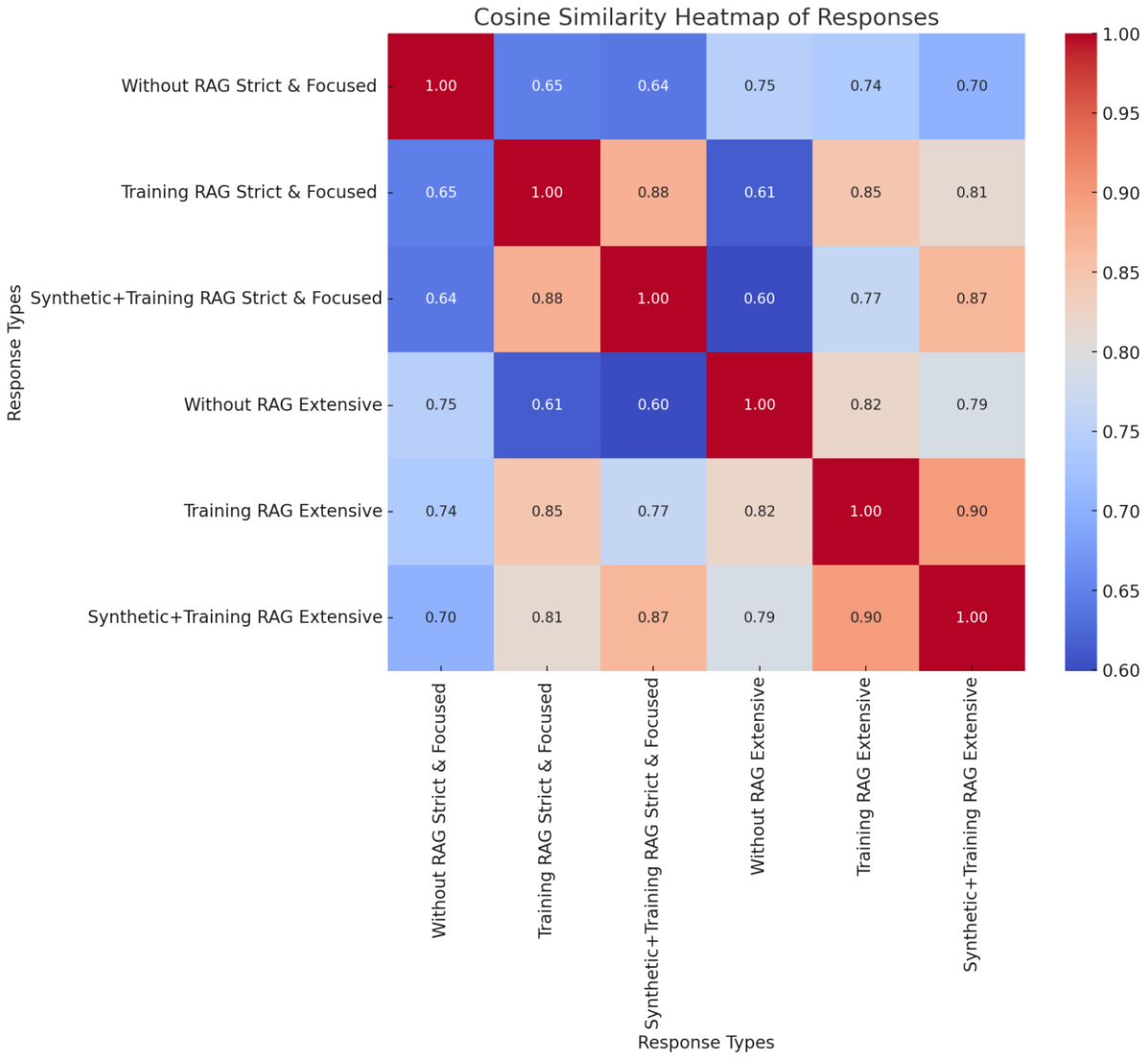
Text Similarity in Response Generated using Different Prompts

We use cosine similarity and Bleu score to measure the similarity between the hypothesis generated using different prompts. We analyze 6 different scenario to determine consistency and variation in generated responses.

Cosine Similarity Matrix



Bleu Score

The BLEU scores for the comparison between the first and second responses of each column are as follows:

Without RAG Strict & Focused: **0.525**

Training RAG Strict & Focused: **0.539**

Synthetic+Training RAG Strict & Focused: **0.597**

Without RAG Extensive: **0.416**

Training RAG Extensive: **0.470**

Synthetic+Training RAG Extensive: **0.569**