# TinyLlama Embeddings on *Me Before You* and Wikipedia Corpus

## 1. Experiment Setup

**Book:** *Me Before You* (Romance)
**Corpus Size:** Middle 2000 sentences extracted
**Model:** TinyLlama
**Total Layers:** 23
**Layers Analyzed:** Layer 14 (2/3rd), Layer 22 (Last)

## 2. Clustering Results: Me Before You

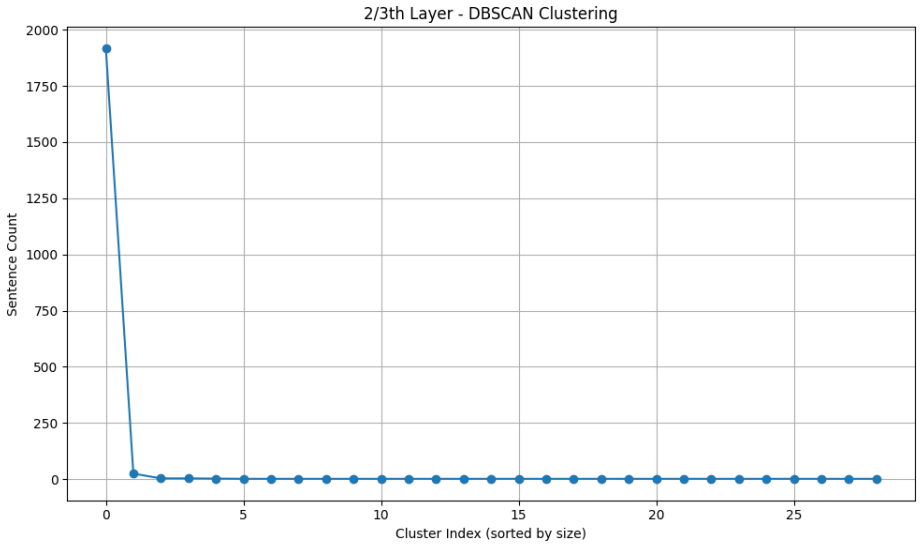
### 2.1 DBSCAN (cosine, $\epsilon = 0.1$)

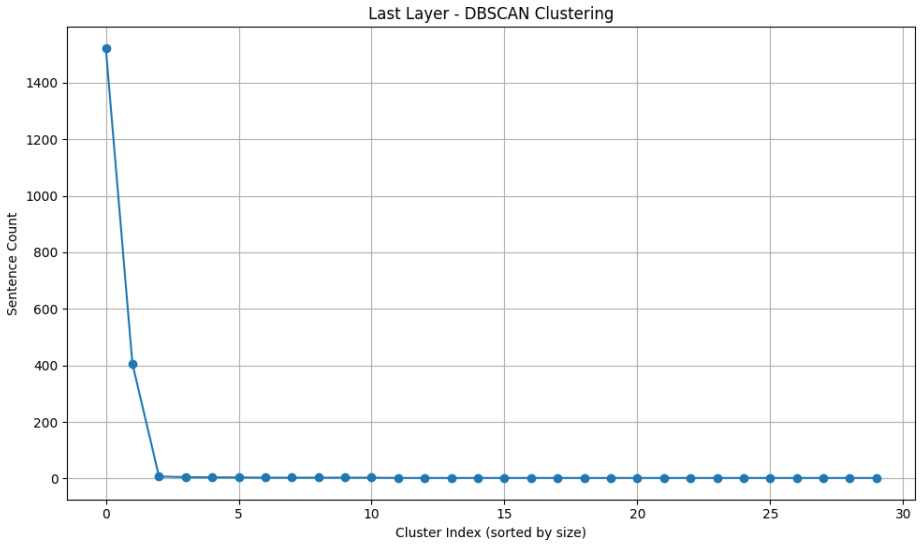
**Cluster Counts:**
Layer 14: 29
Layer 22: 30

**Sentence Counts (sorted):**

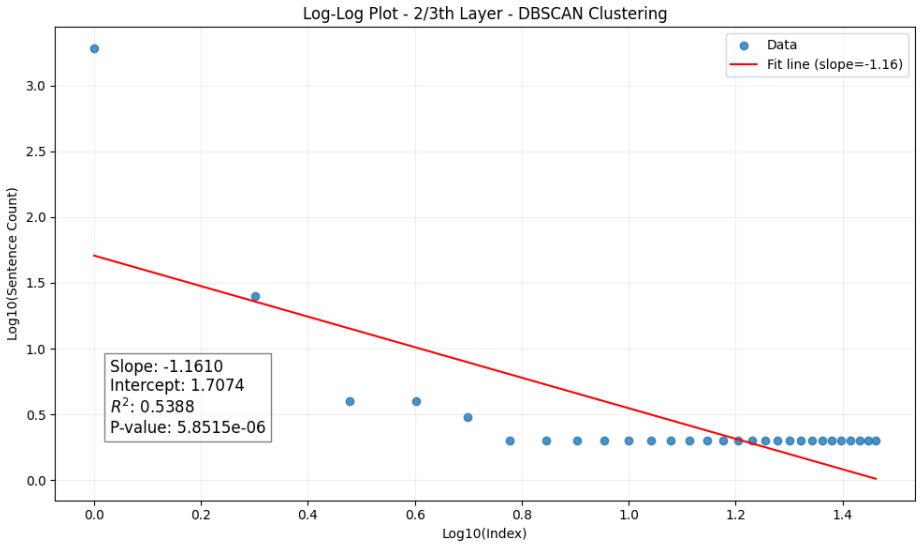
- Layer 14: `[1916, 25, 4, 4, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2]`
- Layer 22: `[1521, 406, 7, 5, 4, 4, 3, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2]`

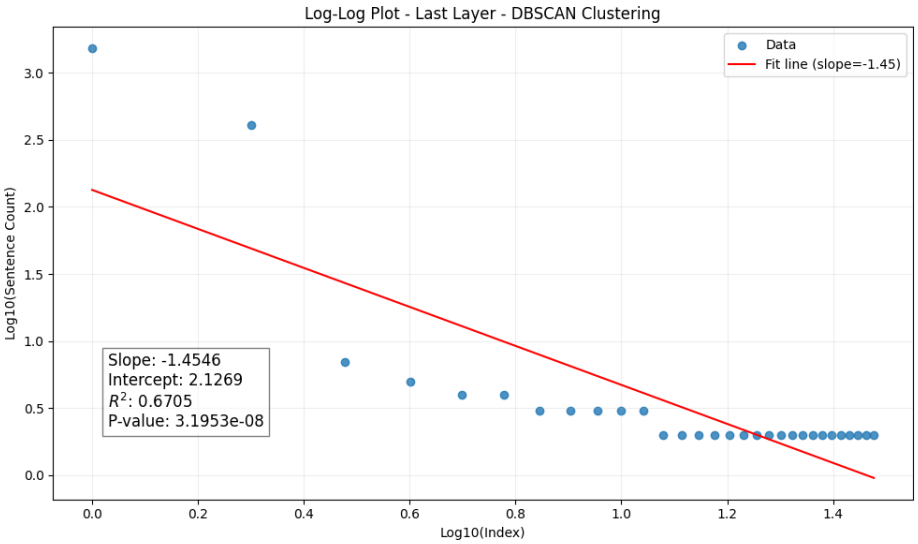
**Sentence Count per Cluster (DBSCAN - 2/3th and Last Layers)**






