [7] and the lack of access to up-to-date or proprietary corporate data.

The fundamental operational process of a RAG system involves vectorizing the user's query, retrieving the most relevant information fragments from a vector database, and subsequently generating a response by a language model based on the retrieved context [10]. Consequently, the accuracy of the information retrieval stage is a critical factor: the quality, completeness, and relevance of the retrieved information directly determine the reliability of the system's final response.

For efficient information retrieval, modern systems rely on dense semantic search methods [8]. Textual data is transformed into dense vectors of real numbers, called embeddings, using specialized neural network models. This approach allows semantically similar texts to be mapped close to each other in a multidimensional feature space, enabling search by meaning rather than merely by exact lexical keyword matches. However, the effectiveness of semantic search is significantly constrained by the architectural and mathematical properties of the embedding models.

Firstly, most modern embedding models impose a strict limit on the length of the input token sequence, typically ranging from 512 to 8192 tokens, due to the quadratic complexity of the attention mechanism [14]. Secondly, there is the fundamental problem of the "information bottleneck" [11]. If an excessively large text fragment containing numerous heterogeneous ideas and facts is fed into the model, the representation of these features is compressed and "averaged" into a single vector. As a result, specific semantic details are lost, critically degrading the vector's ability to be retrieved in response to a narrow, highly specific search query. Given this, the data preprocessing stage – segmenting large documents into smaller fragments prior to vectorization, becomes of paramount importance.

Simple mechanical splitting of text into fixed-length fragments or dividing it exclusively at sentence boundaries often leads to the severance of logical connections. A classic consequence of such an approach is the disruption of coreference: if a subject, such as a person's name, is mentioned in the first segment, and the subsequent segment refers to it only via a pronoun (e.g., "he" or "she"), the semantic content of the second segment becomes isolated and incomplete [6]. During a search query regarding this person's actions, the second fragment will not be retrieved by the system, leading to the loss of critically important context.

Therefore, ensuring the semantic integrity of information during its vectorization is a significant challenge. The development and investigation of novel, semantically grounded methods for adaptive text segmentation is a highly relevant scientific and practical task. Solving this problem will significantly reduce the level of information noise, eliminate the loss of local context, and consequently enhance the overall efficiency of information systems with vector databases by improving the precision and recall metrics of intelligent search.

Analysis of recent research and publications. The effectiveness of modern information systems with vector databases depends on two key stages: the quality of search algorithms and result aggregation, and the quality of data preprocessing. Although significant research efforts are directed toward improving hybrid search methods and rank fusion, particularly by modifying Reciprocal Rank Fusion (RRF) algorithms to mitigate the impact of irrelevant results [1], the fundamental problem remains the quality of the vector representations themselves. No matter how advanced the search algorithms are, their accuracy will be low if the input text was incorrectly divided into segments during the indexing stage.

Existing approaches to effective text preparation and segmentation for information systems can be broadly classified into three main groups: heuristic (rule-based), semantic, and statistical-model-based.

Heuristic approaches, particularly segmenting text into fixed-length fragments or segmenting with partial overlap, are the most common due to their simplicity and low computational complexity [13]. However, as noted by the authors in [12], such methods are "blind" to context: they frequently sever logical connections, paragraphs, and sentences at unnatural boundaries, leading to the loss of critically important information during vector search. A better form of the heuristic approach is structural segmentation – dividing text by natural paragraphs or Markdown structure. Although this partially avoids breaking logical chains, such a method generates fragments of highly uneven length. Furthermore, long paragraphs often contain several distinct semantic blocks, the information density of which is averaged out when transformed into a single vector.

To overcome these drawbacks, semantic segmentation methods have been proposed. Their essence lies in using encoder models to generate vector rep-

resentations of adjacent sentences, followed by calculating their cosine similarity. A drop in similarity below a predefined threshold is interpreted as a topic change. While such approaches are convenient from an engineering perspective and demonstrate a higher quality of baseline context preservation compared to heuristics, they have significant limitations. Firstly, studies [12] indicate their excessive computational cost, as they require the generation of a large number of embeddings during the preprocessing stage. Secondly, as noted in [15], semantic similarity models are often unable to capture subtle logical dependencies between sentences. For instance, a sequence of sentences with a clear logical progression (from general to specific or cause-and-effect) may be incorrectly severed by the algorithm due to the low cosine similarity of their lexical composition, ultimately causing the vector search results to deviate from the core semantic unit.

