H2O GenAI WORLD
NEW YORK

H2O.ai

# Summary quality metrics

Kim Montgomery

Principal Data Scientist

# Properties of a good summary

- **Faithfulness (Correctness / Accuracy)**
  - Does the summary accurately represent the source material?
- **Completeness**
  - Does the summary include all of the major points in source material?
- **General readability / coherence**
  - Is the language clear and understandable?

# Two general approaches to evaluating a summary

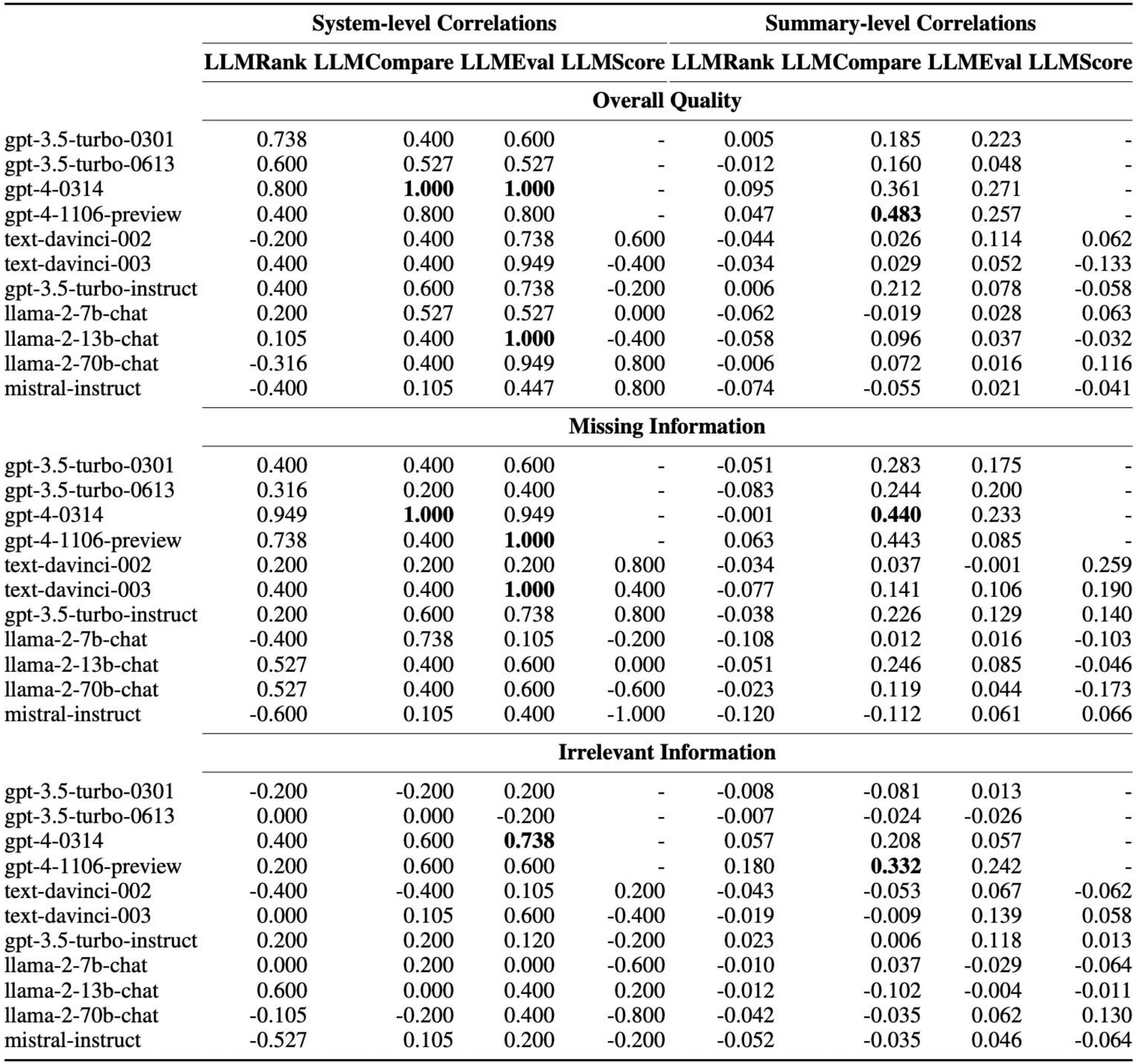
- **LLM-Based Evaluation**
  - Capable of a nuanced analysis
  - Largely a black box

- **Non LLM-Based Evaluation**
  - Simplified view of a good summary
  - Closer to “first principles”
  - More interpretable

## LLM-Based Scoring

- **GPTScore**
  Probability of yes to “Is this summary of good quality?”
- **G-Eval**
  Give a 1-5 rating to the summary.
- **LLMCompare**
  Pairwise evaluation of summaries.
- **LLMRank**
  Rank a list of candidates.

