IEEE SMC 2026

# SA-Chunk: Document Segmentation approach Based on Two-Level Semantic Feature Aggregation

**Liedong Guo[1,2]  Jian Xu[3,*]**

1 Qilu University of Technology (Shandong Academy of Sciences), Jinan
2 Key Lab. of Computing Power Network & Information Security, Ministry of Education
3 Shandong University of Finance and Economics (* corresponding author)



Contact: sdfixj@126.com

# Motivation

## The Problem

•**Semantic Fragmentation**: Fixed-size chunking ignores the inherent logic of text and easily breaks the context.
•**Low Retrieval Accuracy**: Overly fine-grained splitting loses key information, while overly coarse splitting introduces significant noise.
•**Lack of Hierarchical Construction**: Existing learning methods often rely on single features and fail to integrate multi-scale semantic units.

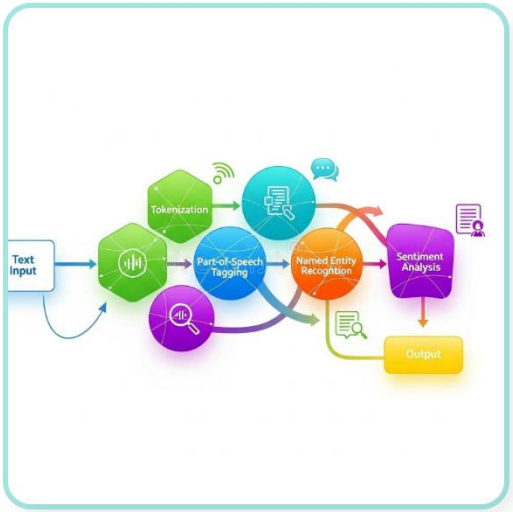
For RAG systems, the quality of text chunking determines retrieval accuracy and generation quality. SA-Chunk aims to solve the semantic fragmentation problem.

## Proposed Solution

•**Propose the SA-Chunk Framework**: Based on two-level semantic feature aggregation, from basic semantic fragments to compound semantic units.
•**Introduce Document-level Semantic Boundary Detection**: Identify boundaries using global semantic relevance to achieve adaptive sentence grouping.
•**Design an LLM-driven Feedback Mechanism**: Dynamically adjust segmentation thresholds to enhance model generalization.

# Background

As NLP technology evolves, intelligent question-answering and content generation have penetrated every industry. However, when processing massive long-form texts, efficiently understanding and modeling their deep semantics remains a core challenge. This challenge is particularly critical in Retrieval-Augmented Generation (RAG) systems, directly determining the system's knowledge accuracy and generation quality.




Core Bottleneck:

The performance of RAG systems heavily depends on the quality of Text Chunking. Chunking serves as the bridge connecting "retrieval" and "generation," and its rationality directly impacts subsequent recall rates and generation effectiveness.

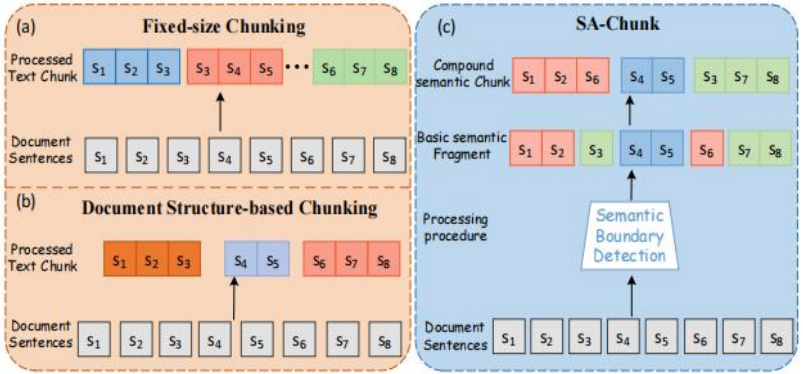
# Method Comparison

## 1 Fixed-size Chunk

Mechanically split by character or token count, completely ignoring the inherent logic of the text. Very likely to cut in the middle of sentences or logical paragraphs, causing semantic fragmentation and breaking the coherence and integrity of the context.