A distinct modern direction in semantic segmentation is the use of Large Language Models as analytical agents (LLM-as-a-Judge). In such approaches, the text division process is delegated directly to the LLM via specially crafted instructions (prompts), tasking the model with analyzing the text and autonomously determining the logical boundaries of paragraphs based on a deep understanding of the context [4]. This method ensures the highest quality of semantic connection preservation and can account for complex narrative structures within a document. However, its practical application in large-scale RAG systems is severely limited by high costs: processing each document requires full response generation from the language model, making this approach unsuitable for the mass indexing of large knowledge bases during preprocessing.

Given these limitations, an alternative and less resource-intensive direction is the use of internal statistical metrics of Large Language Models. Specifically, recent studies [15] have proposed using Perplexity (PPL) and information entropy to determine chunk boundaries. These metrics are based on logits – raw, unnormalized numerical values that the neural network generates for each word (token) from its vocabulary. The magnitude of a logit reflects the model's degree of confidence that this specific token should be the next in the text. After passing through a Softmax function, logits are converted into probabilities. Based on these probabilities, perplexity is calculated, conceptually reflecting the model's overall level of "surprise": the lower it is, the easier it is for the model to predict the next word. The core hypothesis is that local minima of perplexity – moments when the model is most confident in predicting the next token – coincide with the boundaries of logically complete semantic blocks or standard linguistic constructions. This approach is computationally efficient, as it allows topic boundaries to be identified "on the fly" during a single pass of the text through the model.

Despite the promise of using perplexity for text segmentation, the existing approach has a significant flaw that limits its effectiveness in real-world information systems. The problem lies in the high sensitivity of the PPL metric to local linguistic fluctuations. A conventional search for PPL minima reacts to any drop in the complexity of token generation. Such a drop is often caused not by a change in the thematic block, but by the appearance of common phrases, predictable syntactic constructions, or highly frequent proper nouns in the text.

To illustrate this, consider the following example: "The local government implemented strict environmental standards for all industrial enterprises in the region. This initiative received widespread support. The level of atmospheric carbon dioxide pollution decreased to a historic minimum." From a semantic perspective, these three sentences form a single, indivisible logical block describing a cause and its direct consequence. However, the language model evaluates their complexity differently. The first and third sentences are saturated with specific terminology, resulting in relatively higher perplexity scores (2.11 and 3.06, respectively). In contrast, the second sentence is a common journalistic cliché with high predictability, causing its perplexity to plummet (to 1.58), forming a local minimum as depicted in Figure 1. The baseline algorithm falsely identifies this drop as a chunk boundary and mistakenly cuts the text, separating the implemented standards from the results of their action.

As a result, such "noise" in the perplexity signal disrupts the integrity of semantic blocks, as the algorithm is unable to distinguish a minor predictable phrase from a genuine structural transition between document topics. Thus, the problem of developing a mathematically grounded criterion that would allow filtering out minor perplexity fluctuations and accurately identifying the true boundaries of semantic blocks – without resorting to resource-intensive semantic comparison methods or costly iterative calls to Large Language Models for direct text analysis – remains unresolved.

Task statement. Considering the identified problem of excessive sensitivity of existing statistical segmentation methods to local linguistic fluctuations, the main objective of this study is to improve the quality

