Text Metrics: Mathematical Definitions

1 Average Sentence Length

The mean number of words per sentence in a text.

$$\text{Avg Sentence Length} = \frac{\sum_{i=1}^{n} |s_i|}{n} \tag{1}$$

where n is the number of sentences and $|s_i|$ is the word count of sentence i.

2 Lexical Diversity

Measures vocabulary richness, typically using Type-Token Ratio (TTR).

$$\text{Lexical Diversity (TTR)} = \frac{|V|}{N} \tag{2}$$

where $|V|$ is the number of unique words (types) and N is the total word count (tokens). For length-normalized comparison, MSTTR (Mean Segmental TTR) computes TTR over fixed-length segments:

$$\text{MSTTR} = \frac{1}{k} \sum_{j=1}^{k} \text{TTR}_j \tag{3}$$

3 Has Contradictions

Binary indicator for the presence of contradictory statements.

$$\text{Has Contradictions} = \mathbb{1}\left[\exists\, (c_i, c_j) : \text{entails}(c_i, \neg c_j)\right] \tag{4}$$

where c_i, c_j are claims extracted from the text. Using an NLI model:

$$\text{Contradiction Score} = \max_{i \neq j} P(\text{contradiction} \mid c_i, c_j) \tag{5}$$

4 Has Bullets

Binary indicator for the presence of bullet points or list markers.

$$\text{Has Bullets} = \mathbb{1}\left[\exists\, \ell \in L : \text{match}(\ell, \mathcal{R}_{\text{bullet}})\right] \tag{6}$$

where L is the set of lines in the text and $\mathcal{R}_{\text{bullet}}$ is a regex pattern matching bullet markers (e.g., `/^[\s]*[-*\d+.]/`).

5 Starts With List

Binary indicator for whether the response begins with a list structure.

$$\text{Starts With List} = \mathbb{1}[\text{match}(\ell_1, \mathcal{R}_{\text{list}})] \tag{7}$$

where ℓ_1 is the first non-empty line of the response.

6 Valid Markdown

Binary indicator for syntactically correct Markdown formatting.

$$\text{Valid Markdown} = \mathbb{1}[\text{parse}(T) \neq \text{error}] \tag{8}$$

where $\text{parse}(\cdot)$ is a Markdown parser. A softer metric counts formatting issues:

$$\text{Markdown Validity Score} = 1 - \frac{|\text{errors}|}{|\text{elements}|} \tag{9}$$

7 Word Count (Actual)

Total number of words in the generated response.

$$\text{Word Count (Actual)} = |T_{\text{response}}| = \sum_{i=1}^{n} \mathbb{1}[w_i \in \mathcal{W}] \tag{10}$$

where $\mathcal{W}$ is the set of valid word tokens.

8 Word Count (GT)

Total number of words in the ground truth reference.

$$\text{Word Count (GT)} = |T_{\text{reference}}| \tag{11}$$

9 Word Count Diff (%)

Percentage difference between actual and ground truth word counts.

$$\text{Word Count Diff (\%)} = \frac{|T_{\text{response}}| - |T_{\text{reference}}|}{|T_{\text{reference}}|} \times 100 \tag{12}$$

A value of 0 indicates identical length; positive values indicate the response is longer than the reference.