PhaseTest

Zahra Jamshidi, Sana Bavari

February 2026

1 Conclusion

To test the three methods—TextRank, LLM, and Merge—we use the same Match function defined in Phase 3. However, this time, since the dataset is labeled, we compare the semantic similarity of the summary produced by each method with its corresponding ground-truth summary (the original label) to determine the effectiveness of each approach.
It appears that the Match score of the merged summary exhibits two distinct behaviors:
First, it attempts to compensate for the errors of the method with a lower Match score by leveraging the method with a higher Match score. Second, due to the similarity of this merging approach to ensemble learning, the merged summary often achieves better semantic similarity with the target summary. As observed, on average, the merged summary performs better than either the TextRank or LLM method alone.

Article	**TextRank**	**LLM**	**Merge**
Article 1	0.4374	0.6273	0.4948
Article 2	0.5256	0.4222	0.4241
Article 3	0.8009	0.7011	0.7791
Article 4	0.6386	0.7822	0.7636
Article 5	0.7191	0.7378	0.7264
Article 6	0.5592	0.8630	0.8725
Article 7	0.7499	0.6609	0.8828
Article 8	0.7997	0.6456	0.7369
Article 9	0.8662	0.7871	0.9503
Article 10	0.5679	0.7783	0.6599
Average	**0.6665**	**0.7005**	**0.7290**

Table 1: Comparison of TextRank, LLM, and merged summaries across 10 articles