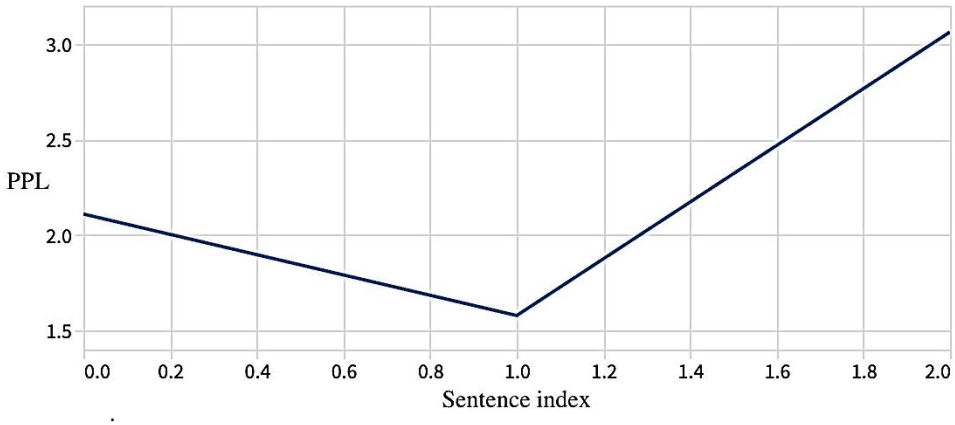

Fig. 1. Local perplexity minimum

of semantic segmentation of unstructured texts for information systems with vector databases by developing a mathematically grounded criterion for filtering "noise" in perplexity metrics.

To achieve this goal, the following scientific and practical tasks are to be solved in this work: to analyze the nature of the occurrence of local perplexity minima during the passage of text through a large language model and to identify patterns that distinguish genuine thematic transitions from minor syntactic or lexical predictabilities; to develop a new metric criterion for identifying the boundaries of semantic blocks, which will reduce the impact of local information noise and increase the accuracy of determining text division points; to build a segmentation algorithm based on the proposed criterion that will ensure the preservation of the semantic integrity of fragments for subsequent vectorization; to conduct an experimental validation of the developed method on benchmark datasets and perform a comparative analysis of its effectiveness against existing segmentation approaches using precision, recall, and F1-score metrics.

Outline of the main material of the study. Within the scope of this study, a novel algorithm for semantic text segmentation for retrieval-augmented generation systems, called Median Change Point Detection (MCPD), was developed and experimentally validated. The proposed method conceptually advances the idea of utilizing the internal "uncertainty" of large language models to detect logical boundaries in text, a concept initially described in the Meta-Chunking framework [15]. This method resolves the fundamental structural flaws of the existing approach by altering the mathematical paradigm of boundary detection and the segment formation strategy.

The theoretical foundation of the MCPD algorithm is based on generating a robust signal of text complexity. Let the input text be pre-divided into a sequence of sentences $S = (s_1,s_2,\ldots,s_n)$. After tokenization, each sentence s_i is represented as a sequence of K_i tokens: $s_i=(t^i_1,t^i_2,\ldots,t^i_{Ki})$. For each token t^i_k, the large language model computes the probability of its occurrence given the entire preceding context $t^i_{<k}$ (all tokens preceding it in the current sentence) and $t_{<i}$ (all tokens from previous sentences). In the context of language models, this probability essentially reflects the degree of predictability of the next word: a high probability, close to 1, indicates that the word is a logical and grammatical continuation of the thought, whereas a low probability, close to 0, signifies a topic change, the introduction of a new term, or an unexpected syntactic turn. To convert these probabilities into a convenient additive metric, the Cross-Entropy Loss is calculated, which reflects the degree of the model's "error" in prediction. For a specific token t_{ik}, it is defined as the negative logarithm of this probability:

$$L(t^i_k) = -\log P_M(t^i_k \mid t^i_{<k}, t_{<i}) \quad (1)$$

where P_M – is the probability distribution generated by model M.

In the baseline Meta-Chunking method, the semantic complexity of a sentence s_i is calculated as the arithmetic mean of the losses of all its tokens. However, this approach has proven to be highly sensitive to statistical outliers. The presence of a single specific term or an unusual proper noun artificially inflates the average score of the entire sentence, generating lexical "noise".

To illustrate this, consider the sentence: "The weather is beautiful today, the sun is shining, and we are walking around the city of Tlaquepaque." Most words in this construction are common and easily predictable for a language model. The model generates them with high probability (e.g., $P(t_k) = 0.9$), accordingly, their loss metric is minimal: $-\log(0.9) \approx 0.046$. However, the last word – "Tlaquepaque"

(the name of a relatively obscure Mexican city) – is semantically unexpected in this context. For this single token, the model will generate an extremely low probability (e.g., $P(t_k) = 0.0001$), leading to a sharp spike in the loss metric: $-\log(0.0001) \approx 4.0$. Consequently, due to one specific token, the entire sentence receives an artificially inflated complexity score.