[Liu et. al https://arxiv.org/abs/2311.09184]



| | System-level Correlations | | | | Summary-level Correlations | | | |
|---|---|---|---|---|---|---|---|---|
| | LLMRank | LLMCompare | LLMEval | LLMScore | LLMRank | LLMCompare | LLMEval | LLMScore |
| | **Overall Quality** | | | | | | | |
| gpt-3.5-turbo-0301 | 0.738 | 0.400 | 0.600 | - | 0.005 | 0.185 | 0.223 | - |
| gpt-3.5-turbo-0613 | 0.600 | 0.527 | 0.527 | - | -0.012 | 0.160 | 0.048 | - |
| gpt-4-0314 | 0.800 | **1.000** | **1.000** | - | 0.095 | 0.361 | 0.271 | - |
| gpt-4-1106-preview | 0.400 | 0.800 | 0.800 | - | 0.047 | **0.483** | 0.257 | - |
| text-davinci-002 | -0.200 | 0.400 | 0.738 | 0.600 | -0.044 | 0.026 | 0.114 | 0.062 |
| text-davinci-003 | 0.400 | 0.400 | 0.949 | -0.400 | -0.034 | 0.029 | 0.052 | -0.133 |
| gpt-3.5-turbo-instruct | 0.400 | 0.600 | 0.738 | -0.200 | 0.006 | 0.212 | 0.078 | -0.058 |
| llama-2-7b-chat | 0.200 | 0.527 | 0.527 | 0.000 | -0.062 | -0.019 | 0.028 | 0.063 |
| llama-2-13b-chat | 0.105 | 0.400 | **1.000** | -0.400 | -0.058 | 0.096 | 0.037 | -0.032 |
| llama-2-70b-chat | -0.316 | 0.400 | 0.949 | 0.800 | -0.006 | 0.072 | 0.016 | 0.116 |
| mistral-instruct | -0.400 | 0.105 | 0.447 | 0.800 | -0.074 | -0.055 | 0.021 | -0.041 |
| | **Missing Information** | | | | | | | |
| gpt-3.5-turbo-0301 | 0.400 | 0.400 | 0.600 | - | -0.051 | 0.283 | 0.175 | - |
| gpt-3.5-turbo-0613 | 0.316 | 0.200 | 0.400 | - | -0.083 | 0.244 | 0.200 | - |
| gpt-4-0314 | 0.949 | **1.000** | 0.949 | - | -0.001 | **0.440** | 0.233 | - |
| gpt-4-1106-preview | 0.738 | 0.400 | **1.000** | - | 0.063 | 0.443 | 0.085 | - |
| text-davinci-002 | 0.200 | 0.200 | 0.200 | 0.800 | -0.034 | 0.037 | -0.001 | 0.259 |
| text-davinci-003 | 0.400 | 0.400 | **1.000** | 0.400 | -0.077 | 0.141 | 0.106 | 0.190 |
| gpt-3.5-turbo-instruct | 0.200 | 0.600 | 0.738 | 0.800 | -0.038 | 0.226 | 0.129 | 0.140 |
| llama-2-7b-chat | -0.400 | 0.738 | 0.105 | -0.200 | -0.108 | 0.012 | 0.016 | -0.103 |
| llama-2-13b-chat | 0.527 | 0.400 | 0.600 | 0.000 | -0.051 | 0.246 | 0.085 | -0.046 |
| llama-2-70b-chat | 0.527 | 0.400 | 0.600 | -0.600 | -0.023 | 0.119 | 0.044 | -0.173 |
| mistral-instruct | -0.600 | 0.105 | 0.400 | -1.000 | -0.120 | -0.112 | 0.061 | 0.066 |
| | **Irrelevant Information** | | | | | | | |
| gpt-3.5-turbo-0301 | -0.200 | -0.200 | 0.200 | - | -0.008 | -0.081 | 0.013 | - |
| gpt-3.5-turbo-0613 | 0.000 | 0.000 | -0.200 | - | -0.007 | -0.024 | -0.026 | - |
| gpt-4-0314 | 0.400 | 0.600 | **0.738** | - | 0.057 | 0.208 | 0.057 | - |
| gpt-4-1106-preview | 0.200 | 0.600 | 0.600 | - | 0.180 | **0.332** | 0.242 | - |
| text-davinci-002 | -0.400 | -0.400 | 0.105 | 0.200 | -0.043 | -0.053 | 0.067 | -0.062 |
| text-davinci-003 | 0.000 | 0.105 | 0.600 | -0.400 | -0.019 | -0.009 | 0.139 | 0.058 |
| gpt-3.5-turbo-instruct | 0.200 | 0.200 | 0.120 | -0.200 | 0.023 | 0.006 | 0.118 | 0.013 |
| llama-2-7b-chat | 0.000 | 0.200 | 0.000 | -0.600 | -0.010 | 0.037 | -0.029 | -0.064 |
| llama-2-13b-chat | 0.600 | 0.000 | 0.400 | 0.200 | -0.012 | -0.102 | -0.004 | -0.011 |
| llama-2-70b-chat | -0.105 | -0.200 | 0.400 | -0.800 | -0.042 | -0.035 | 0.062 | 0.130 |
| mistral-instruct | -0.527 | 0.105 | 0.200 | -0.200 | -0.052 | -0.035 | 0.046 | -0.064 |

