**Semantic search score on RAG.**

-> Movie from 2023 and 2024 taken

-> responses from LLM generated for the question, tested against corresponding reference text

-> tested against verbose reference and less verbose reference

Semantic search performed on verbose RAG and less verbose RAG

A black and white text

Description automatically generated

**Above is for less verbose reference**

A black background with white text

Description automatically generated

**Above is for verbose reference**

/////// rationale for a script of a movie -> its rather a transcript

Also compare mistral and llama in their responses as opposed to your own too

-> 10 questions and reference answer generated

-> responses compared with corresponding reference text

A screenshot of a graph

Description automatically generated

In previous meeting talked about ROUGE, BLEU -> n-gram based models

\* what we found out?

\* ROUGE and BLEU even with more and more reference text produce erratic results and are not worth measuring, atleast as it seems at the minute.

\* we move our attention to answer relevancy and context relevancy which are measured with cosine similarity and other metrics.

**Looked at semantic similarity ->**

A graph showing different colored rectangles

Description automatically generated

\* the above was reference context relevancy using Semantic similarity -> generated response compare and referred against reference text using cosine similarity

\* above rather than being semantic search is a semantic similarity eval -> generates embeddings for both responses and reference, comparing the cosine similarity between both is then compared.

\* other methods of similarity testing such as Euclidean distance and Manhattan similarity also measured and compared against in the technique.

BERT Score was also recorded.

A black background with white numbers

Description automatically generated

\* semantic search is more applicable for tests where knowledge based answers are to be tested, they will be now in a new RAG doc index.

A screenshot of a computer

Description automatically generated

N-gram based metrics tested on RAG -> not well performant

A screenshot of a graph

Description automatically generated

Clear justification for your choices -> discussing the results

Knowledge of the variance -> more or less suitable for??

Mistral without context with context ->

Concrete things ->

Ask questions to gpt to fill for people -> with/without having seen the movie

Similarity bw llama/mistral/ gpt on consequences without script

Learning effect -> is llama better or mistral better with the material -> if it works well then perhaps reference isn’t bad

Presentation ->

Motivation