In contrast, the proposed MCPD method forms a set of losses for all tokens in the sentence $L_i = \{L(t^i_1), L(t^i_2), \dots, L(t^i_K)\}$, and the sentence signal is defined as the median of this set. Mathematically, if the elements of set L_i are sorted in ascending order ($L_{(1)} \le L_{(2)} \le \cdots \le L_{(Ki)}$), the median is calculated as:

$$MedianLoss(s_i) = \begin{cases} L_{((K_i+1)/2)}, & \text{if } K_i \text{ is odd} \\ (L_{(K_i/2)} + L_{(K_i/2+1)})/2, & \text{if } K_i \text{ is even} \end{cases} \quad (2)$$

The use of the median acts as a natural low-pass filter: it allows the algorithm to ignore lexical anomalies and analyze exclusively the "background" syntactic and semantic complexity.

Once each sentence has received its robust signal, the task arises to find the optimal cut points that will divide the sequence into logically coherent fragments. Existing approaches use "greedy" algorithms, where a cut is made when the difference between the signals of two adjacent sentences exceeds a rigidly defined threshold, for example, $\Delta PPL \ge 0.3$. This approach is unstable because it reacts to local noise and fails to "see" the overall picture of the text, leading to false fragmentation and high domain dependency.

To overcome this problem, in the MCPD method, the segmentation task is formalized as a classical change point detection problem in a time series, solved using the Pruned Exact Linear Time (PELT) algorithm [9]. The goal of PELT is to divide the text into segments such that each segment is as homogeneous as possible in terms of complexity, without creating too many small fragments. Mathematically, this reduces to minimizing the objective function:

$$\min\tau,m\left(\sum_{i=1}^{m+1} C\left(y_{(\tau_{i-1}+1):\tau_i}\right) + \beta f(m)\right) \quad (3)$$

where, m - is the number of cut points dividing the text into m + 1 segments; τi - is the index of the i -th cut point (with τ0 = 0, and τm + 1 = N, where N - is the total number of sentences); $y_{(\tau_{i-1}+1):\tau_i}$ - is the sequence of sentence perplexity medians within the i -th segment; C - is a cost function measuring the level of internal variance within a single segment; $\beta f(m)$ is a penalty function for creating new cuts, which prevents excessive text fragmentation and false triggering of the algorithm on local information noise.

In this work, the sum of squared deviations from the mean is chosen as the cost function C, and the penalty function is linear ($f(m) = m$), where each new cut increases the total cost by a fixed constant β. The proposed objective function implements a strict balancing mechanism. The algorithm executes a text cut only if the reduction in internal variance resulting from the division exceeds the fixed cost of the new cut β. Due to this, PELT ignores minor local perplexity fluctuations and reacts exclusively to significant, global shifts in the baseline level of text complexity.

Calculating all possible combinations of cuts would take exponential time. However, the innovativeness of PELT lies in its mathematically proven pruning mechanism. The algorithm sequentially evaluates each sentence and, if it determines that a certain point in the past can mathematically never become part of the optimal set of cuts, it permanently removes it from memory. Thanks to this, PELT guarantees finding the absolute exact global optimum at a speed linearly dependent on the text length – O(N).

The effectiveness of this approach is clearly demonstrated by the dynamics of perplexity changes depending on the text topic, as illustrated in Figure 2. The graph clearly traces the structural blocks of the document: sentences 0 to 35 describe historical events and have their own baseline complexity level; the subsequent block, sentences 35-73, is dedicated to demographics, accompanied by a change in the overall signal variance. This is followed by transitions to the economy block, sentences 73-91, and the government description, sentences 92-113. As seen in the figure, local "spikes" (noise) are present within each thematic block, yet the PELT algorithm successfully ignores them, fixing cuts only at the points of global change in the statistical characteristics of the signal between these macro-blocks.