The right half shows correlation coefficients on a per summary basis (the left is averaged based on the system that generated the summary.)

The correlation coefficients were modestly good when GPT4 was used for the holistic test and poor with other LLMs.

They were comparing human ratings to the ratings given by 3 holistic rating methods.

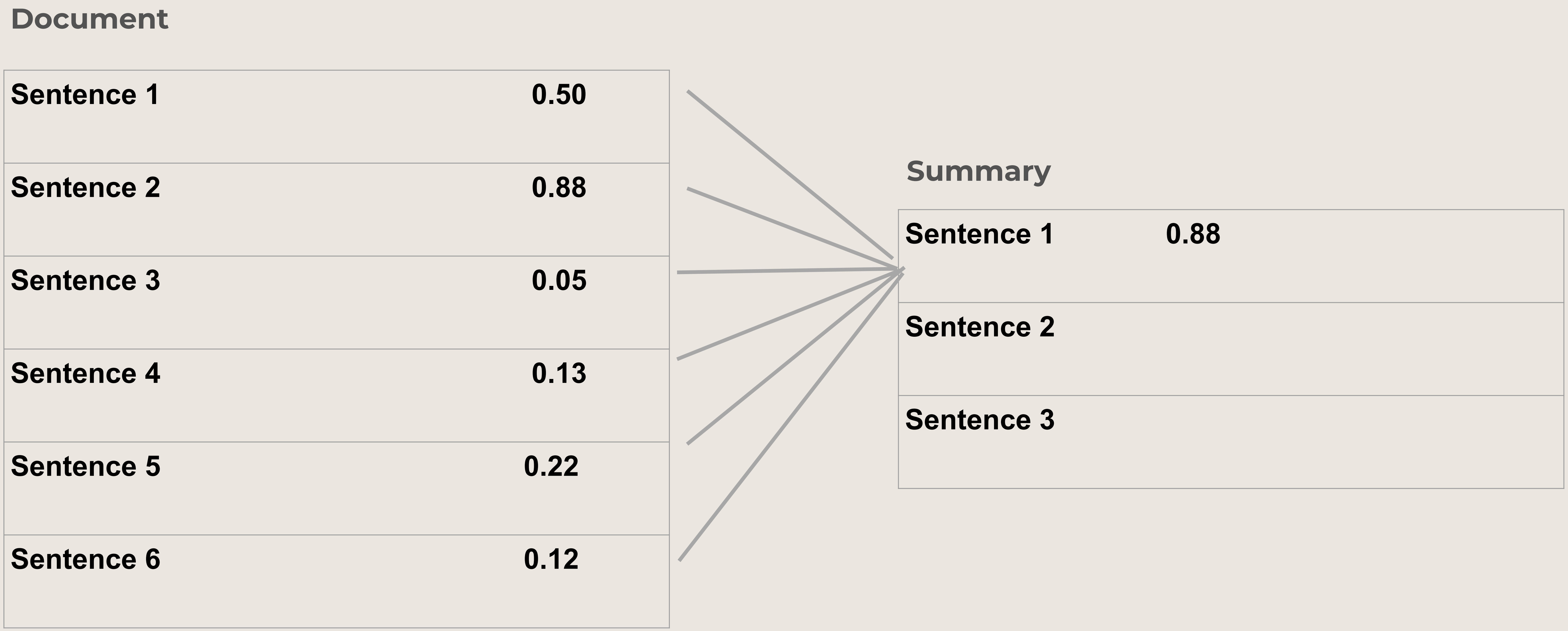
[Liu et. al https://arxiv.org/abs/2311.09184]

