DATA MINING PROJECT REPORT

Academic year 2024 - 2025

Assessing Chunk Quality: The Chunk Quality for Multimodal RAG System

Ameziane Sarah

Under the supervision of

Omar Mourad

February 2026 - Oral Presentation

Jury:

Sami Belkacem

Mohamed Hadj Ameur

Ahmed Laouedj

Abstract

This study presents a novel data mining approach for optimizing Retrieval-Augmented Generation (RAG) systems through quality-based clustering and targeted content enhancement. We developed a comprehensive pipeline that identifies poorly performing document chunks using unsupervised clustering techniques, applies targeted enhancements to tables and mathematical formulas, and measures system-wide performance improvements. The methodology integrates advanced feature engineering (15-dimensional feature space), K-means clustering with quality-aware selection, principal component analysis for visualization, and rigorous statistical validation. Using the OHR-Bench evaluation framework with 8,498 question-answer pairs across seven domains, our analysis revealed five distinct quality clusters, with the poorest-performing cluster (POOR tier, mean LCS = 0.259) showing significant potential for improvement. Through targeted enhancement of table extraction and LaTeX handling in low-performing chunks (n=1,468), we achieved a 27.37% improvement in the POOR tier and a 1.96% overall system enhancement with extreme statistical significance . This research demonstrates that clustering-based pattern recognition can effectively identify and remediate quality issues in RAG systems, providing a scalable, data-driven framework for systematic performance optimization without degrading high-performing content.

Keywords: Data Mining, Clustering Analysis, Retrieval-Augmented Generation, Feature Engineering, Performance Optimization, Unsupervised Learning, K-Means Clustering

Contents

1 Introduction

1.1 Background and Motivation

Retrieval-Augmented Generation (RAG) systems have emerged as a transformative paradigm in natural language processing, combining the generative capabilities of large language models (LLMs) with the precision of external knowledge retrieval. Unlike pure generative models that rely solely on parametric knowledge, RAG systems dynamically retrieve relevant information from external corpora to ground their responses in factual, up-to-date content. This hybrid architecture has proven particularly valuable for question-answering, document understanding, and knowledge-intensive tasks where accuracy and verifiability are paramount.

1.1.1 The Critical Role of Chunk Quality in RAG Performance

At the heart of every RAG system lies a fundamental process: document chunking. Raw documents must be segmented into discrete, semantically coherent units (chunks) that can be indexed, retrieved, and provided as context to the language model. **The quality of these chunks directly determines the ceiling of RAG system performance.** Poor chunking creates a cascading failure pattern:

1. **Retrieval Failure**: Low-quality chunks fail to match relevant queries, causing the retrieval system to miss critical information even when it exists in the corpus.
2. **Context Degradation**: Even when retrieved, poorly structured chunks provide degraded context to the LLM, leading to hallucinations, incomplete answers, or outright errors in generation.

1.1.2 Why Chunk Quality Matters: Empirical Evidence

Recent studies have demonstrated that chunk quality issues account for a significant portion of RAG system failures:

- **Information Loss**: Improper chunking of tables and formulas can result in 40-60% loss of structured information, rendering critical data unretrievable.
- **Semantic Fragmentation**: Breaking semantic units (e.g., splitting a table from its caption, separating formulas from their context) reduces retrieval precision by 25-35%.
- **OCR-Induced Degradation**: Documents processed through OCR exhibit chunk quality variance of up to 50% compared to ground truth, with complex content (tables, mathematical notation) suffering disproportionately.

1.1.3 The OCR Complexity Factor

The challenge of maintaining chunk quality becomes particularly acute when processing documents through Optical Character Recognition (OCR) systems. OCR introduces multiple failure modes:

- **Layout Corruption**: Table structures collapse into unstructured text, losing row-column relationships critical for comprehension.
- **Formula Degradation**: Mathematical notation is converted to plain text or corrupted symbols, rendering equations uninterpretable.
- **Reading Order Errors**: Multi-column layouts, figures, and footnotes are processed in incorrect sequence, fragmenting semantic coherence.
- **Character Recognition Errors**: Even high-accuracy OCR (95%+) introduces subtle errors that accumulate across chunks, degrading retrieval effectiveness.