The physical essence of this process can be illustrated through an analogy with driving a car, where the text acts as the road, and token perplexity acts as the "tension" level of the language model. Local lexical noise (e.g., the appearance of a rare word) resembles a pothole on a smooth highway: the system registers a sharp jump in the PPL metric, but immediately after passing this word, the complexity returns to the initial baseline level. The PELT algorithm, analyzing the variance over a broader window of sentences, recognizes this spike as a temporary anomaly and does not create a cut. Conversely, a genuine topic change (e.g., a transition from a general description to complex terminology) acts like turning onto a winding mountain road: the complexity level rises and remains consistently high throughout the subsequent fragment. In this case, the algorithm detects a sustained shift in the baseline perplexity level (e.g., a change in the mean value from 12.0 to 17.5) and mathematically justifies dividing the semantic blocks at exactly this point.

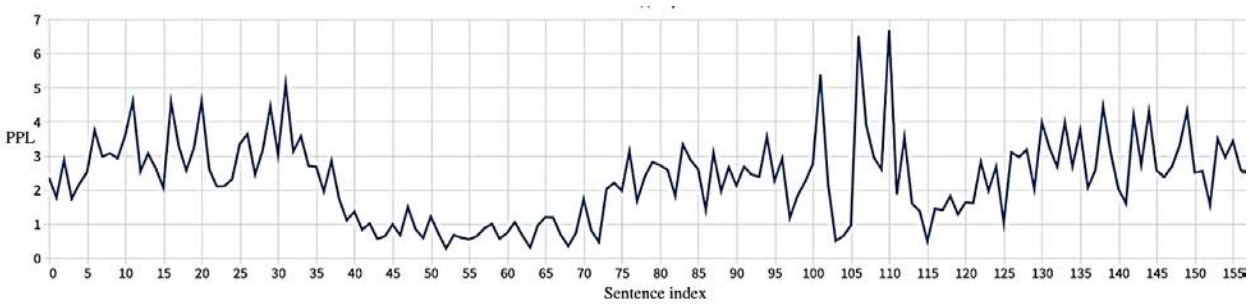

Fig. 2. Dynamics of perplexity changes across structural thematic blocks of the document

An important aspect of the practical application of the PELT algorithm is rethinking the approach to chunk length and the need to adapt the mathematical model to the engineering constraints of vector databases. Since classical PELT optimizes only variance and does not account for the physical size of the blocks, an adaptive buffering strategy was developed in this study. The algorithm accumulates input sentences into a buffer whose size is twice the target maximum segment length, for example, 2×256 tokens. This window size was not chosen randomly: it is the minimum sufficient size for the algorithm to "see" a broader context and compare the variance of the current block with the next potential block, while maintaining low RAM consumption and high computational speed compared to analyzing the entire document at once. Subsequently, PELT analyzes the accumulated array and generates a set of optimal cut points. From this set, the algorithm selects the point that forms the largest semantically coherent segment not exceeding the specified limit.

However, traditional systems tend to artificially and strictly limit segment length, often resorting to forced cutting of long paragraphs or mechanical merging of small fragments if a cut point is not found. In contrast, the MCPD method relies on the logic that semantic continuity of the text is the absolute priority. In cases where PELT mathematically proves that an entire text array within the limit is absolutely homogeneous and contains no logical breaks (e.g., a block of 400 or 500 tokens), a "relaxed limit" strategy is applied. Instead of "blindly" cutting the text in half, the algorithm ignores the strict limit and leaves such a block undivided, executing a cut only at the first detected semantic boundary. It is significantly better to form a larger segment that retains all necessary interconnected information than to artificially tear it apart. Despite the increased text volume, modern vectorization models are capable of effectively capturing the meaning of long but semantically homogeneous passages without losing semantic focus in the generated vector. Similarly, generative models in RAG systems operate much more accurately when they receive the complete authorial context without losing logical connections.

To objectively confirm the theoretical advantages of the method, a series of experiments was conducted on three diverse datasets. The algorithm's ability to accurately determine authorial section boundaries was tested on Wikipedia articles from the WikiSection dataset [2] using the Boundary F1 metric. The impact of segmentation on the quality of answer generation in RAG systems was evaluated using the QA F1 metric on complex scientific articles from the Qasper dataset [3], as well as on multi-hop logical reasoning tasks from the 2WikiMultihopQA dataset [5]. The Qwen2-1.5B-Instruct model was used as the baseline model for both generating logits and formulating final answers.