## 2 Structure-based Chunk

Reliant on explicit structural features like titles and paragraphs for segmentation. But it performs very poorly on unstructured text or content with sparse structure, making it hard to capture deeper semantic connections and logical flow.




## 3 SA-Chunk

To tackle the issues of semantic disconnection and fragmentation in traditional document segmentation, this paper introduces SA-Chunk, an innovative document segmentation framework based on a two-level semantic feature aggregation, aiming to fundamentally improve the semantic coherence in handling long texts.

# This paper's contributions

01 Hierarchical Aggregation Framework

Pioneering a dual-layer aggregation model of "Basic Segments – Composite Units," it progressively integrates semantic features to effectively alleviate semantic fragmentation caused by traditional segmentation, ensuring the semantic integrity and logical coherence of text chunks.

02 Adaptive Boundary Detection

Introducing a unified document-level semantic scoring mechanism that adaptively evaluates sentence importance and topic boundaries, enabling dynamic sentence grouping and precise identification of semantic breakpoints, avoiding rigid mechanical splitting.

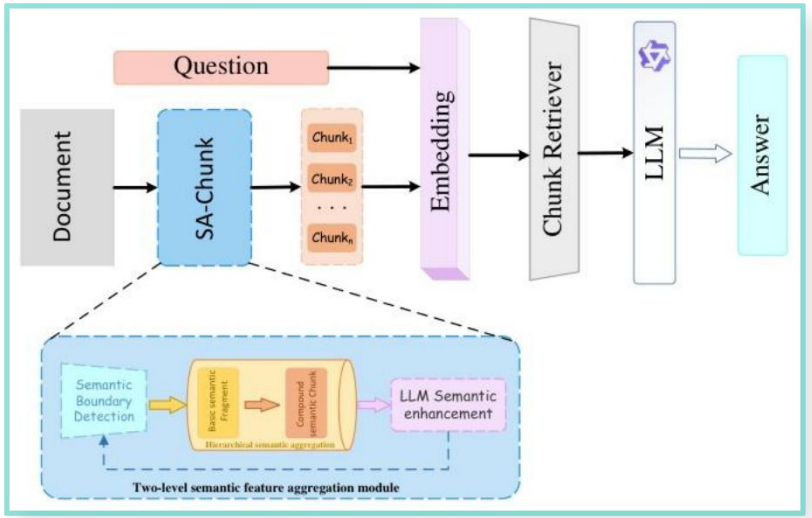
03 LLM-Driven Feedback Optimization

Leveraging the comprehension capabilities of large language models to build a closed-loop feedback mechanism, dynamically adjusting segmentation thresholds and aggregation strategies, effectively reducing semantic hallucinations during segmentation, and enhancing the model's generalization across diverse domains.