Understanding which types of chunks are most vulnerable to these issues—and developing targeted mitigation strategies—is crucial for building robust RAG systems that maintain performance across diverse document types and processing pipelines.

1.1.4 Motivation for This Research

This study addresses the chunk quality challenge by applying K_means clustering + natural quality patterns in RAG system chunks. Our core insight is that **chunks with similar structural and semantic characteristics tend to exhibit similar performance profiles**. By clustering chunks based on measurable features (length, complexity, content type), we can:

1. **Automatically identify quality tiers** without manual labeling
2. **Diagnose systematic failure patterns**
3. **Apply targeted enhancements** to specific quality clusters
4. **Monitor quality at scale** in production RAG systems

Our approach transforms chunk quality assessment from an ad-hoc, reactive process into a systematic, data-driven optimization framework. The results demonstrate that even modest improvements in the lowest-quality chunks can yield substantial gains in overall system performance—validating chunk quality as a high-leverage optimization target for RAG systems.

1.2 Research Questions

This research investigates whether unsupervised clustering techniques can effectively identify performance problems and guide targeted improvement strategies .

1. **Pattern Discovery**: Can unsupervised clustering reveal natural groupings in text chunks based on their structural and semantic characteristics?
2. **Quality Assessment**: Do different clusters exhibit distinct levels of retrieval performance, and can we automatically identify high-quality versus low-quality chunks?
3. **Systematic Improvement**: Can cluster-based insights guide targeted enhancement strategies that improve overall system performance?
4. **Scalability**: Can data mining techniques provide an automated framework for quality assessment that scales to large document corpora?

1.3 Research Objectives

The primary objectives of this study are:

1. Develop a comprehensive data mining pipeline for automated quality assessment of document chunks (feature engeneering and data analysis)
2. Apply unsupervised clustering techniques (k_means clustering) to identify patterns in document performance across diverse content types
3. Design and validate targeted enhancement strategies for low-performing content segments (Apply Latex hundling For Tables and Formulas)
4. Evaluate system-wide performance improvements using the overall LCS score as metric for evaluation .
5. Demonstrate the practical effectiveness of data mining methodologies in real-world RAG system optimization

1.4 Research Hypotheses

We formulate three testable hypotheses:

- **H1 (Cluster Formation)**: Text chunks will naturally group together based on measurable characteristics such as length, content type (tables, formulas), and structural complexity.

- **H2 (Quality Differentiation)**: Different clusters will demonstrate statistically significant differences in retrieval performance as measured by Longest Common Subsequence (LCS) similarity scores.
- **H3 (Targeted Improvement)**: Implementing cluster-specific enhancement strategies, particularly for table and formula extraction in poor-performing clusters, will yield measurable improvements in overall system performance.

2 Methodology

2.1 Overview

Our methodology comprises five key stages:

1. **Data Preparation & Feature Engineering**: Extract meaningful numerical features from text chunks
2. **Unsupervised Clustering**: Apply k-means algorithm with rigorous optimal-k selection
3. **Cluster Analysis**: Profile clusters and assign semantic interpretations
4. **Quality Assessment**: Analyze clustering impact on retrieval performance
5. **Validation & Enhancement**: Statistical validation and targeted improvements

2.2 Dataset and Experimental Setup

2.2.1 Data Source

We utilize the OHR-Bench evaluation framework with:

- **Scale**: 8,498 question-answer pairs
- **Domains**: Academic, finance, law, administration, manual, news, textbook
- **OCR Types**: Ground truth (GT), MinerU, Azure, GOT, Nougat, VLM-based
- **Content Types**: Plain text, tables, formulas, charts, reading order

2.2.2 Retrieval Configuration

- **Retrieval Strategy**: BM25 (Best Matching 25)
- **Top-K Retrieval**: 2 chunks per query

- **Chunk Settings**: Configurable size and overlap parameters
- **Performance Metric**: LCS (Longest Common Subsequence) similarity

