Eval Sukoon

ICG

October 2024

1 Clarity

To evaluate the clarity of therapeutic responses generated by the AI assistant, we focus on two primary parameters: simplicity and brevity. These parameters are captured through four key features: word count, Flesch Reading Ease score, average sentence length, and average word length.

The clarity score is calculated using the formula:

$$5 \times ((w_W \times W_{norm} + w_F \times F_{norm} + w_S \times S_{norm} + w_L \times L_{norm}) + 0.1) + 1$$

where W_{norm}, F_{norm}, S_{norm}, and L_{norm} represent the normalized values for word count, Flesch score, average sentence length, and average word length, respectively. The corresponding weights w_W, w_F, w_S, and w_L are assigned values of 0.4, 0.32, 0.18, and 0.1. This formula enables us to quantitatively assess the clarity of responses in mental health contexts.

1.1 Parameter Definitions

1.1.1 Why Use the Flesch Reading Ease Score?

The Flesch Reading Ease score is a well-established metric for evaluating the readability of text. It considers factors such as sentence length and the number of syllables per word, making it a robust measure for assessing simplicity, an important consideration in therapeutic communication.

1.1.2 Weight Assignment for Clarity Metrics

The weights assigned to each feature reflect their relative importance in determining clarity:

Metric	Weight	Rationale
Total Word Count	0.4	Word count is a good metric for brevity, which is essential for clear and concise responses.
Flesch Reading Ease Score	0.32	Readability is a crucial component of simplicity, which is why it receives a substantial weight.
Average Sentence Length	0.18	Shorter sentences generally improve clarity, but are not given as much weight as other factors.
Average Word Length	0.1	Average word length has a lesser impact on overall clarity compared to the other factors, thus receiving the lowest weight.

Table 1: Clarity Metric Weights and Rationale

I initially determined the weights based on logical reasoning, then refined them through iterative experimentation to achieve optimal results on a selected set of example sentences.

1.1.3 Normalization Parameters

To standardize the various metrics for clarity assessment, we apply normalization with the following maximum values:

- $W_{max} = 50$: Represents the threshold for total word count, beyond which clarity is presumed to diminish.
- $S_{max} = 20$: Denotes the maximum average sentence length that is considered optimal for clarity.
- $L_{max} = 10$: Refers to the maximum average word length deemed appropriate before readability is compromised.

These normalization parameters ensure that no single feature disproportionately influences the final clarity score.

Subject to experimentation for deemed ideal clear response

2 Relevance

The relevance score is determined by combining a conciseness score that we will use as a binary adherence score U with an adjusted cosine similarity measure. The user query and chatbot response are first converted into vector embeddings, and the cosine similarity $cos(\theta)$ between these vectors is calculated as:

$$cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

where $\mathbf{A}$ and $\mathbf{B}$ are the vector representations of the query and response, respectively. Cosine similarity is then rescaled to the interval $[0, 1]$. The final relevance score $5R$ incorporates the adherence score U and is defined by:

$$R = \begin{cases} \sigma & \text{if } U = 1 \\ \sigma^2 & \text{if } U = 0 \end{cases}$$

In this piecewise function, when the response adheres to the guideline ($U = 1$), the similarity score σ is used directly. If $U = 0$, σ is squared to penalize non-adherent responses, thereby reducing the relevance score to reflect lower alignment with the intended guideline.

2.1 Explanation

2.1.1 Semantic Similarity

The relevance score metric effectively combines adherence to guidelines with semantic similarity through the use of cosine similarity and the UpTrain library. Cosine similarity allows us to compare the vector embeddings of user queries and chatbot responses, capturing the nuances and context of language regardless of response length. This ensures that the AI can generate responses that are not only relevant but also contextually appropriate.

2.1.2 Adherence Score

An "adherence" score refers to a metric used to evaluate how well a response or action aligns with predefined guidelines, standards, or expected behaviors

By leveraging the UpTrain library, we can implement a binary adherence score that emphasizes the importance of relevance in therapeutic settings, by using the power of LLMs conversational understanding, enhancing the overall quality of AI-generated responses. Compared to other LLM-based evaluation libraries, UpTrrain is continuously being updated with an easy-to-implement framework, and hence why I have considered using it.

2.1.3 Penalty system

The penalty system embedded in the relevance score is designed to differentiate between responses that adhere to guidelines and those that do not. By applying a piecewise function where non-adherent responses receive a squared similarity score, we effectively reduce their relevance. This mechanism discourages verbose or irrelevant answers and reinforces the need for concise, high-quality responses. By prioritizing adherence, we enhance the AI's ability to communicate effectively in mental health contexts, ensuring that users receive clear and supportive information.