## Log-Log Plot (DBSCAN - 2/3rd and Last Layers)





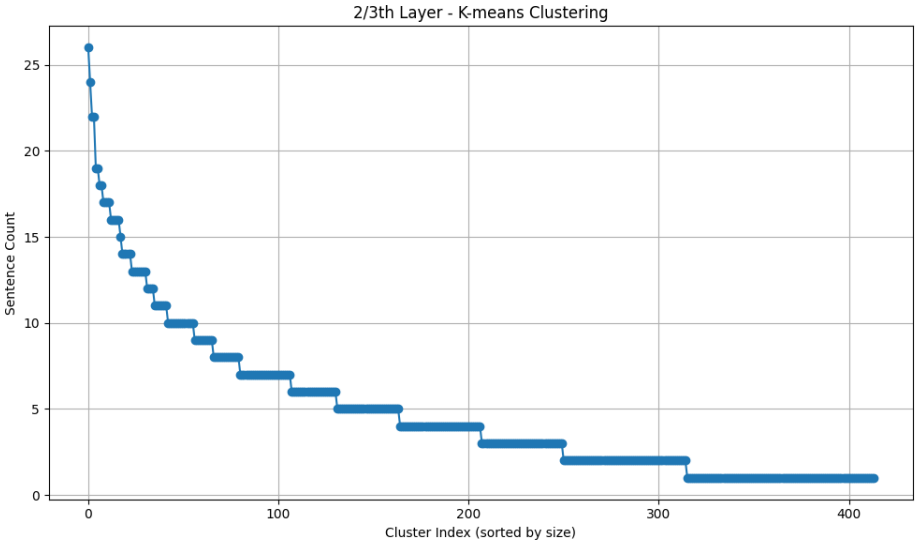

## 2.2 KMeans (knee detection)

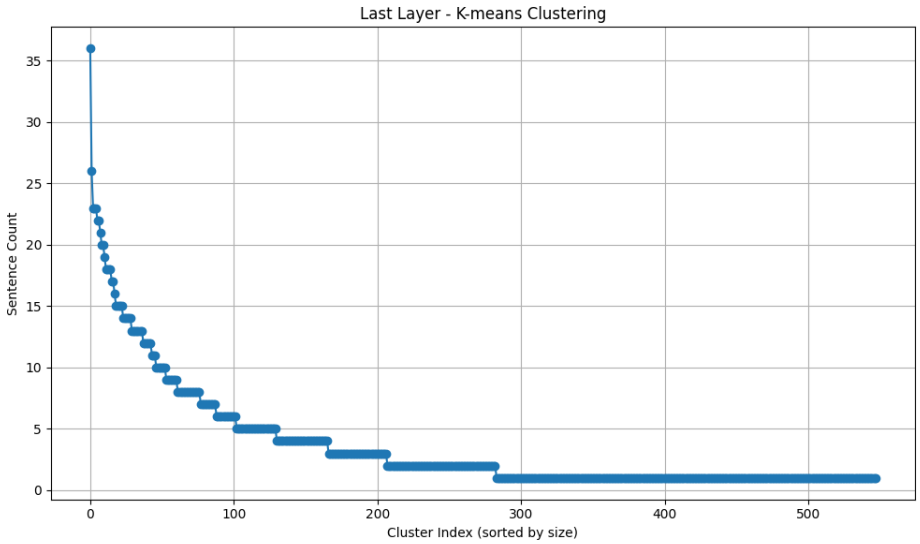
**Cluster Counts:**
Layer 14: 414 clusters, WCSS=484.243
Layer 22: 548 clusters, WCSS=142.662

**Sentence Counts (sorted):**

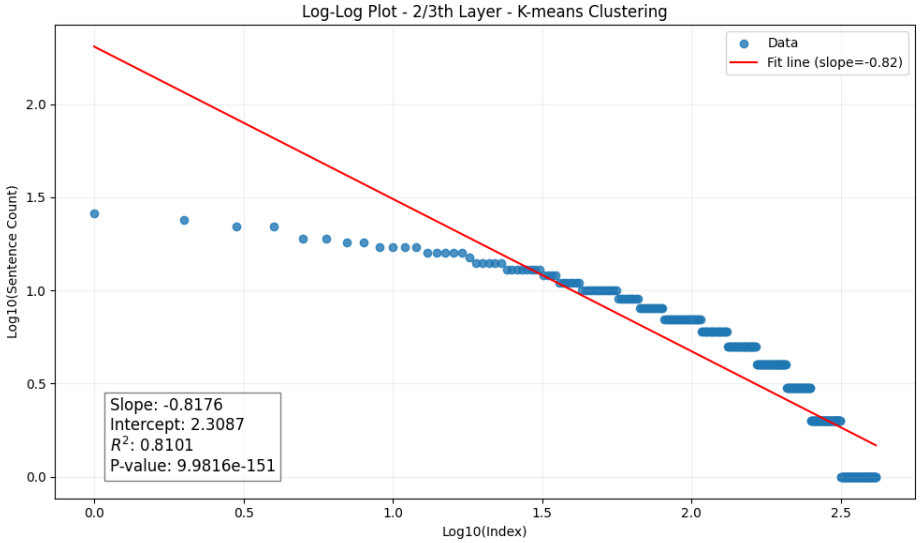
- Layer 14: `[26, 24, 22, 22, 19, 19, 18, 18, 17, 17, 17, 17, 16, 16, 16, 16, 16, 15, 14, 14, 14, 14, 14, 13, 13, 13, 13, 13, 13, 13, 13, 12, 12, 12, 12, 11, 11, 11, 11, 11, 11, 11, 10, 10, 10, 10, ..., 1]`
- Layer 22: `[36, 26, 23, 23, 23, 22, 22, 21, 20, 20, 19, 18, 18, 18, 18, 17, 17, 16, 15, 15, 15, 15, 15, 14, 14, 14, 14, 14, 14, 13, 13, 13, 13, 13, 13, 13, 13, 12, 12, 12, 12,..., 1]`

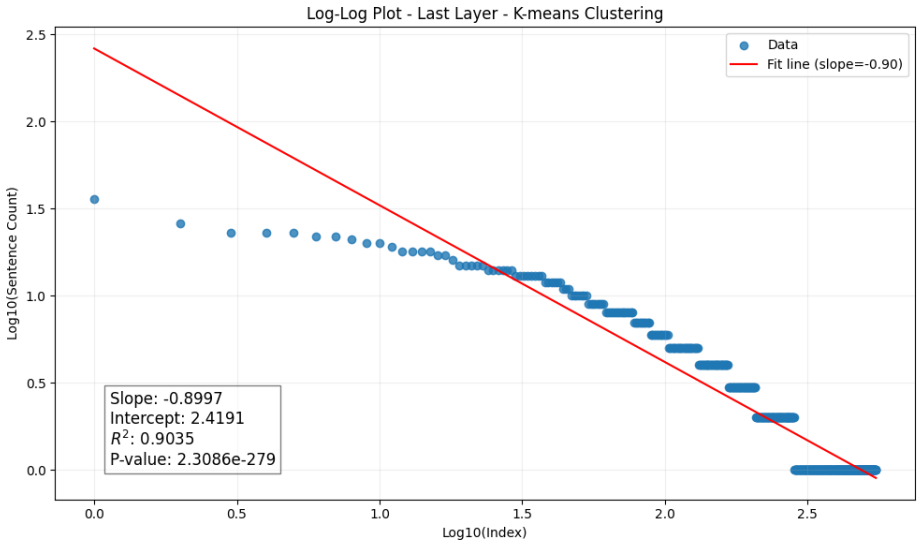
**Sentence Count per Cluster (KMeans - 2/3th and Last Layers)**






## Log-Log Plot (KMeans - 2/3rd and Last Layers)





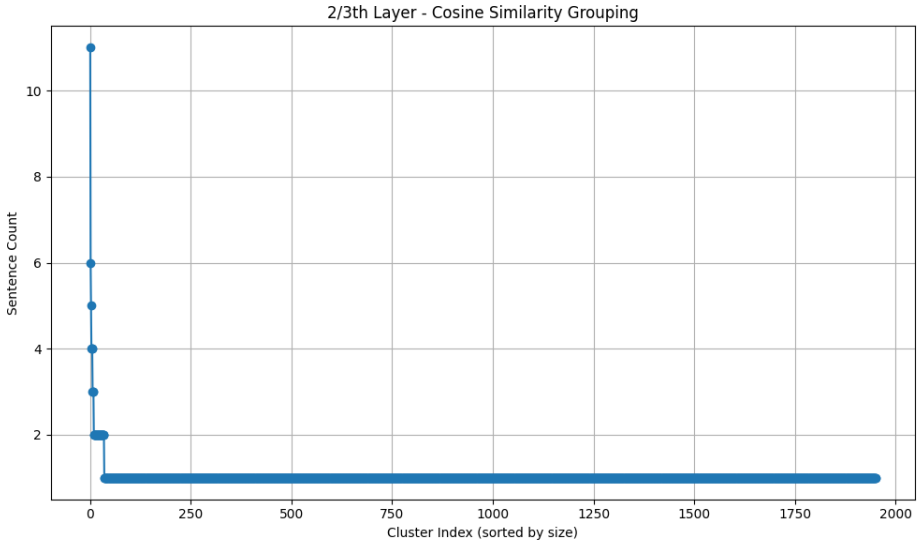

## 2.3 Cosine Similarity Grouping (threshold=0.9)

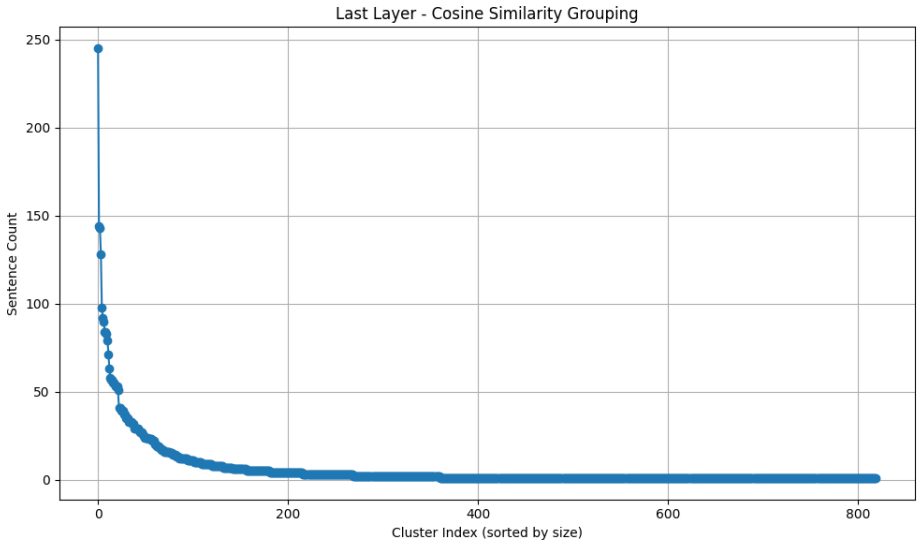
**Group Counts:**
Layer 14: 1952
Layer 22: 820