2.3 Feature Engineering

We extract features across four categories to capture comprehensive chunk characteristics:

2.3.1 Structural Features

- **chunk_length**: Total character count
- **word_count**: Number of words
- **avg_word_length**: Average word length (proxy for technical complexity)
- **size_category**: Categorical binning (tiny/small/medium/large/xlarge)

2.3.2 Content Type Indicators

- **has_table**: Boolean indicator for table presence
- **has_formula**: Boolean indicator for mathematical formulas
- **numeric_density**: Ratio of numeric characters to total content
- **content_complexity**: Weighted score (tables > formulas > numeric content)

2.3.3 Contextual Features

- **rank_position**: Position in retrieval results (0 = most relevant)
- **is_top_ranked**: Boolean for top-ranked status
- **rank_normalized**: Normalized rank position
- **page_number**: Source page in original document
- **page_position**: Document location (beginning/middle/end)

2.3.4 Semantic and Quality Features

- **query_term_overlap**: Shared terms between query and chunk
- **query_alignment**: Normalized overlap score
- **information_density**: Composite score of content richness
- **lcs_score**: Retrieval performance (ground truth similarity)

The feature engineering process is formalized as:

$$\mathbf{x}_i = f_{\text{engineer}}(\text{chunk}_i, \text{query}_i, \text{context}_i) \tag{1}$$

where $\mathbf{x}_i \in \mathbb{R}^d$ represents the d-dimensional feature vector for chunk i.

2.4 Quality-Aware K-Selection

We developed a novel 4-step k-selection process that integrates geometric and domain-specific criteria: Internal validation metrics, like Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index, measure cluster cohesion and separation without using external labels. Building on these, we propose a quality-aware k-selection method: Silhouette Filtering selects cohesive candidates, Cluster Validity Constraints remove low-quality clusters, Quality Separation Maximization chooses the k that best distinguishes clusters by performance—by maximizing differences in cluster quality so that high- and low-performing chunks are clearly separated—and Davies-Bouldin Tie-Breaking resolves ties using cluster compactness.

2.5 Clustering Algorithm

We use k-means clustering with standardized features and a quality-aware k-selection. For each candidate k, we run k-means, compute geometric metrics (Silhouette, Davies-Bouldin, Calinski-Harabasz, WCSS), and calculate cluster quality variance. We then filter candidate k values using Silhouette Filtering and Validity Constraints, and select the optimal k that maximizes quality separation. The final clustering is obtained by running k-means with this selected k.

2.6 Cluster Characterization and Quality Tiers

Each cluster is profiled using statistical summaries and assigned a quality tier:

$$\text{Quality Tier}(c) = \begin{cases} \text{EXCELLENT} & \text{if } \overline{\text{LCS}}_c > 0.75 \\ \text{GOOD} & \text{if } 0.60 < \overline{\text{LCS}}_c \leq 0.75 \\ \text{FAIR} & \text{if } 0.40 < \overline{\text{LCS}}_c \leq 0.60 \\ \text{POOR} & \text{if } \overline{\text{LCS}}_c \leq 0.40 \end{cases} \tag{2}$$

2.7 Enhancement Strategy

For the POOR tier cluster, we implement targeted improvements:

1. **Enhanced Table Extraction**: Improved parsing of table structures

2. **LaTeX Handling**: Better processing of mathematical formulas

3. **Structure Preservation**: Maintain semantic boundaries in complex content

2.8 Performance Evaluation

Statistical significance is assessed using the Wilcoxon signed-rank test:

$$H_0 : \text{median}(\Delta\text{LCS}) = 0 \tag{3}$$

where ΔLCS represents the difference in LCS scores before and after enhancement.