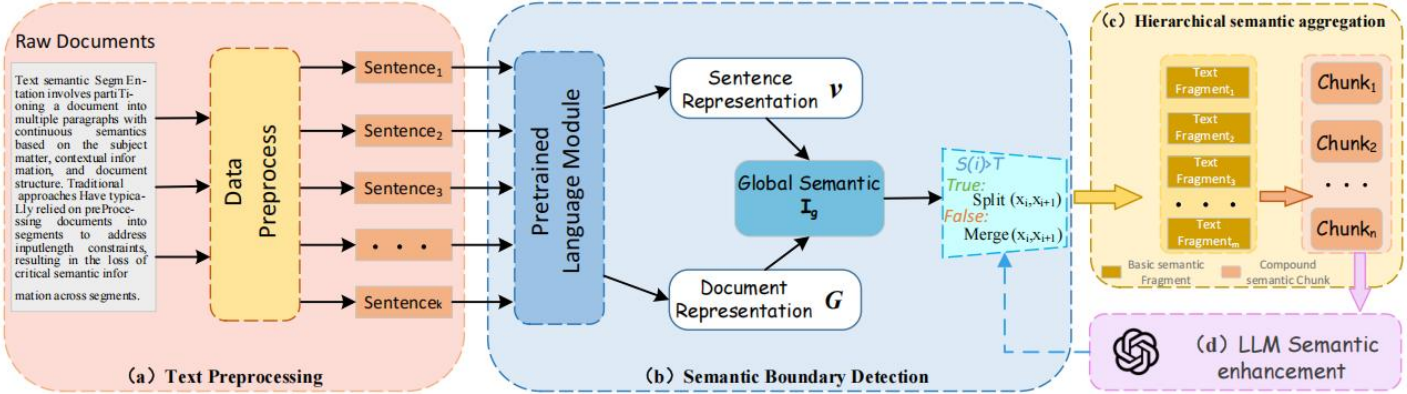
# Framework Overview

RAG System Integration and Performance Enhancement
SA-Chunk replaces traditional rigid chunking strategies by leveraging an innovative Two-Layer Semantic Feature Aggregation (TSFA) mechanism to generate semantically complete and logically coherent text chunks. After these high-quality chunks are encoded and stored in the vector database, they can more accurately match user query intent during the retrieval phase, providing highly relevant context for Large Language Models (LLM). This effectively resolves the semantic fragmentation issues caused by traditional chunking, significantly improving the accuracy of question answering and knowledge coverage.



# Methodology

## Four Core Module Architecture




The framework is driven by a closed loop of four key stages: document preprocessing cleans noisy data, semantic boundary detection precisely identifies natural segmentation points in the text, hierarchical semantic aggregation achieves deep fusion of multi-granularity features, and finally, LLM semantic enhancement ensures the semantic integrity and coherence of the generated text chunks.

# Methodology

01 Document Pre-processing

- Coreference Resolution: Replaces pronouns with their antecedents to eliminate referential ambiguity, ensuring the coherence and integrity of textual semantics.
- Sentence Boundary Detection: Performs precise segmentation based on punctuation and syntactic rules, generating an ordered sequence of independent sentences to solidify the processing foundation.
- Backtracking Truncation Strategy: Intelligently truncates text at sentence boundaries to accommodate model input length constraints, avoiding rigid semantic fragmentation.

02 Semantic Boundary Detection

- Global Semantic Vector (G) Pools the full-document vector to extract holistic semantic features, anchoring the core theme of the document.
- Sentence–Global Similarity Calculates the similarity between individual sentences and the global vector, quantifying the relevance of local segments to the main topic.
- Boundary Score Calculation Captures abrupt similarity changes between adjacent sentences; the larger the mutation, the higher the probability of a boundary.
- Threshold-Based Boundary Determination When the score exceeds the threshold, it is determined to be a boundary, achieving precise topic segmentation.

# Methodology

### 03Hierarchical Semantic Aggregation

Merges basic semantic segments into richer and more coherent composite units. Through iterative merging, it constructs a multi-granularity semantic structure, enhancing the holism of text understanding.

- Similarity Calculation and Merge Decision

Calculates the similarity between adjacent semantic segments; if it exceeds the preset aggregation threshold, a merge is executed to form larger semantic chunks.

- Iterative Aggregation Process

Repeatedly performs merge operations until no new segments can be merged, constructing a multi-level semantic system ranging from fine-grained to coarse-grained.

### 04LLM Semantic Enhancement

Introduces a large model feedback mechanism to dynamically optimize segmentation thresholds, addressing the insufficient generalization capability of fixed thresholds, and improving the accuracy of semantic segmentation for complex texts.

- Initial Segmentation and Intelligent Evaluation

Generates a plan based on the initial threshold, and leverages LLM to perform end-to-end semantic consistency scoring on the segmented chunks.

