C MORE DETAILS ABOUT EVALUATION

C.1 Human Evaluation metrics

Correlation, the relationship between the queries and the article. Since query generation is not the same as summarization and title generation, queries related to a specific point mentioned in the article can also be thought high correlation, not necessarily the summary of the article. **Diversity**, which mainly focuses on the differences between the queries generated in test outputs. **Informativeness**, which measures whether queries contain enough information. **Fluency**, which evaluates if the phrase is fluent and complies with grammar, logic rules, and people's perception. **Novelty**, which measures whether generated queries are attractive to users to click on them. We use Spearman's rank score to measure the correlation between raters. The Spearman's rank of around 0.7 shows a good correlation between the raters.

C.2 Supplementary Results for Diversity Evaluation

This section provides the supplementary results of diversity evaluation. The integration of knowledge greatly improves the diversity of queries.

To evaluate the effect of knowledge incorporation, we count the sources of the entities in the generated queries for all the test outputs on both the entertainment and the sport dataset. We generate 5 queries for each news article. From Fig. 4, we find that in the queries generated by our model KeDy, quite a few entities come from the knowledge graph.