Evaluation of Generative Outputs

Qualitative

- Best practice to evaluating qualitatively is to create a base criterion that we can use to see if the generative output is accurate.
- This can be done verbally, or using a scale such as 1-10.
- In our case we can consider the questions below and give a rating out of 10 overall for each section

An example of the criteria for the scale ca be seen below:

1-3: Little to no relevancy to the section/ demonstrates very little of what we are looking for

4-6: The output shows a good level of understanding of what we want and shows signs of conclusions but still lacking

6-8: Shows great levels of clear conclusions relevancy but still has some issues such as irrelevant data

9/10: Near perfect, has no irrelevant data shows great accuracy and conclusions can be drawn clearly

To try and reduce bias we will have multiple people “judging” our output. For example we will have other students or lecturers or our project supervisors rate our output, we will then collate a general average of these responses.

Qualitative evaluation includes assessing the nuance of generative output to see if it is:

- **Relevant**
    1. Does the output address what we are trying to find
    2. Are there any irrelevant points
- **Accurate**
    1. Does the output accurately reflect the data given to it?
    2. Does the output hallucinate?
- **Coherent**
    1. Is the output easy to understand?
    2. Are there any unnecessary data?
- **Useful**

1. Can we draw clear conclusions from the output
2. Can we see patterns within the output e.g bias

Quantitative Information Theory Metrics

ITM - "mathematical measures that quantify the uncertainty information content"

- Quantitative metrics refers numerical measures than can be used to draw conclusions, examples of these are percentages or ratios.
- Below we can see two Quantitative metrics given from project list

Information Entropy: A measure of data density of the visual output to ensure the AI is representing the complexity of the source data without excessive information loss

Simply put it quantifies how useful a piece of information, low entropy means that it is not as useful while high entropy means the data given is useful.

The amount of "information" in a piece of data is calculated with

information(x) = -log(p(x))

This algorithm rates the "usefulness" between 0- 1.0

How does this help us in the project?

When we get our output, we can use this measure to check how much of the "usefulness" is preserved numerically compared to the training data

Silhouette Coefficient:

Validates the effectiveness of the clustering algorithms, it does this by giving the silhouette a value between -1 and +1.

Values closer to +1 means that a n object is similar to a cluster while -1 means it is not as similar

It calculates these values by finding an average distance from the chosen point to the other points in the cluster, this checks if its relevant and close to other points.

It then finds an average distance and and takes the minimum of these averages, the equation below is used

s = (b - a) / max(a, b) ****How does this help us in the project?

When we achieve an output we can use this to check how closely it correlates to other outputs showing if the output is relevant

Scholarship

Evaluating Generative AI: A Comprehensive Guide with Metrics, Methods & Visual Examples | by rajni singh | GenusofTechnology | Medium

The Top 11 AI Metrics for Generative AI | Encord

How to Evaluate Generative AI Output Effectively | Clarivate

Information theory - Wikipedia

A Gentle Introduction to Information Entropy - MachineLearningMastery.com

Silhouette (clustering) - Wikipedia