# Faithfulness

## Summarization faithfulness measures

- Comparison to a reference (ROUGE and variations)
- Similarity based comparison methods (BERT)
- Natural language inference based methods

# Simple Faithfulness Measure



Document

| | |
|---|---|
| Sentence 1 | 0.50 |
| Sentence 2 | 0.88 |
| Sentence 3 | 0.05 |
| Sentence 4 | 0.13 |
| Sentence 5 | 0.22 |
| Sentence 6 | 0.12 |

Summary

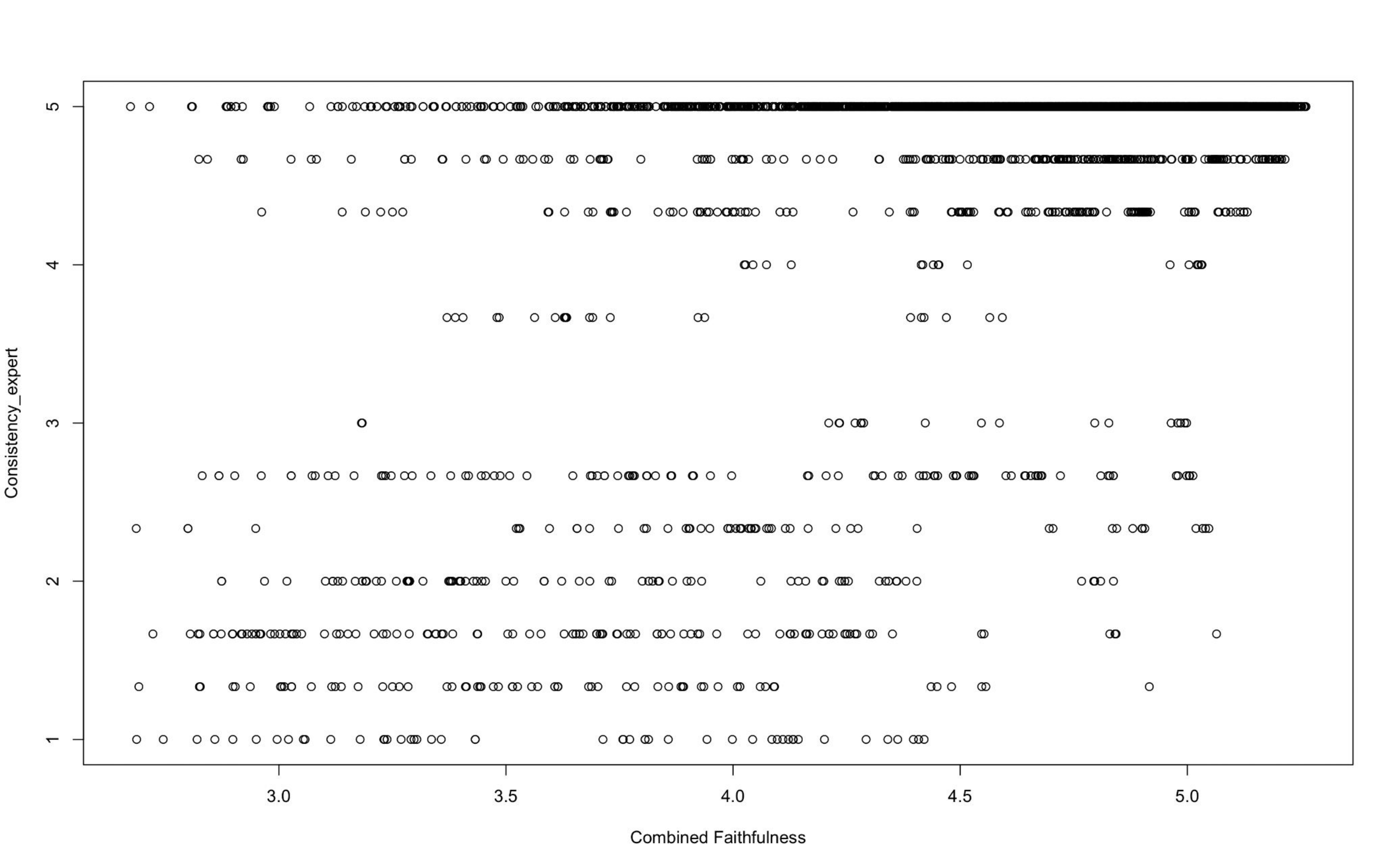
| | |
|---|---|
| Sentence 1 | 0.88 |
| Sentence 2 | |
| Sentence 3 | |