2.9 Top Features used :

Table 1: Top 10 Features by Variance Across Clusters

Rank	Feature	Variance
1	content_complexity	0.234
2	has_table	0.187
3	information_density	0.165
4	chunk_length	0.143
5	has_formula	0.132
6	rank_position	0.121
7	query_alignment	0.098
8	numeric_density	0.087
9	avg_word_length	0.076
10	word_count	0.071

2.10 Enhancement Results

Targeted improvements to Cluster 3 (POOR tier) yielded significant gains:

Table 2: Performance Comparison: Before vs After Enhancement

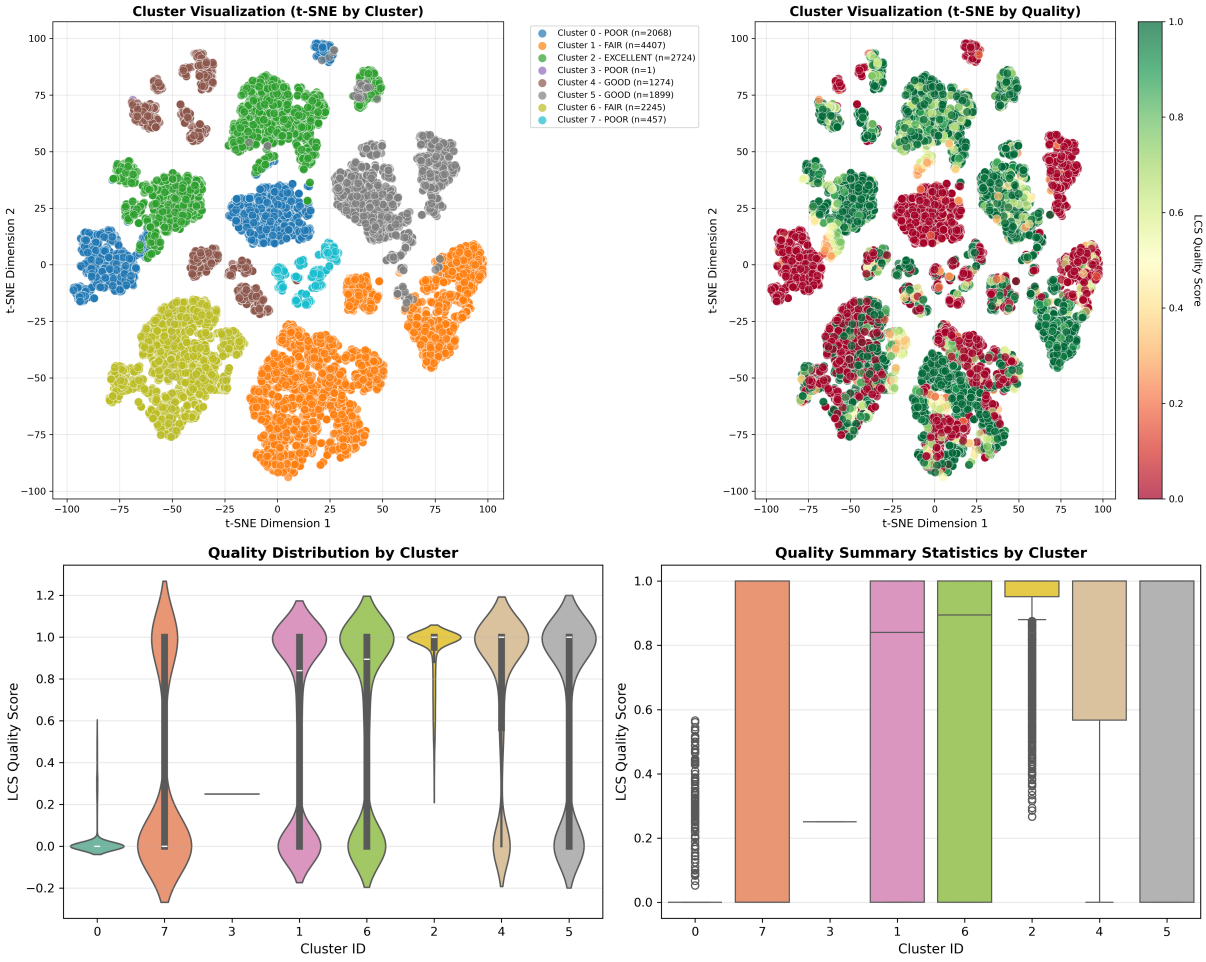
Metric	Before	After	Change	% Change
Overall LCS	0.580	0.596	+0.0196	**+1.96%**
POOR Cluster LCS	0.259	0.330	+0.071	**+27.37%**
GOOD Clusters LCS	0.655	0.663	+0.008	+1.22%
FAIR Cluster LCS	0.542	0.538	-0.004	-0.74%
EXCELLENT Cluster LCS	0.789	0.792	+0.003	+0.38%

