



Trigger-word Cosine Similarity to "false"

Embeddings: Gemma-2-9B Final Layer