**Sample Sentence Counts:**

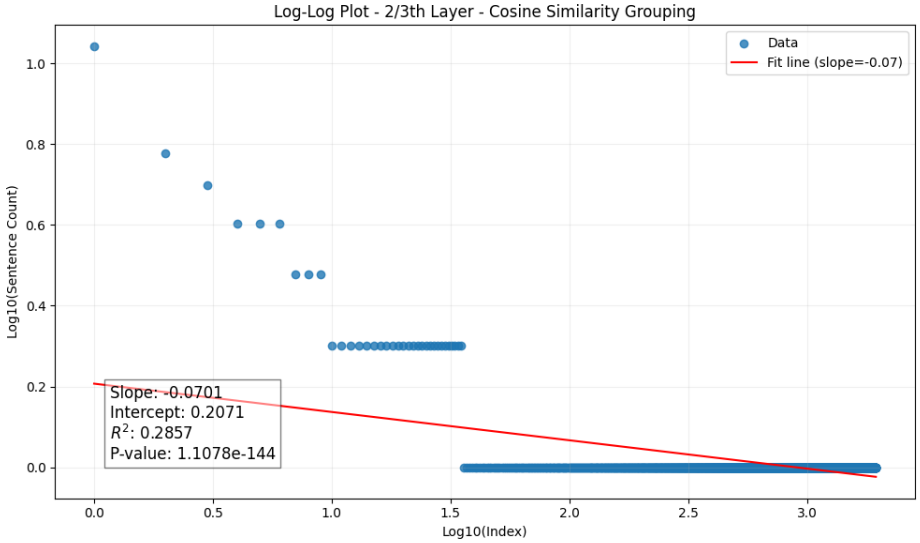
- Layer 14: `[11, 6, 5, 4, 4, 4, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, ...]`
- Layer 22: `[245, 144, 143, 128, 98, 92, 90, 84, 84, 83, 79, 71, 63, 58, 57, 57, 55, 55, 54, 53, 53, 51, 41, 41, 39, 39, 39, 37, 37, 35, 35, 35, 33, 33, 33, 33, 32, 32, 29, 29, ...]`

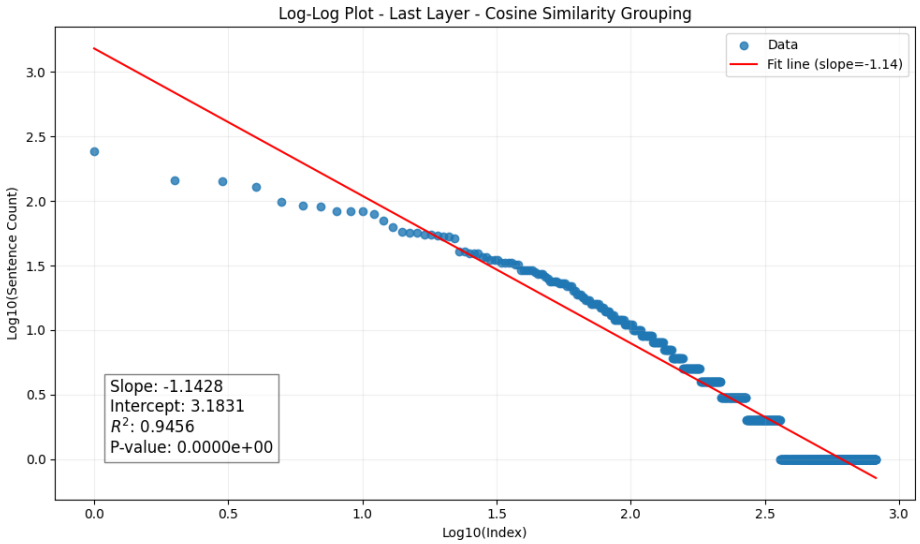
**Sentence Count per Cluster (Cosine Similarity Grouping - 2/3th and Last Layers)**






## Log-Log Plot (Cosine Similarity - 2/3rd and Last Layers)





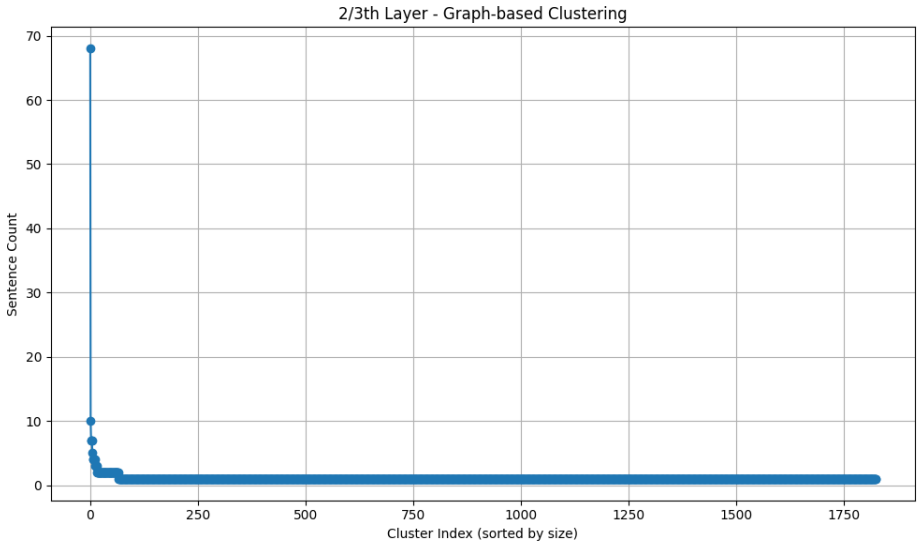

## 2.4 Graph-Based Grouping (threshold=0.85)

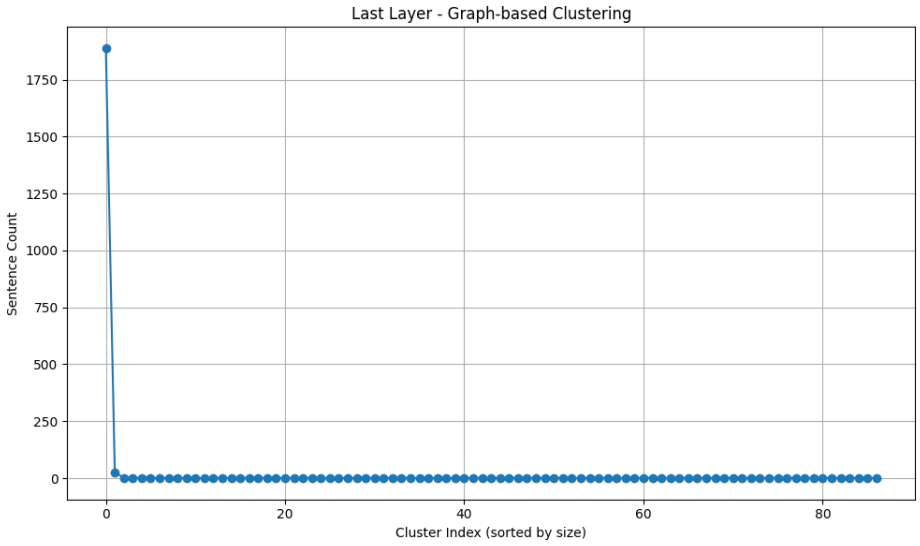
**Group Counts:**
Layer 14: 1826
Layer 22: 87

**Sample Sentence Counts:**

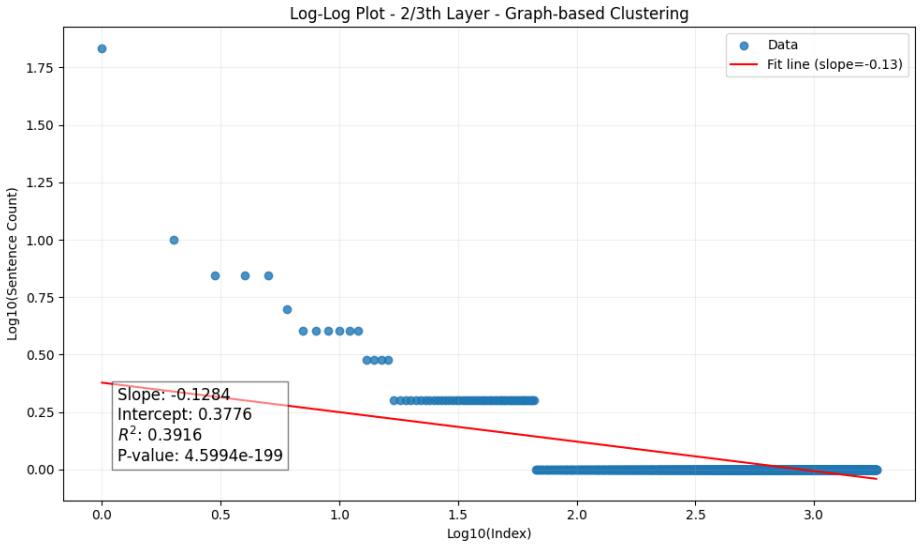
- Layer 14: `[68, 10, 7, 7, 7, 5, 4, 4, 4, 4, 4, 4, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...]`
- Layer 22: `[1887, 25, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...]`

