**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. Each metric provides insight into different aspects of topic quality and interpretability. Our analysis demonstrates that while performance varies across datasets, the LLM approach generally delivers superior results on key semantic metrics.

**Density and Distinctiveness**

Overall density measures the cohesiveness of documents within each topic cluster. Higher density values indicate that documents assigned to the same topic are more similar to each other. Across our three datasets, the traditional approach showed higher density in MLScholar (0.1652 vs. 0.1249) and APT-Technical (0.0900 vs. 0.0839), while the LLM approach performed better in APT-Diverse (0.0839 vs. 0.0796).

Overall distinctiveness, which measures how well-separated different topics are from one another, reveals a more favorable picture for the LLM approach. The LLM model produced significantly more distinctive topics in two datasets, with particularly strong performance in MLScholar (0.4182 vs. 0.2124) and APT-Diverse (0.2959 vs. 0.0976). This suggests that while traditional methods may create slightly more cohesive clusters, LLM-generated topics are more clearly differentiated from each other, likely due to the model's stronger semantic understanding.

**Topic Coherence**

Topic coherence measures the semantic relatedness of words within each topic, with higher values indicating more interpretable topics. To measure this, we used UCI coherence, which is based on pointwise mutual information between word pairs. It is particularly valuable as it measures co-occurrence patterns that reflect meaningful semantic relationships.

The LLM approach demonstrated superior performance in UCI coherence across all datasets. The higher scores of the LLM approach on this metric suggests that it produces topics with stronger semantic integrity and interpretability.

**Label Entropy and Label Quality Analysis**

Label entropy, which measures the balance of document distribution across topics, was consistently higher for the LLM approach across all datasets. In APT-Diverse, the LLM model achieved an entropy of 2.9537 compared to 2.2125 for the traditional approach. This indicates that the LLM model creates a more balanced document distribution, preventing over-concentration in a few dominant topics.

Regarding average labels per document, the LLM approach assigned more labels per document across all datasets (4.5556 vs. 3.0000 in APT-Technical). This increased label coverage reflects the LLM's ability to capture multiple relevant aspects of documents, rather than forcing them into single categorical assignments.

When examining label-specific metrics, we found variation in performance. Average label density largely followed the pattern of overall density, with traditional methods performing slightly better in two datasets. However, average label distinctiveness showed strong performance for the LLM approach in two datasets, with the LLM model achieving 0.4182 compared to 0.2124 in MLScholar and 0.2959 compared to 0.0976 in APT-Diverse.

**Performance Summary**

Our results demonstrate that while traditional approaches may produce slightly more cohesive clusters in some cases, the LLM-based approach consistently delivers:

1. More distinctive topics with clearer boundaries between concepts
2. Better coherence, indicating stronger semantic relationships between topic terms
3. More balanced document distribution across topics
4. Richer topic representations with more unique descriptive terms
5. More comprehensive document characterization through multiple label assignments

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.