The F1 Score, which is the harmonic mean of precision and recall, was chosen as the primary criterion for evaluating generation quality. In the context of RAG systems, a high F1 score indicates that the segmentation algorithm forms text fragments that are so high-quality and semantically clean that the language model can find all the necessary information for an answer within them, without being distracted by irrelevant "noise" from adjacent paragraphs.

The obtained results are presented in Table 1. They demonstrate that the use of rigid local thresholds makes the baseline method fundamentally unstable. During testing on the WikiSection dataset, it achieved its maximum result (Boundary F1 = 14.21%) at a high cut threshold of θ=0.9, whereas on the Qasper scientific texts, the threshold had to be reduced 18-fold – to θ=0.05 – to achieve the maximum (QA F1 = 18.82%). Such adjustment makes it impossible to use the baseline method in production systems without prior fine-tuning for each text type. In contrast, the developed MCPD method proved to be universal and demonstrated high effectiveness without additional adaptation, using a single set of baseline settings across all datasets.

The experiment on the WikiSection dataset confirmed the undeniable superiority of global optimization over local optimization: the MCPD algorithm achieved a Boundary F1 score of 23.81%, which outperforms the best possible result of the baseline PPL Chunking method by 67%. The preservation of semantic integrity without artificial fragmentation had a direct positive impact on the quality of final

answer generation. On the Qasper scientific article dataset, the MCPD method provided an F1 Score of 19.29%, bypassing the baseline PPL Chunking method. Similar stability (F1 = 18.12%) was observed on the 2WikiMultihopQA dataset compared to 18.06% for the baseline approach. It is worth noting that the increase in the overall score in the MCPD method is achieved primarily due to an increase in precision, which confirms the algorithm's ability to filter out irrelevant context.

Table 1

Comparison of the effectiveness of the PPL Chunking method and the MCPD method

Datset	Metric	PPL Chunking	MCPD
WikiSection	Boundary F1	14.21% (at θ = 0.90)	23.81%
2WikiMultihopQA	QA F1	18.82% (at θ = 0.05)	19.29%
Qasper	QA F1	18.06% (at θ = 0.30)	18.12%

Conclusions. This study solves the relevant scientific and practical task of improving the quality of semantic segmentation of unstructured texts for information systems with vector databases. Based on the conducted analysis, it was revealed that existing statistical methods relying on the search for local perplexity minima are excessively sensitive to lexical "noise", inevitably leading to false and excessive text fragmentation. To overcome this problem, a novel algorithm, Median Change Point Detection (MCPD), was developed. Its key innovation is the transition from local heuristic thresholds to statistically grounded global optimization using the PELT algorithm, as well as the use of the median of token losses as a robust signal. This allowed for the effective filtering of minor linguistic fluctuations and the accurate identification of genuine thematic transitions.

Experimental validation on three diverse datasets confirmed the high effectiveness of the proposed approach. The results proved that the MCPD algorithm is universal and operates stably with a single set of settings, eliminating the need to tune thresholds for specific text types, unlike the baseline method. Specifically, on the task of determining authorial section boundaries (WikiSection), the developed method achieved a Boundary F1 score of 23.81%, which outperforms the best possible result of the baseline approach by 67%. Furthermore, preserving the semantic integrity of fragments without artificial cutting or mechanical merging ensured an increase in the quality of final answer generation in RAG systems on complex scientific and multi-hop texts.

Prospects for further research in this direction encompass several key tasks. Firstly, there are plans to develop a mechanism for the dynamic, adaptive calculation of the penalty for the PELT algorithm depending on the overall length and structural density of the input document. Secondly, it is expedient to investigate the impact of the language model's size and architecture – particularly the use of more compact and faster LLMs – on the quality of logit generation and the ultimate segmentation accuracy. Finally, an important step will be expanding the experimental base to evaluate the effectiveness of the MCPD algorithm on multilingual text corpora and integrating the method into real-time data streaming pipelines.