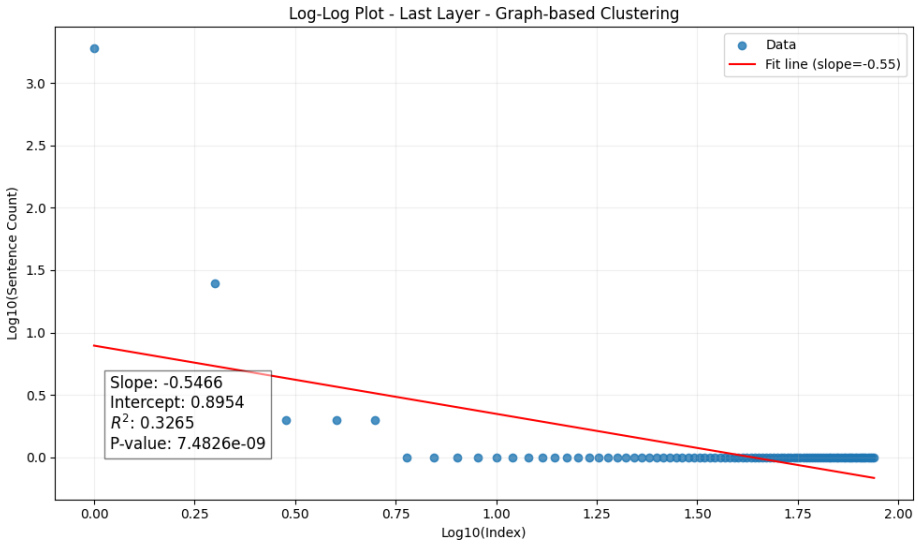
**Sentence Count per Cluster (Graph-Based Grouping - 2/3th and Last Layers)**






## Log-Log Plot (Graph-Based Grouping - 2/3rd and Last Layers)





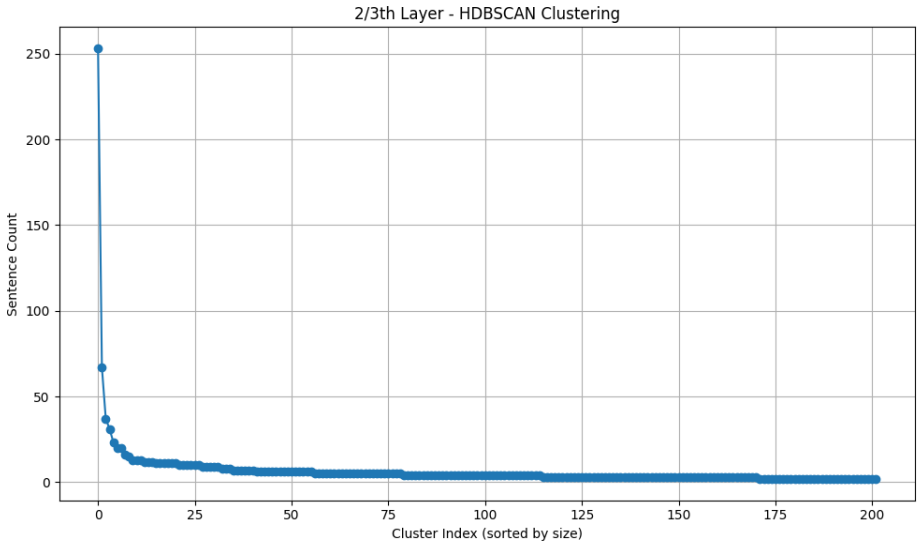

## 2.5 HDBSCAN (n_neighbors=30, n_components=20)

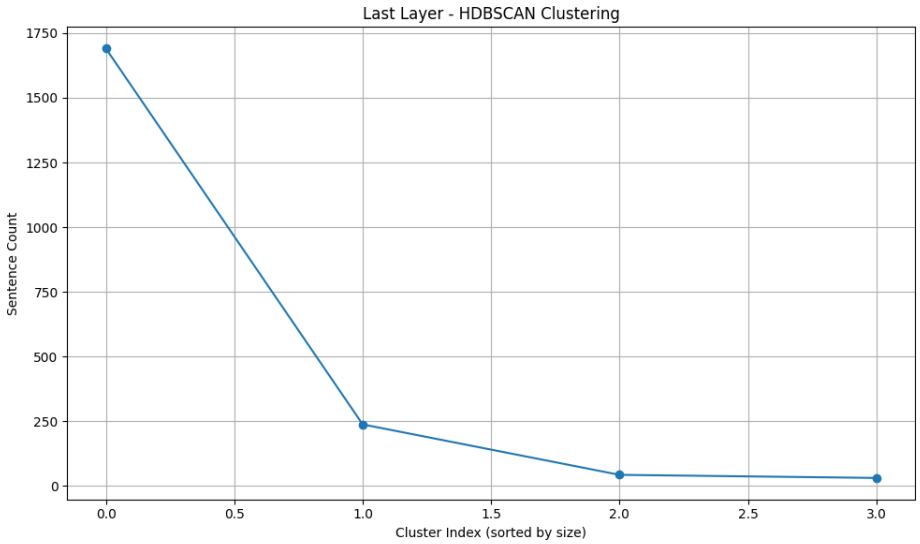
**Cluster Counts:**
Layer 14: 202
Layer 22: 4

**Sentence Counts (sample):**

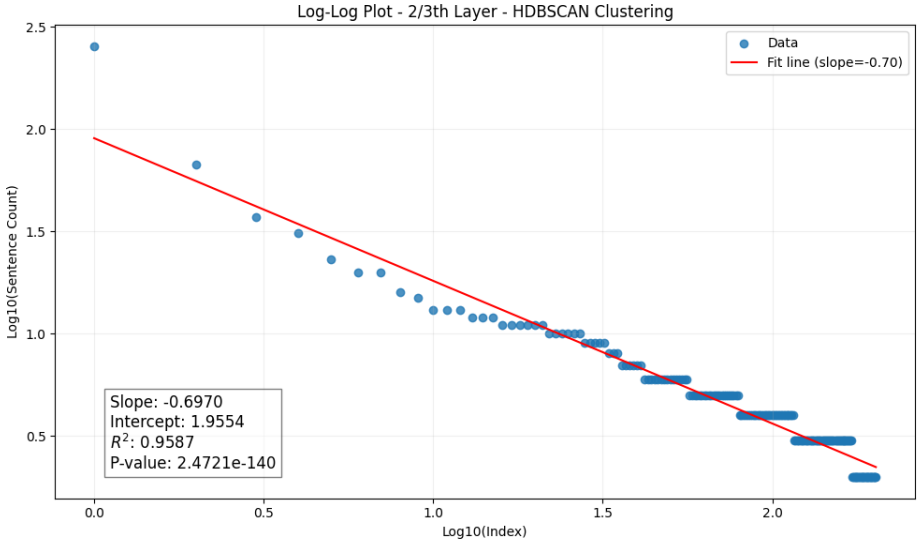
- Layer 14: `[253, 67, 37, 31, 23, 20, 20, 16, 15, 13, 13, 13, 12, 12, 12, 11, 11, 11, 11, 11, 11, 10, 10, 10, 10, 10, 10, 9, 9, 9, 9, 9, 8, 8, 8, 7, 7, 7, 7, 7, ...]`
- Layer 22: `[1691, 237, 42, 30]`

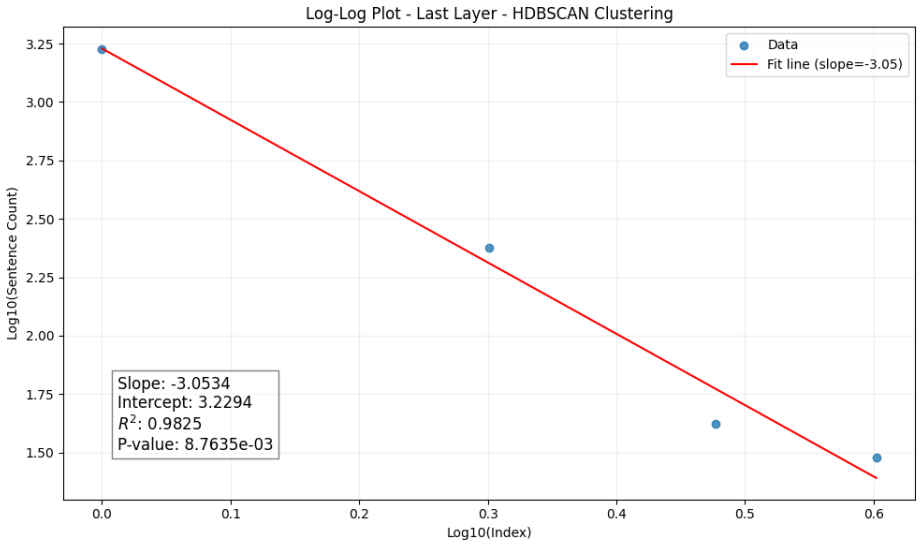
**Sentence Count per Cluster (HDBSCAN - 2/3th and Last Layers)**






