Part 3.9 — Final Recommendation: Embedding Strategy Lock-In

Objective

The objective of Part 3 was to identify the embedding configuration that produces the clearest and most separable semantic signal for GICS sector classification. Using silhouette scores with known sector labels, we quantitatively evaluated text inputs and embedding models to determine the optimal strategy for downstream semantic search and clustering.

Input Text Selection

Final Choice: SUMMARY_only

Across all seven embedding models, summary-based inputs consistently outperformed full Wikipedia content. The improvement was systematic and substantial, indicating that summaries effectively distill sector-relevant semantic information while removing noise introduced by long-form narrative content and silent truncation. Despite being significantly shorter, summaries produced more coherent and separable sector clusters.

Embedding Model Selection

Final Choice: nomic-ai/nomic-embed-text-v1.5

When paired with summarized inputs, the Nomic embedding model outperformed all generalist alternatives, achieving the highest silhouette scores in the experiment. While Nomic performed poorly on raw wiki content, its superior performance on summaries highlights its strength when provided with high signal-to-noise inputs.

Task Prefix Selection

Final Choice: classification:

Among the four Nomic task prefixes tested, the classification: prefix achieved the highest silhouette score (0.0839). This result aligns with the nature of the task, as sector identification is fundamentally a classification problem. Prefix conditioning materially altered embedding geometry, making this choice critical rather than cosmetic.

Chunking and Aggregation

Final Choice: None

Chunking strategies were unnecessary given that summarized inputs comfortably fit within model context limits. Avoiding chunking reduces pipeline complexity while preserving optimal semantic performance.

Final Locked Embedding Configuration

- **Input Text:** SUMMARY_only
- **Embedding Model:** nomic-ai/nomic-embed-text-v1.5
- **Task Prefix:** classification:
- **Chunking Strategy:** None
- **Aggregation:** N/A