- Dynamic Threshold Optimization

Iteratively adjusts the threshold by minimizing a quality loss function, balancing intra-chunk coherence with boundary precision.

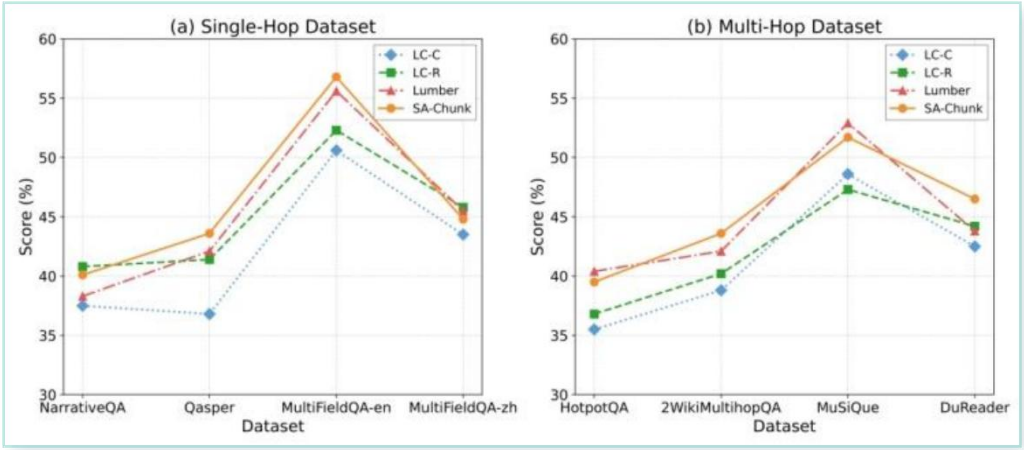
- Domain Adaptive Capability

Automatically adapts to the characteristics of texts from different domains, achieving more robust semantic segmentation and aggregation effects.

# RAG Benchmark Test

To fully test how effective SA-Chunk is in real RAG scenarios, we ran a systematic evaluation on 8 datasets from the LongBench benchmark, comparing it with mainstream chunking methods to verify its advantages in both performance and efficiency.

- Performance Comprehensively Surpasses Mainstream Baselines
- Lower Inference Cost and Higher Efficiency
- Hierarchical Semantic Aggregation Resolves Fragmentation Issues



# Ablation study

To check how effective each part of SA-Chunk is, we break down the model's performance through ablation experiments, digging into how much the core modules contribute to the final result and confirming whether the two-layer semantic feature aggregation mechanism is really necessary.

Experiments show that models based on Longformer significantly outperform traditional BERT and RoBERTa encoders on long-text processing tasks. Longformer-Large achieves the best performance across all evaluation metrics through its efficient long-range dependency capturing capability, making it the optimal backbone architecture choice for this study.

| SA-Chunk Base Model | WIKI-727k | | |
|---|---|---|---|
| | P | R | F1 |
| Longformer-Base | **80.21** | **73.64** | **76.78** |
| w/o TSFA | 80.01 | 73.06 | 76.37 |
| Longformer-Large | **82.62** | **76.45** | **79.41** |
| w/o TSFA | 79.84 | 75.63 | 77.67 |
| BERT-Base | 72.41 | **71.89** | **72.15** |
| w/o TSFA | **74.56** | 68.23 | 71.25 |

Removing the TSFA module leads to a significant degradation in model performance. On the strongest Longformer-Large backbone, the addition of TSFA improves the F1 score by nearly 2 percentage points, effectively addressing the limitations of single-feature representation in complex semantic boundary detection, and validating the effectiveness of the two-layer aggregation mechanism.

# Conclusion

This paper proposes a novel document segmentation model called SA-Chunk, aiming to alleviate the semantic fragmentation problem faced by traditional methods when processing complex texts by layered aggregation from basic semantic fragments to composite semantic units, providing a new solution for understanding and processing long texts.