## Log-Log Plot (HDBSCAN - 2/3rd and Last Layers)






# 3. Clustering Results: Wikipedia Corpus

**Dataset:** Wikipedia Sentences (Kaggle)

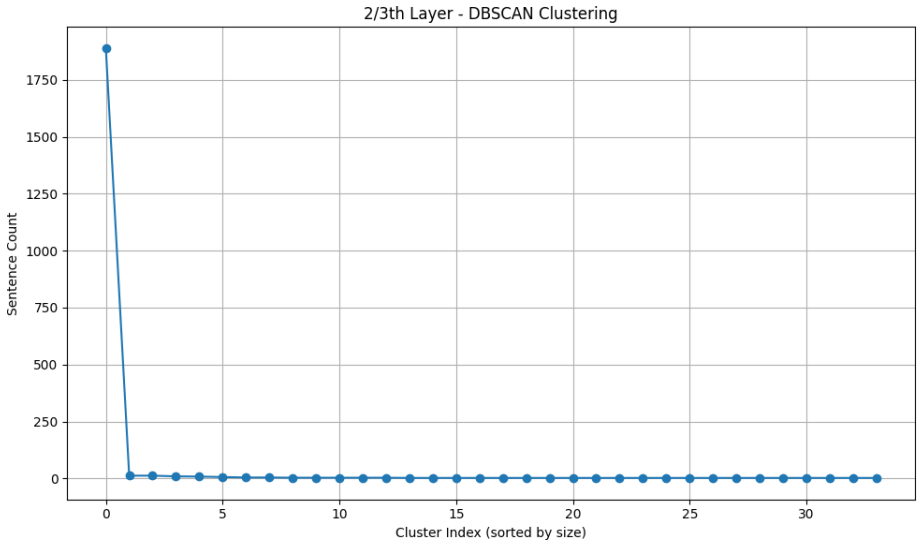
## 3.1 DBSCAN (cosine, $\epsilon = 0.1$)

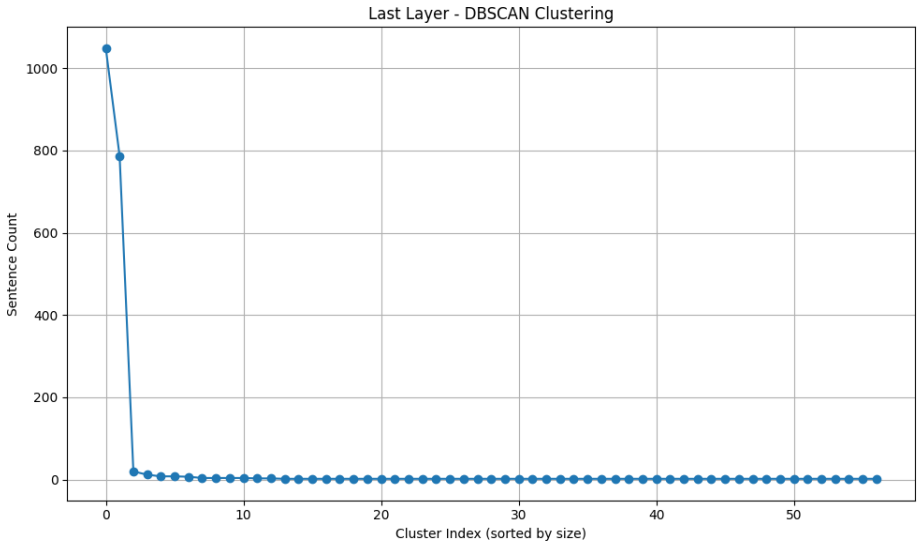
**Cluster Counts:**
Layer 14: 34
Layer 22: 57

**Sentence Counts:**

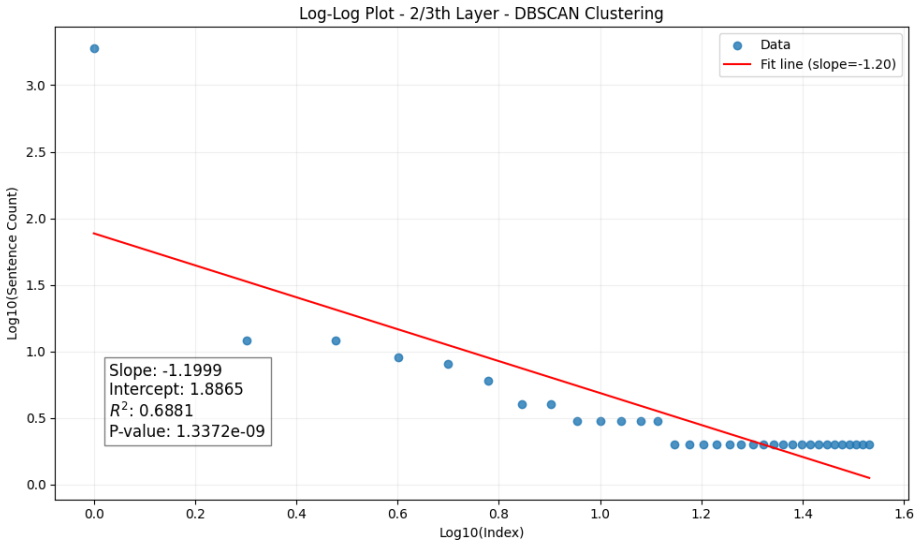
- Layer 14:
- Layer 14: `[1888, 12, 12, 9, 8, 6, 4, 4, 3, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2]`
- Layer 22: `[1048, 787, 20, 12, 8, 8, 7, 4, 4, 4, 4, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...]`

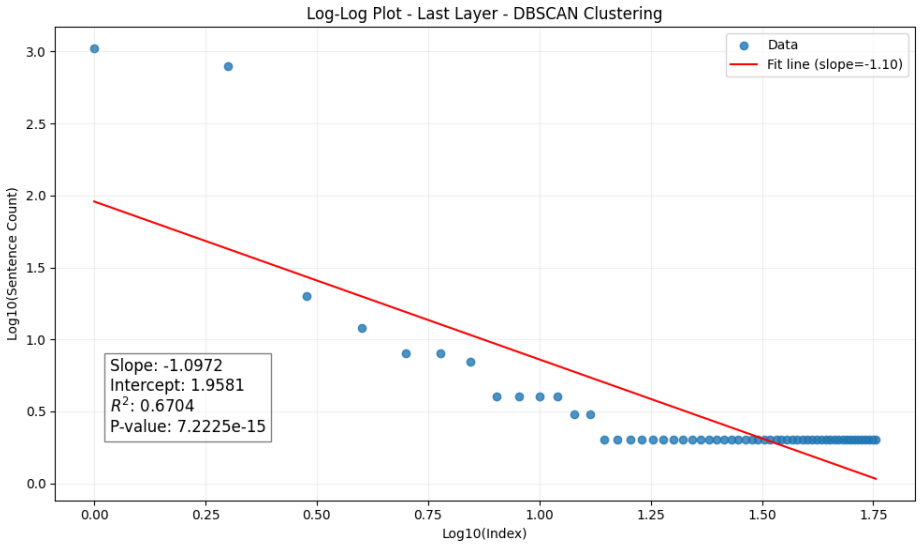
**Sentence Count per Cluster (DBSCAN - 2/3rd and Last Layers)**




**Log-Log Plot (DBSCAN - 2/3rd and Last Layers)**



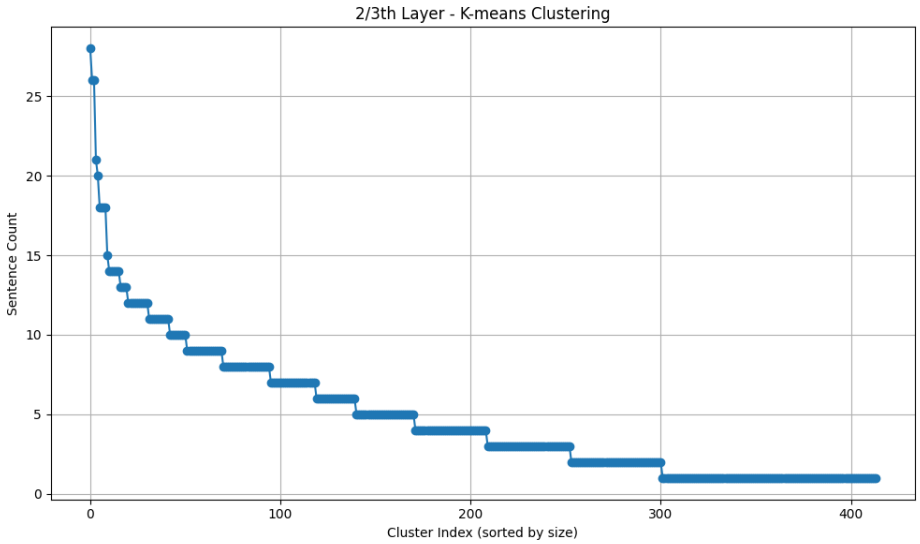

## 3.2 KMeans (knee detection)

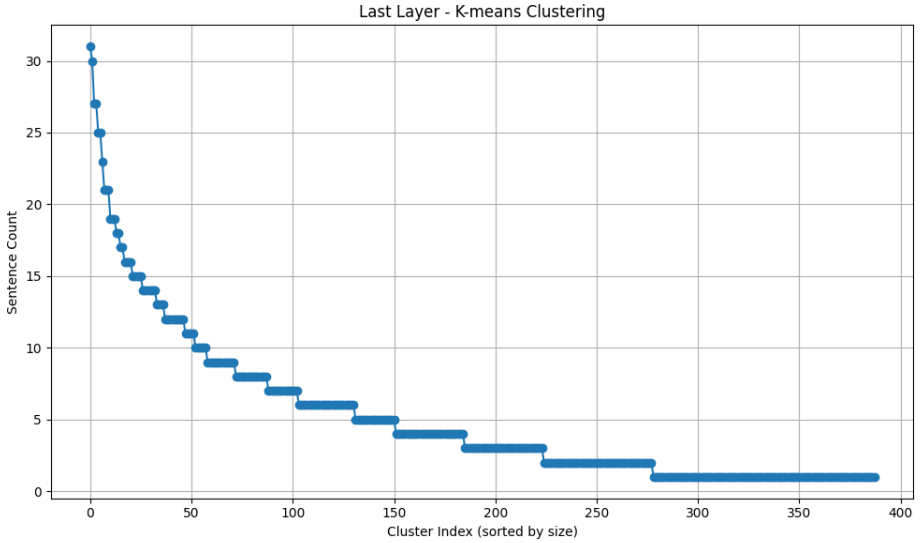
**Cluster Counts:**
Layer 14: 414, WCSS=420.938
Layer 22: 388, WCSS=226.830