Bibliography:

1. Білий М., Крилов Є. Модифікація методу Reciprocal Rank Fusion для поліпшення результатів гібридного пошуку у інформаційних системах з векторними базами даних. Інформаційні технології та суспільство. 2025. № 2 (17). С. 14–19. DOI: https://doi.org/10.32689/maup.it.2025.2.2.

2. Arnold S., Schneider R., Cudré-Mauroux P., Gers F. A., Löser A. SECTOR: A Neural Model for Coherent Topic Segmentation and Classification. arXiv preprint arXiv:1902.04793. 2019. P. 1–16. DOI: https://doi.org/10.48550/arXiv.1902.04793.

3. Dasigi P., Lo K., Beltagy I., Cohan A., Smith N. A., Gardner M. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. arXiv preprint arXiv:2105.03011. 2021. P. 1–18. DOI: https://doi.org/10.48550/arXiv.2105.03011.

4. Duarte A. V., Marques J., Graça M., Freire M., Li L., Oliveira A. L. LumberChunker: Long-Form Narrative Document Segmentation. arXiv preprint arXiv:2406.17526. 2024. P. 1–16. DOI: https://doi.org/10.48550/arXiv.2406.17526.

5. Ho X., Nguyen A. K., Sugawara S., Aizawa A. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. arXiv preprint arXiv:2011.01060. 2020. P. 1–17. DOI: https://doi.org/10.48550/arXiv.2011.01060.

6. Jang Y., Hong S., Son J., Park S., Park C., Lim H. From Ambiguity to Accuracy: The Transformative Effect of Coreference Resolution on Retrieval-Augmented Generation systems. arXiv preprint arXiv:2507.07847. 2025. P. 1–15. DOI: https://doi.org/10.48550/arXiv.2507.07847.

7. Ji Z., Lee N., Frieske R., Yu T., Su D., Xu Y., Ishii E., Bang Y. J., Madotto A., Fung P. Survey of Hallucination in Natural Language Generation. arXiv preprint arXiv:2202.03629. 2022. P. 1–43. DOI: https://doi.org/10.48550/arXiv.2202.03629.

8. Karpukhin V., Oguz B., Min S., Lewis P., Wu L., Edunov S., Chen D., Yih W. Dense Passage Retrieval for Open-Domain Question Answering. arXiv preprint arXiv:2004.04906. 2020. P. 1–14. DOI: https://doi.org/10.48550/arXiv.2004.04906.

9. Killick R., Fearnhead P., Eckley I. A. Optimal Detection of Changepoints With a Linear Computational

Cost. arXiv preprint arXiv:1101.1438. 2011. P. 1–25. DOI: https://doi.org/10.48550/arXiv.1101.1438.
10. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W., Rocktäschel T., Riedel S., Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401. 2020. P. 1–19. DOI: https://doi.org/10.48550/arXiv.2005.11401.
11. Liu N. F., Lin K., Hewitt J., Paranjape A., Bevilacqua M., Petroni F., Liang P. Lost in the Middle: How Language Models Use Long Contexts. arXiv preprint arXiv:2307.03172. 2023. P. 1–20. DOI: https://doi.org/10.48550/arXiv.2307.03172.
12. Qu R., Tu R., Bao F. Is Semantic Chunking Worth the Computational Cost? arXiv preprint arXiv:2410.13070. 2024. P. 1–23. DOI: https://doi.org/10.48550/arXiv.2410.13070.
13. Shaukat M. A., Adnan M., Kuhn C. C. N. A Systematic Investigation of Document Chunking Strategies and Embedding Sensitivity. arXiv preprint arXiv:2603.06976. 2026. P. 1–25. DOI: https://doi.org/10.48550/arXiv.2603.06976.
14. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. Attention Is All You Need. arXiv preprint arXiv:1706.03762. 2017. P. 1–15. DOI: https://doi.org/10.48550/arXiv.1706.03762.
15. Zhao J., Ji Z., Feng Y., Qi P., Niu S., Tang B., Xiong F., Li Z. Meta-Chunking: Learning Text Segmentation and Semantic Completion via Logical Perception. arXiv preprint arXiv:2410.12788. 2024. P. 1–18. DOI: https://doi.org/10.48550/arXiv.2410.12788.

