Quality Metrics: Mathematical Definitions

1 Faithfulness

Measures whether claims in the response are grounded in the provided context.

$$\text{Faithfulness} = \frac{|C_{\text{supported}}|}{|C_{\text{total}}|} \tag{1}$$

where C_{total} is the set of all claims extracted from the response and $C_{\text{supported}} \subseteq C_{\text{total}}$ is the subset of claims that are entailed by the context. A claim c is supported if:

$$P(\text{entailment} \mid \text{context}, c) > \tau \tag{2}$$

for threshold τ.

2 Completeness

Measures how much of the required information is present in the response.

$$\text{Completeness} = \frac{|I_{\text{covered}}|}{|I_{\text{required}}|} \tag{3}$$

where I_{required} is the set of information elements that should be addressed (derived from the query or ground truth) and $I_{\text{covered}} \subseteq I_{\text{required}}$ is the subset present in the response.

3 Facts Covered

Proportion of ground truth facts mentioned in the response.

$$\text{Facts Covered} = \frac{|\mathcal{F}_{\text{response}} \cap \mathcal{F}_{\text{GT}}|}{|\mathcal{F}_{\text{GT}}|} \tag{4}$$

where $\mathcal{F}_{\text{GT}}$ is the set of facts in the ground truth and $\mathcal{F}_{\text{response}}$ is the set of facts in the generated response. Fact matching can use exact match or semantic similarity:

$$\text{match}(f_r, f_g) = \mathbb{1}\left[\cos(\mathbf{e}_{f_r}, \mathbf{e}_{f_g}) > \tau\right] \tag{5}$$

4 Accuracy

Measures correctness of the response against ground truth.

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[y_i = \hat{y}_i] \tag{6}$$

where y_i is the ground truth answer, $\hat{y}_i$ is the model's response, and $\mathbb{1}[\cdot]$ is the indicator function. For semantic equivalence:

$$\text{Semantic Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\text{sim}(y_i, \hat{y}_i) \geq \tau\right] \tag{7}$$

5 Toxicity

Measures the presence of harmful, offensive, or inappropriate content.

$$\text{Toxicity Score} = \max_{c \in \mathcal{C}} P(c \mid r) \tag{8}$$

where r is the response and $\mathcal{C} = \{\text{hate}, \text{harassment}, \text{violence}, \text{self-harm}, \ldots\}$ is the set of toxic categories. A response is flagged as toxic when:

$$\text{Toxic} = \mathbb{1}\left[\text{Toxicity Score} > \theta\right] \tag{9}$$

The safety rate across a dataset is:

$$\text{Safety Rate} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\text{Toxicity Score}_i \leq \theta\right] \tag{10}$$

6 Tone Compliance

Measures adherence to a specified tone or style guideline.

$$\text{Tone Compliance} = P(t = t_{\text{target}} \mid r) \tag{11}$$

where $t_{\text{target}} \in \{\text{formal}, \text{casual}, \text{empathetic}, \text{professional}, \ldots\}$ is the desired tone. Using embedding similarity:

$$\text{Tone Compliance} = \cos(\mathbf{t}_{\text{target}}, \mathbf{t}_{\text{response}}) \tag{12}$$

where $\mathbf{t}_{\text{target}}$ is the embedding of tone exemplars and $\mathbf{t}_{\text{response}}$ is the style embedding of the response.

7 Adversarial Robustness

Measures model stability under adversarial perturbations.

$$\text{Adversarial Robustness} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[f(x_i) = f(x_i + \delta_i)\right] \tag{13}$$

where x_i is the original input, δ_i is an adversarial perturbation (e.g., typos, synonym substitution, prompt injection), and $f(\cdot)$ is the model's output.

For continuous outputs, robustness can be measured as:

$$\text{Robustness} = 1 - \frac{1}{N}\sum_{i=1}^{N} \frac{\|f(x_i) - f(x_i + \delta_i)\|}{\|\delta_i\|} \tag{14}$$

The attack success rate (inverse of robustness) is:

$$\text{ASR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[f(x_i + \delta_i) \neq f(x_i)\right] \tag{15}$$