**Sentence Counts (sample):**

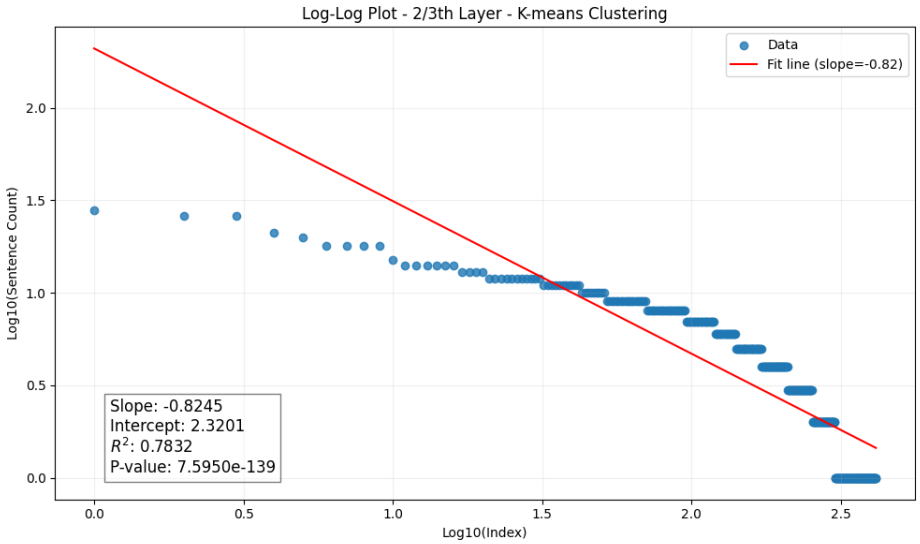
- Layer 14: `[28, 26, 26, 21, 20, 18, 18, 18, 18, 15, 14, 14, 14, 14, 14, 14, 13, 13, 13, 13, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 11, 11, 11, 11, 11, 11, 11, 11, 11, ...]`
- Layer 22: `[31, 30, 27, 27, 25, 25, 23, 21, 21, 21, 19, 19, 19, 18, 18, 17, 17, 16, 16, 16, 16, 15, 15, 15, 15, 15, 14, 14, 14, 14, 14, 14, 14, 13, 13, 13, 13, 12, 12, 12, ...]`

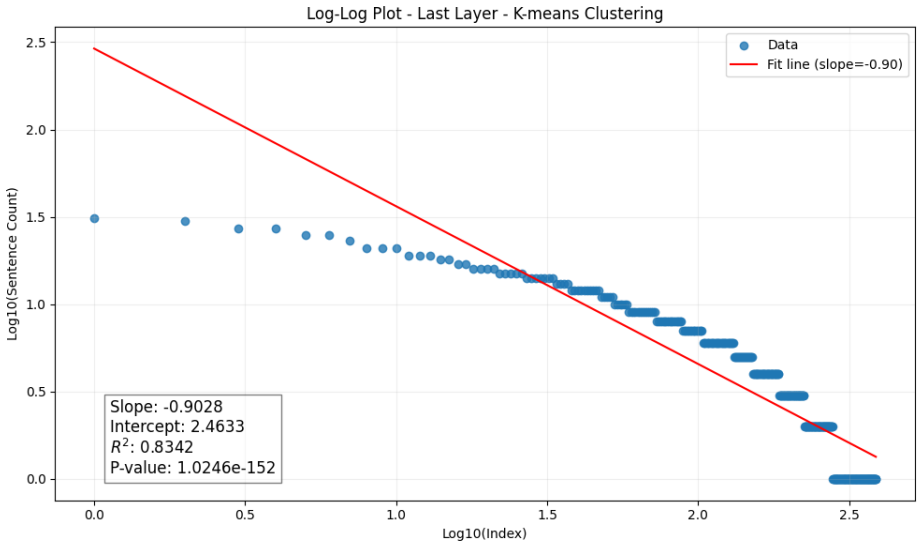
**Sentence Count per Cluster (KMeans - 2/3rd and Last Layers)**




**Log-Log Plot (KMeans - 2/3rd and Last Layers)**



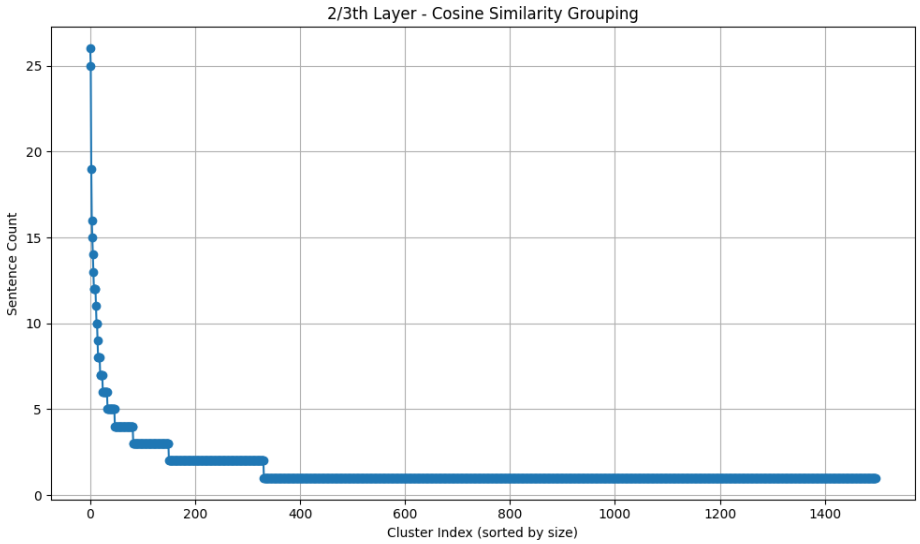

### 3.3 Cosine Similarity Grouping (threshold=0.9)

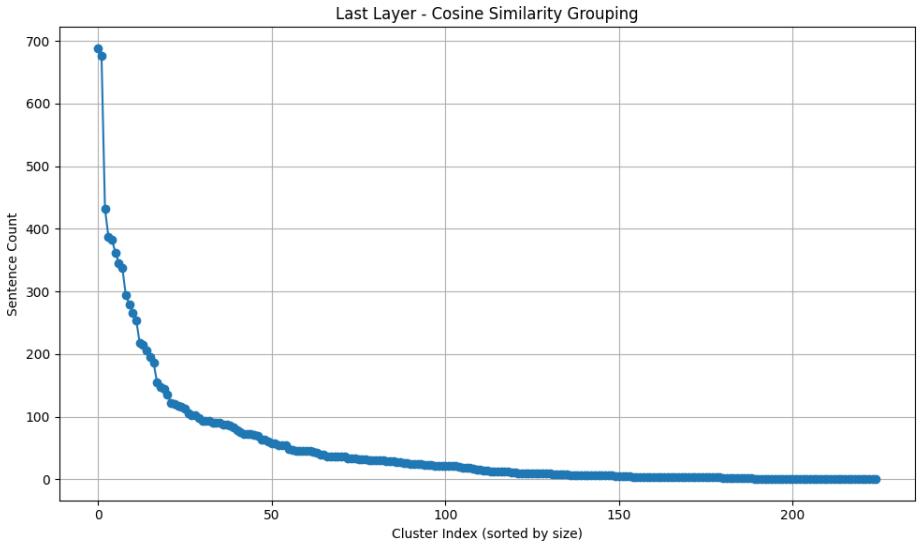
**Group Counts:**
Layer 14: 1498
Layer 22: 225