Білий М. О., Крилов Є. В. MEDIAN CHANGE POINT DETECTION: МЕТОД СЕМАНТИЧНОЇ СЕГМЕНТАЦІЇ ТЕКСТІВ ДЛЯ ІНФОРМАЦІЙНИХ СИСТЕМ З ВЕКТОРНИМИ БАЗАМИ ДАНИХ

У статті розглядається актуальна проблема підвищення ефективності інформаційних систем із векторними базами даних шляхом вдосконалення методів попередньої підготовки та семантичної сегментації неструктурованих текстових даних. Якість функціонування таких систем, зокрема тих, що побудовані на архітектурі генерації, доповненої пошуком (Retrieval-Augmented Generation, RAG), критично залежить від того, наскільки точно вхідний текст поділяється на логічно завершені фрагменти. Існуючі евристичні методи (фіксована довжина, поділ за абзацами) часто розривають семантичні зв'язки та порушують кореферентність, тоді як семантичні підходи на базі моделей-енкодерів є обчислювально надто ресурсоємними. Альтернативні статистичні методи, що використовують показник перплексії (Perplexity, PPL) великих мовних моделей для пошуку меж тем, демонструють високу швидкодію, проте мають суттєвий фундаментальний недолік. Вони є вкрай чутливими до локальних мовних флуктуацій: поява специфічних термінів, абревіатур чи рідкісних власних назв викликає різкі стрибки показника втрат (Cross-Entropy Loss), що хибно ідентифікуються базовими алгоритмами як межі смислових блоків, призводячи до надмірної та неприродної фрагментації тексту.

Для вирішення цієї проблеми у дослідженні розроблено та експериментально обґрунтовано новий алгоритм семантичного сегментування – Median Change Point Detection (MCPD). Запропонований метод кардинально змінює математичну парадигму виявлення меж. По-перше, замість арифметичного середнього використовується медіана втрат токенів речення, що діє як природний фільтр низьких частот і дозволяє ігнорувати точкові лексичні аномалії, виокремлюючи справжній базовий рівень синтаксичної складності. По-друге, алгоритм відмовляється від нестабільних локальних порогів розрізу на користь глобальної оптимізації за допомогою алгоритму Pruned Exact Linear Time (PELT). PELT аналізує дисперсію складності на широких вікнах тексту та математично точно визначає точки стійкої зміни теми за лінійний час O(N), ігноруючи тимчасові сплески "шуму". Крім того, MCPD реалізує нову стратегію формування сегментів, яка надає абсолютний пріоритет збереженню семантичної цілісності тексту над жорсткими обмеженнями довжини чанків.

Експериментальна перевірка алгоритму проводилася на трьох різнопланових наборах даних: WikiSection, Qasper та 2WikiMultihopQA з використанням моделі Qwen2-1.5B-Instruct. Результати дослідження переконливо довели, що базовий метод на основі PPL є фундаментально нестабільним і потребує ручного налаштування порогів (від 0.9 до 0.05) під кожен окремий домен. Натомість розроблений метод MCPD виявився повністю універсальним і працював з єдиними налаштуваннями на всіх датасетах. На завданні визначення авторських меж розділів (WikiSection) алгоритм MCPD досяг показника Boundary F1 на рівні 23.81%, що на 67% перевершує найкращий можливий результат базового методу. Збереження логічної неперервності фрагментів також забезпечило підвищення якості фінальної генерації відповідей (QA F1) на складних наукових статтях та завданнях із багатокроковим логічним виведенням. Здобуті результати підтверджують, що перехід до статистично обґрунтованої глобальної оптимізації та використання робастних сигналів дозволяє створити високоефективний інструмент підготовки даних для сучасних інформаційно-пошукових систем.

Ключові слова: *генерація доповнена пошуком, чанкінг, перплексія, векторні бази даних, інформаційні системи.*

Дата першого надходження статті до видання: 07.03.2026
Дата прийняття статті до друку після рецензування: 03.03.2026
Дата публікації (оприлюднення) статті 11.05.2026