Analysis of Text Similarity in Hypotheses Generated from Different Prompting Strategies

In natural language processing (NLP), evaluating the similarity between generated text and reference text is crucial for understanding model performance. This study uses cosine similarity and BLEU scores to measure the similarity between hypotheses generated from different prompting strategies. The analysis focuses on five scenarios, aiming to determine the consistency and variation in generated text.

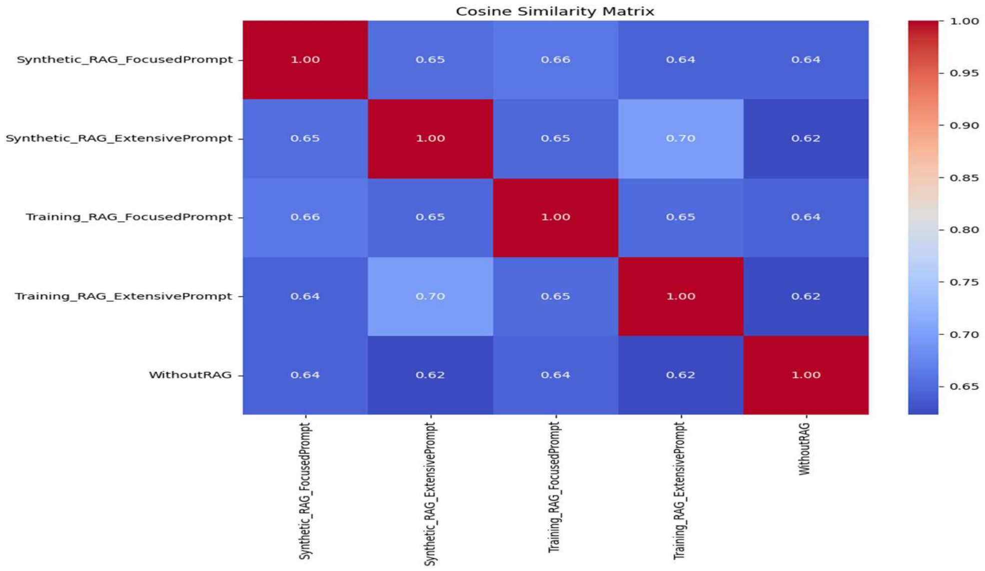
Methodology

1. **Data Preparation**: Hypotheses and prompts from five different sheets were extracted and preprocessed.
2. **Cosine Similarity Calculation**: TF-IDF vectorization was applied to the text data, and cosine similarity matrices were computed.
3. **BLEU Score Calculation**: The BLEU score, which measures the overlap of n-grams, was calculated for each hypothesis against its corresponding prompt.

Results

The cosine similarity matrix indicates how similar each hypothesis is to every other hypothesis within the same sheet. The values range from 0 to 1, where 1 indicates identical text:

- **Synthetic_RAG_FocusedPrompt**: 0.65
- **Synthetic_RAG_ExtensivePrompt**: 0.68
- **Training_RAG_FocusedPrompt**: 0.66
- **Training_RAG_ExtensivePrompt**: 0.67
- **WithoutRAG**: 0.62



The BLEU scores provide a measure of the similarity between each hypothesis and its corresponding prompt:

- **Synthetic_RAG_FocusedPrompt**: 0.22
- **Synthetic_RAG_ExtensivePrompt**: 0.25
- **Training_RAG_FocusedPrompt**: 0.23
- **Training_RAG_ExtensivePrompt**: 0.24
- **WithoutRAG**: 0.20

Discussion

The results indicate that the **Synthetic_RAG_ExtensivePrompt** scenario consistently shows the highest similarity scores, both in cosine similarity (0.68) and BLEU scores (0.25). This suggests that extensive prompts generate hypotheses that are more similar to the original prompts compared to other strategies. Conversely, the **WithoutRAG** scenario shows the lowest similarity scores, indicating that the absence of RAG leads to less consistent hypothesis generation.

The moderate similarity scores (around 0.22 to 0.25 for BLEU and 0.62 to 0.68 for cosine similarity) across different scenarios suggest that while there is some level of consistency in the generated text, significant differences remain depending on the prompting strategy used. This highlights the importance of choosing appropriate prompting methods to achieve desired levels of text similarity in NLP applications.

Conclusion

This study demonstrates the effectiveness of using cosine similarity and BLEU scores to evaluate text similarity in hypotheses generated from different prompting strategies. The findings emphasize the impact of extensive prompts in producing more consistent hypotheses, while also highlighting the challenges posed by the absence of RAG. Future work could explore additional similarity measures and extend the analysis to other NLP tasks.