**Sentence Counts (sample):**

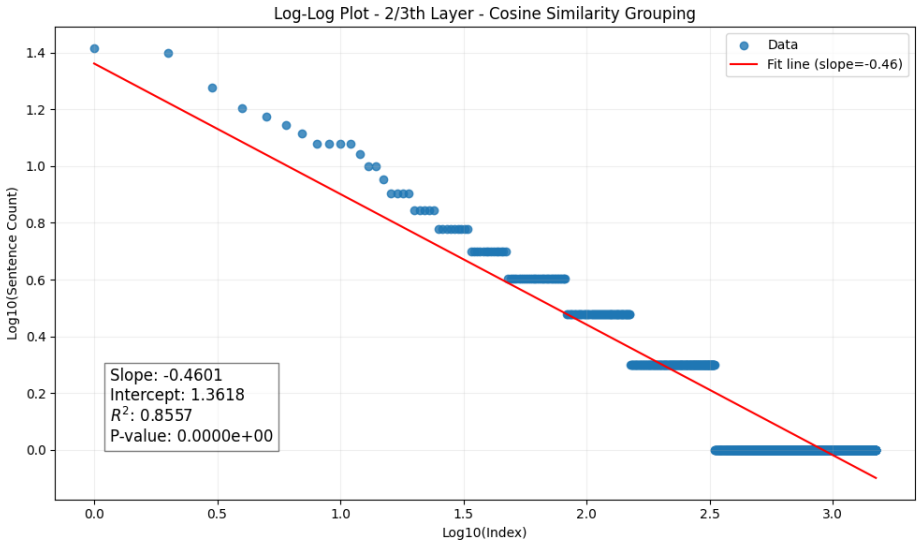
- Layer 14: `[26, 25, 19, 16, 15, 14, 13, 12, 12, 12, 12, 11, 10, 10, 9, 8, 8, 8, 8, 7, 7, 7, 7, 7, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 5, 5, 5, 5, 5, 5, ...]`
- Layer 22: `[688, 677, 432, 387, 383, 362, 346, 338, 294, 280, 266, 254, 218, 215, 206, 195, 186, 155, 147, 144, 136, 122, 120, 118, 116, 113, 105, 103, 102, 98, 93, 93, 93, 91, 90, 90, 88, 88, 86, 83, ...]`

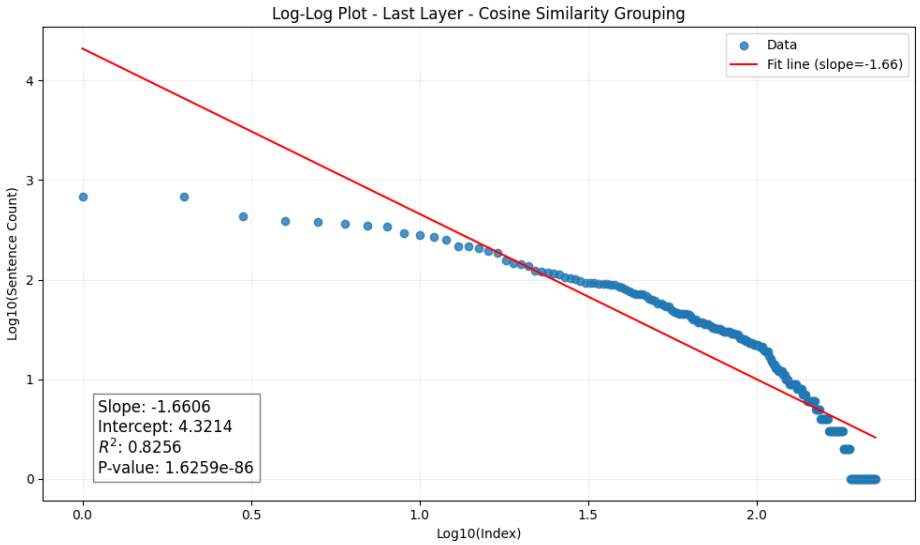
**Sentence Count per Cluster (Cosine Similarity - 2/3rd and Last Layers)**




**Log-Log Plot (Cosine Similarity - 2/3rd and Last Layers)**



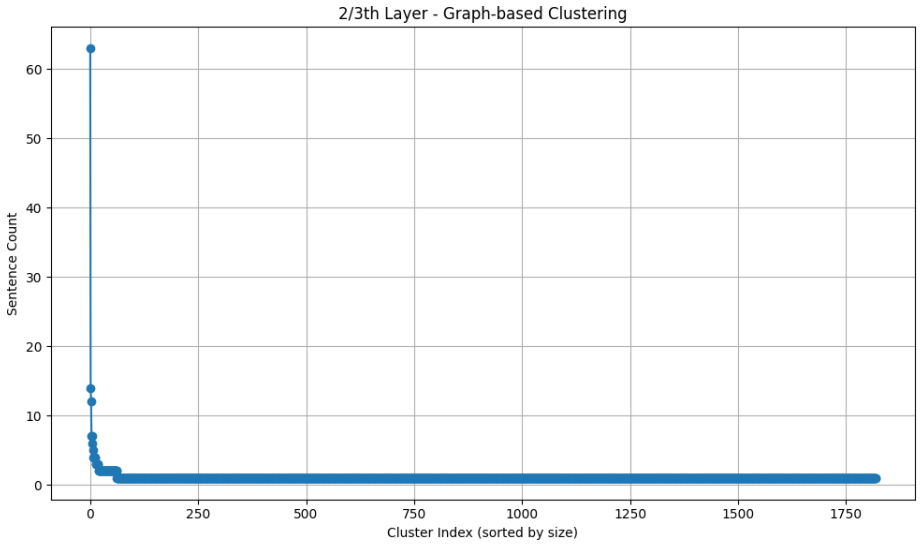

## 3.4 Graph-Based Grouping (threshold=0.85)

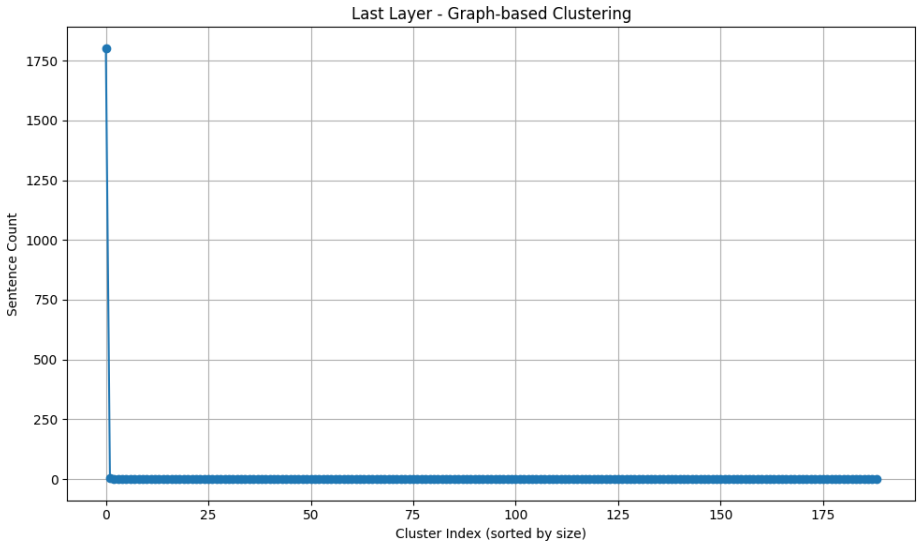
**Group Counts:**
Layer 14: 1820
Layer 22: 189

**Sentence Counts (sample):**

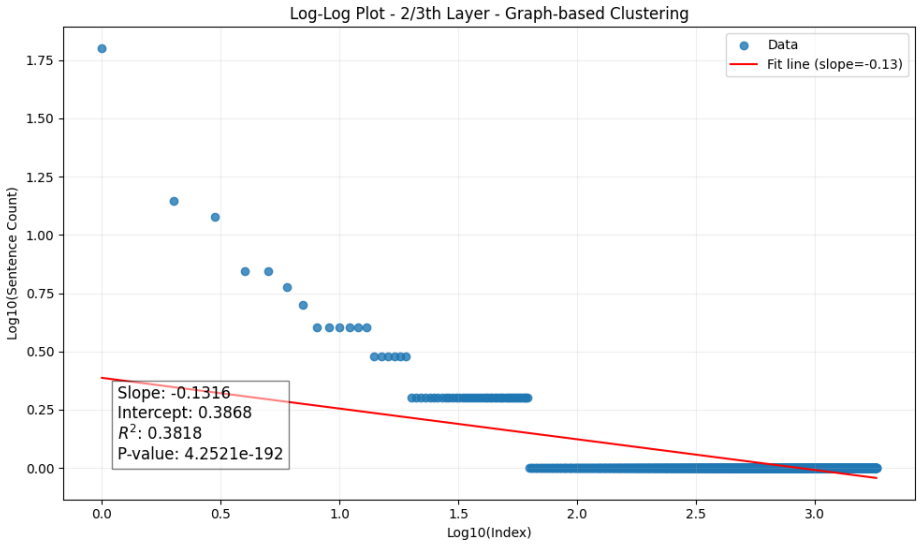
- Layer 14: `[63, 14, 12, 7, 7, 6, 5, 4, 4, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...]`
- Layer 22: `[1801, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...]`