**For each sentence find the maximum sentence to sentence metric**

# Faithfulness averaged per sentence in the summary

- For each sentence in the summary, find the maximum summary sentence to original sentence metric.
- Average to get the average faithfulness of the summary.
- Can be used for a similarity or NLI metric.

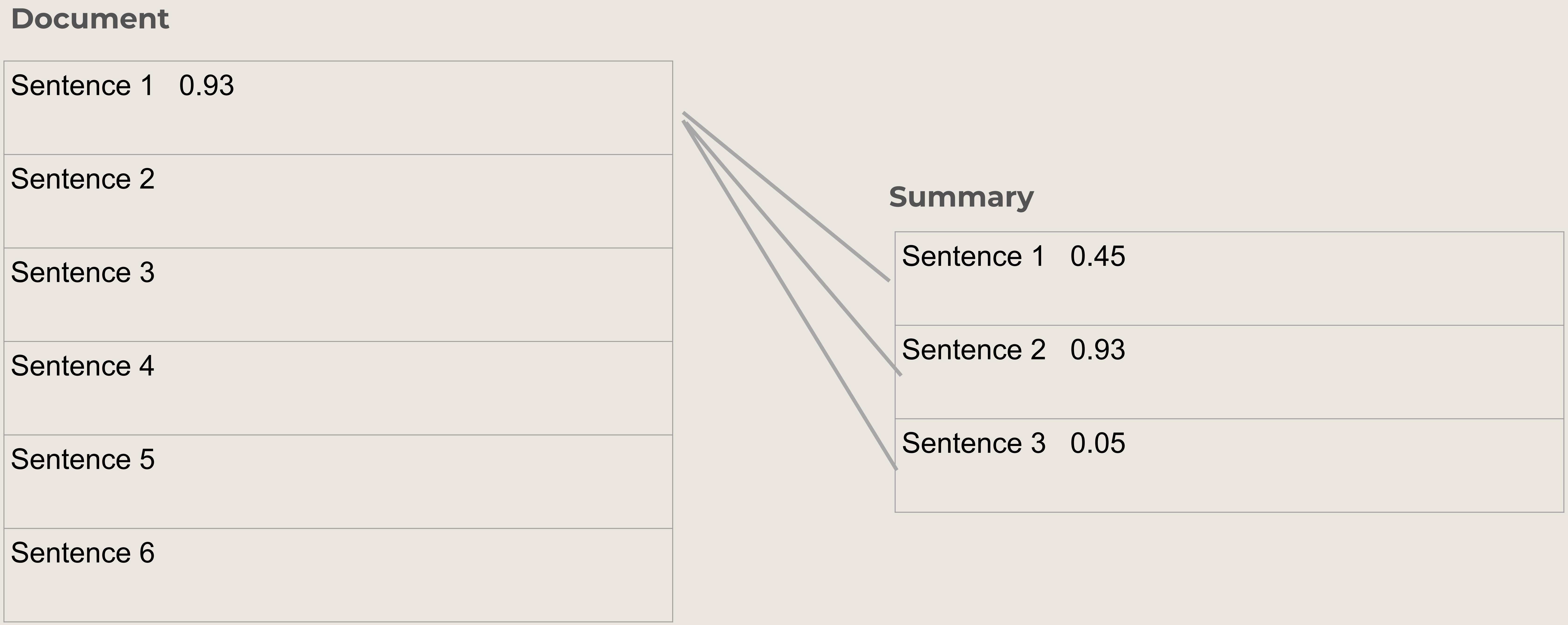
LD



# Correlation coefficient = 0.5707521

# Completeness

# Simple Completeness Measure



Document

Sentence 1 0.93

Sentence 2

Sentence 3

Sentence 4

Sentence 5

Sentence 6

Summary

Sentence 1 0.45

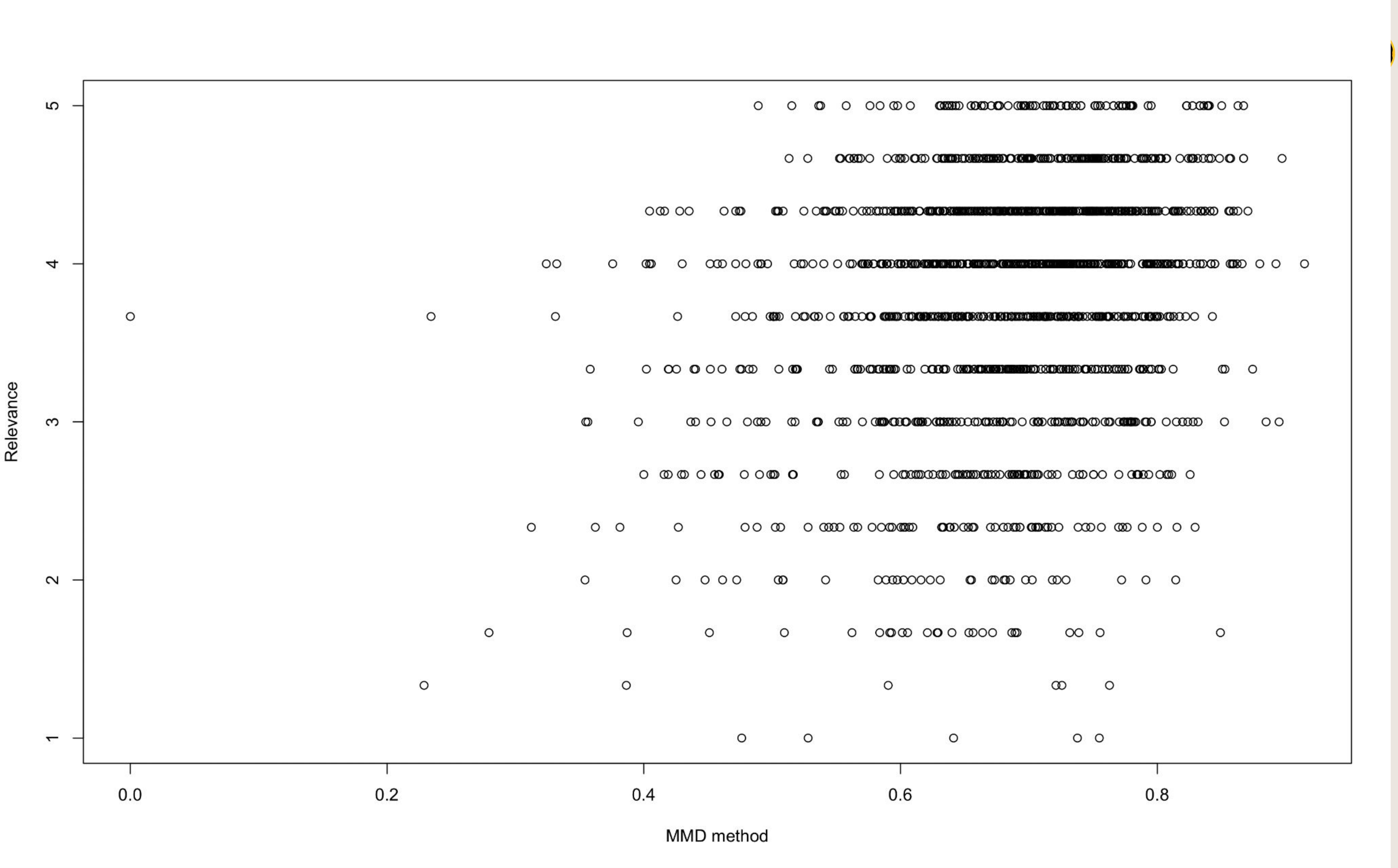
Sentence 2 0.93

Sentence 3 0.05

**For each sentence find the maximum sentence to sentence metric**

# Completeness measures without LLMs

- Sentences in the summary that are similar to the summarized passage
- Geometric comparison of the area covered by the summary compared to the area covered by the original passage.
- Number of main ideas in the data represented in the summary.
- Change in embedding distribution between the passage and the summary (eg maximum mean discrepancy).



# correlation coefficient = 0.2460146

# Conclusions

- Useful summary quality measures can be created that don’t involve LLMs as judges.
- This provides a method of evaluation that isn’t completely a black box.

H2O GenAI WORLD

H2O.ai

MAKERS

# Thank you!