**Overview of Evaluation Metrics**

We evaluated our LLM-based topic modeling approach against traditional statistical methods using multiple datasets and a comprehensive set of metrics. The LLM approach consistently creates more distinctive topic boundaries with clearer conceptual separation, stronger semantic coherence within topics, more balanced document distribution, more interpretable descriptive labels, and more comprehensive document characterization through multiple label assignments.

**Density and Distinctiveness**

Density metrics, which measure the cohesiveness of documents within topic clusters, showed mixed results across datasets. The traditional approach demonstrated higher density in two out of three datasets: MLScholar (0.1652 vs. 0.1249) and APT-Technical (0.0900 vs. 0.0839), while the LLM performed better in APT-Diverse (0.0839 vs. 0.0796). This suggests that traditional methods sometimes create slightly more cohesive document clusters.

Distinctiveness metrics, measuring separation between different topics, revealed significantly stronger performance for the LLM approach. Particularly impressive results were observed in MLScholar (0.4182 vs. 0.2124) and APT-Diverse (0.2959 vs. 0.0976). This indicates that the LLM creates more clearly differentiated topics, likely due to its stronger semantic understanding of content.

**Topic Coherence**

Our analysis revealed consistent superiority of the LLM approach across all coherence metrics in the three datasets examined. In the MLScholar dataset, LLM-based clustering achieved a CV Coherence score of 0.3564 vs. 0.3317 for the traditional approach. More substantial differences were observed in the APT-Diverse dataset, where the LLM approach demonstrated significantly better coherence under the UCI Coherence (-0.4966 vs. -0.8452). Similarly, in the APT-Technical dataset, the LLM approach outperformed the traditional method on all three coherence metrics.

The magnitude of coherence improvement varied notably across dataset types, with the APT-Diverse dataset showing the largest gains. This suggests the LLM approach is particularly effective at maintaining topic coherence when handling documents with varied formatting and reporting styles. Interestingly, while the absolute values of NPMI Coherence metrics were relatively small across all datasets (ranging from -0.0133 to -0.0593), the LLM approach consistently achieved better scores, with improvements of 0.0063, 0.0045, and 0.0096 across the three datasets respectively. The consistently superior performance on all three coherence metrics provides robust evidence that the LLM-based approach generates more semantically coherent topic clusters compared to traditional methods, regardless of domain or dataset homogeneity.

**Label Entropy and Label Quality Analysis**

While the traditional approach delivers higher density scores in two datasets (0.1652 vs 0.0501 in MLScholar and 0.0900 vs. 0.0840 in APT-Technical), the LLM method consistently achieves superior distinctiveness across all three datasets, with particularly significant margins in MLScholar (0.4721 vs. 0.2124). This pattern indicates that traditional clustering prioritizes tightly grouped documents, whereas LLM-based clustering creates more clearly delineated topic boundaries that minimize semantic overlap between categories.

The superior distinctiveness and higher label entropy values of the LLM approach translate directly to practical advantages in information management systems. Well-separated topic clusters enable more precise information retrieval, clearer navigation pathways, and reduced ambiguity when categorizing new documents. The balanced information distribution confirmed by consistently higher entropy values (3.32 vs. 2.19) in MLScholar demonstrates that the LLM method avoids creating dominant categories that could obscure meaningful distinctions, ultimately delivering a more intuitive and usable document organization system aligned with how human experts conceptualize information domains.

NOTES:  
  
**DS1** → **MLScholar**: A dataset of machine learning research papers showcasing diverse topics across the field.

**DS2** → **APT-Diverse**: A heterogeneous dataset of Advanced Persistent Threat documentation, characterized by varied formatting, technical depth, and reporting styles.

**DS3** → **APT-Technical**: A homogeneous collection of Advanced Persistent Threat analyses with consistent technical specifications and standardized reporting formats.