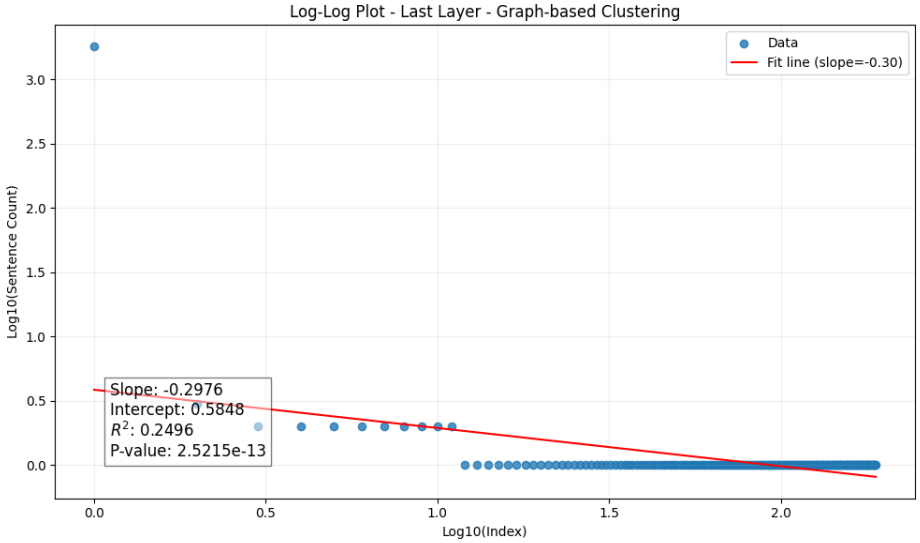
**Sentence Count per Cluster (Graph-Based Grouping - 2/3rd and Last Layers)**




**Log-Log Plot (Graph-Based Grouping - 2/3rd and Last Layers)**



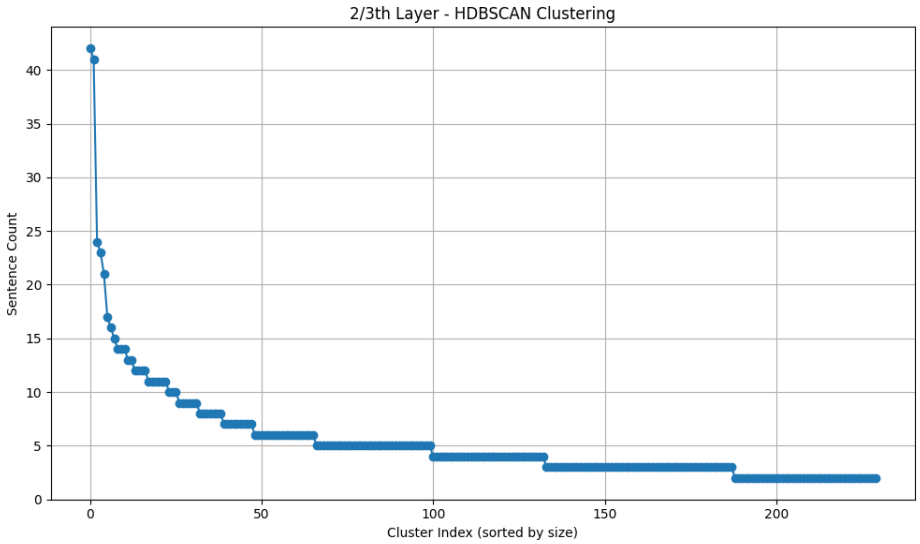

## 3.5 HDBSCAN (n_neighbors=30, n_components=20)

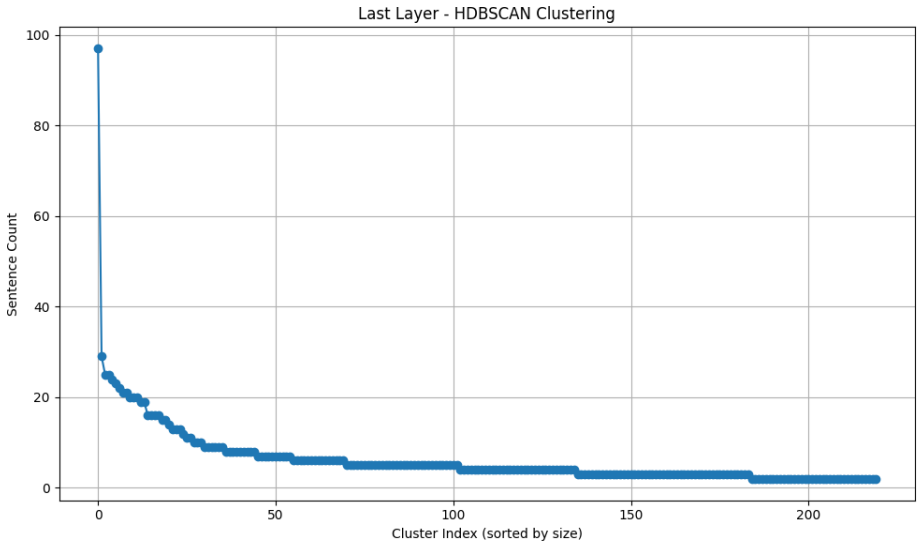
**Cluster Counts:**
Layer 14: 1243
Layer 22: 1393

**Sample Sentence Counts:**

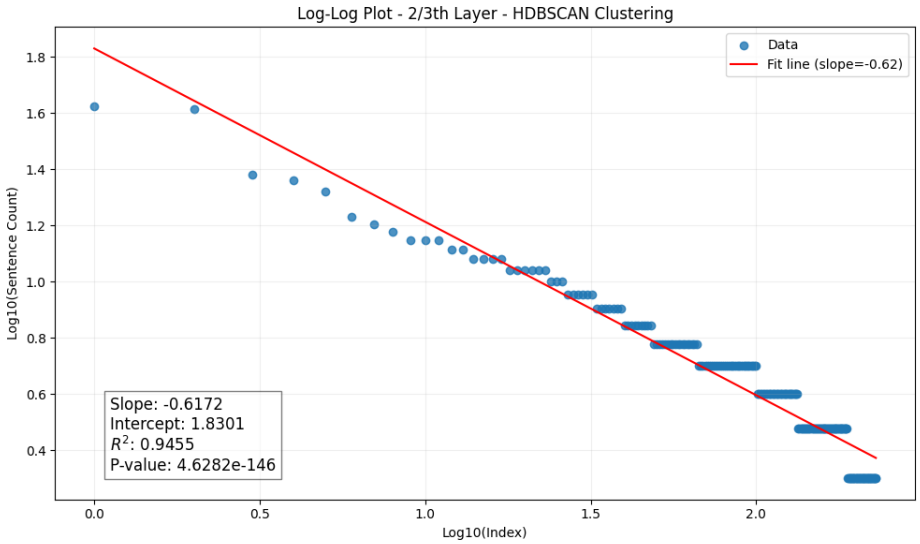
- Layer 14: `[42, 41, 24, 23, 21, 17, 16, 15, 14, 14, 14, 13, 13, 12, 12, 12, 12, 11, 11, 11, 11, 11, 11, 10, 10, 10, 9, 9, 9, 9, 9, 9, 8, 8, 8, 8, 8, 8, 8, 7, ...]`
- Layer 22: `[97, 29, 25, 25, 24, 23, 22, 21, 21, 20, 20, 20, 19, 19, 16, 16, 16, 16, 15, 15, 14, 13, 13, 13, 12, 11, 11, 10, 10, 10, 9, 9, 9, 9, 9, 9, 9, 8, 8, 8, ...]`

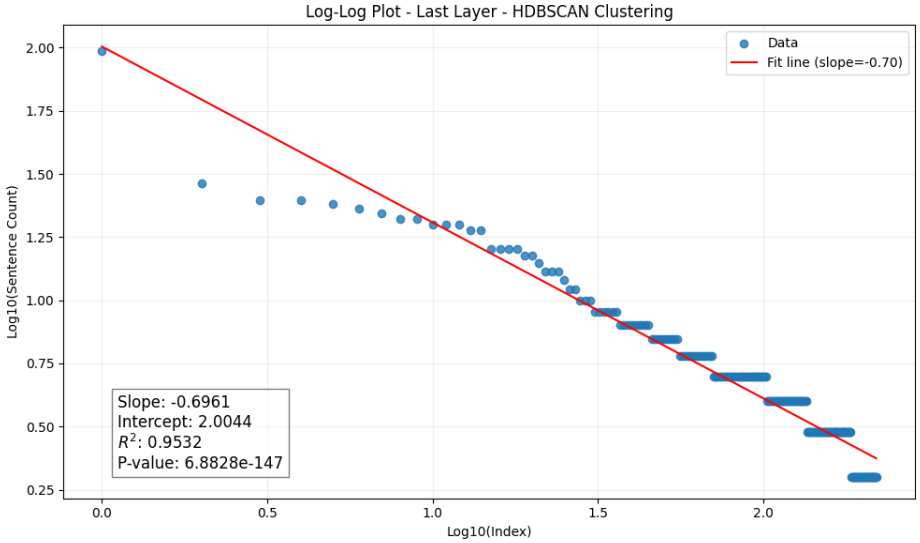
**Sentence Count per Cluster (HDBSCAN - 2/3rd and Last Layers)**




**Log-Log Plot (HDBSCAN - 2/3rd and Last Layers)**


Log-Log Plot - 2/3th Layer - HDBSCAN Clustering
Data
Fit line (slope=-0.62)
Slope: -0.6172
Intercept: 1.8301
$R^2$: 0.9455
P-value: 4.6282e-146
Log10(Sentence Count)
Log10(Index)


Log-Log Plot - Last Layer - HDBSCAN Clustering
Data
Fit line (slope=-0.70)
Slope: -0.6961
Intercept: 2.0044
$R^2$: 0.9532
P-value: 6.8828e-147
Log10(Sentence Count)
Log10(Index)