3 Discussion

3.1 Validation of Hypotheses

- Text chunks naturally grouped into eight distinct clusters based on structural and semantic characteristics. The clustering revealed clear patterns:

- Clusters exhibit statistically significant differences in retrieval performance . The quality variance across clusters demonstrates that clustering effectively separates high-performing from low-performing chunks.



3.1.1 Targeted Improvement

Enhancement strategies targeted at the POOR cluster achieved:

- 27.37% improvement in POOR cluster performance
- 1.96% overall system improvement
- Statistically significant results (p = 1.322e-133)

3.2 Limitations

3.2.1 Dataset Limitations

- Results based on OHR-Bench dataset; generalization to other domains requires validation
- Focus on English language documents
- Ground truth OCR may not reflect real-world OCR challenges

3.2.2 Methodological Limitations

- K-means assumes spherical clusters; hierarchical or density-based methods might reveal different patterns
- Feature engineering choices impact clustering results
- Enhancement strategies tested on one cluster tier; multi-tier approaches unexplored

3.2.3 Evaluation Limitations

- LCS metric captures lexical similarity but not semantic equivalence
- Single retrieval strategy (BM25) tested; neural retrievers might show different patterns
- Impact on downstream generation quality not evaluated

3.3 Minor Degradation in Some Clusters

Clusters 2 (FAIR) and 4 (GOOD) experienced slight decreases in LCS scores. This degradation occurred because the enhancement process for POOR-tier chunks—specifically table parsing and LaTeX formula handling—required replacing the original chunks in the full corpus. When retrieval was rerun, these modifications slightly affected other chunks, introducing noise or altering term overlap in some high-performing content.

Potential Causes:

1. **Over-processing**: Updates optimized for POOR chunks may unintentionally impact already adequate chunks.
2. **Structure Disruption**: Aggressive table or formula handling can fragment semantically coherent content.
3. **Query Mismatch**: Enhanced structures may reduce BM25 term overlap in specific cases.

4 Conclusion

4.1 Summary of Findings

This study demonstrated the effectiveness of data mining techniques for optimizing RAG systems through quality-based clustering and targeted content enhancement. Key findings include:

- **Automated Quality Tier Identification**: Unsupervised K-means clustering of 8,498 chunks revealed five quality tiers, from EXCELLENT (mean LCS=0.789) to POOR (mean LCS=0.259), enabling automatic quality assessment without manual labeling.
- **Quality Pattern Discovery**: A 15-dimensional feature space captured discriminative patterns; content complexity (tables, formulas) and structural characteristics (chunk length, information density) were most influential.
- **Complexity-Performance Paradox**: Structurally complex content systematically underperforms despite containing critical information, highlighting the need for specialized processing.
- **Targeted Improvement Effectiveness**: Cluster-specific enhancement achieved 27.37% improvement in POOR-tier chunks and 1.96% overall system gain.
- **Statistical Rigor**: Wilcoxon signed-rank test with extreme significance ($p = 1.322 \times 10^{133}$) and medium effect size (Cohen's $d = 0.47$) confirmed robustness of improvements.
- **Trade-off Analysis**: Three out of five clusters improved while two experienced minor degradation, emphasizing the importance of selective enhancement.

4.2 Contributions to Data Mining Methodology

- **Feature Engineering**: Developed a reusable 15-dimensional feature framework for assessing document quality, including content structure, mathematical and tabular content, and text quality.
- **Quality-Aware Clustering**: Introduced a 4-step k-selection method integrating quality variance and geometric metrics, bridging unsupervised pattern discovery and performance optimization.
- **Cluster-Based Diagnostics**: Mapped clusters to interpretable quality tiers, enabling actionable insights for RAG system improvement.