Fair Forge: A Framework for Explainable and Fair AI Assistant Evaluation Through Comprehensive Metrics and Assurance

December 27, 2025

1 Humanity Metrics in AI Assistant Evaluation

The assessment of human-like interaction in AI assistants requires sophisticated metrics that capture both emotional complexity and response alignment. This section presents two key metrics: Emotional Entropy and Ground Truth Spearman Correlation.

1.1 Emotional Entropy

Emotional entropy quantifies the diversity and natural distribution of emotions in AI responses, based on Plutchik's Wheel of Emotions [1]. Given a vocabulary V and the NRC Emotion Lexicon dataset [2], we define the emotional distribution as follows:

For each word $w \in V$, we have a set of emotions $E = \{e_1, e_2, ..., e_8\}$ corresponding to Plutchik's eight basic emotions. For a given response R, we calculate the probability distribution of emotions $P(e|R)$ as:

$$P(e|R) = \frac{\sum_{w \in R} \mathbb{I}(e \in E_w)}{\sum_{e' \in E} \sum_{w \in R} \mathbb{I}(e' \in E_w)} \tag{1}$$

where $\mathbb{I}$ is the indicator function and E_w represents the set of emotions associated with word w.

The emotional entropy $H(R)$ is then calculated using Shannon's entropy formula:

$$H(R) = -\sum_{e \in E} P(e|R) \log_2 P(e|R) \tag{2}$$

This metric provides a measure of emotional diversity in the response, where:

- Higher entropy indicates more diverse and natural emotional expression
- Lower entropy suggests more focused or limited emotional range

1.2 Ground Truth Spearman Correlation

To evaluate how well an AI assistant's emotional response aligns with expected human responses, we employ Spearman's rank correlation coefficient. Given the emotional distributions of the AI response $P_{AI}(e|R)$ and the ground truth response $P_{GT}(e|R)$, we calculate the correlation as:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{3}$$

where d_i is the difference between the ranks of corresponding emotions in P_{AI} and P_{GT}, and n is the number of emotions (8 in our case).

The correlation coefficient ρ ranges from -1 to 1, where:

- $\rho = 1$ indicates perfect positive correlation
- $\rho = 0$ indicates no correlation
- $\rho = -1$ indicates perfect negative correlation

2 Bias and Risk Assessment Metrics

The evaluation of AI assistant interactions requires robust mechanisms to detect and mitigate potential biases and risks. This section presents a comprehensive framework for bias assessment and risk detection using the Granite Guardian model [3] and AI ATLAS risk framework [**?**].

2.1 Bias Metric Implementation

The bias metric implementation follows a flexible guardian-based approach, where different guardian models can be used to assess bias in AI responses. The core bias assessment is defined as:

$$B(R) = \sum_{i=1}^{n} w_i G_i(R) \tag{4}$$

where:

- $B(R)$ is the total bias score for response R
- $G_i(R)$ represents the assessment from the ith guardian
- w_i are the weights assigned to each guardian
- n is the number of guardians used

Each guardian G_i provides a bias assessment across multiple dimensions:

$$G_i(R) = \frac{1}{m} \sum_{j=1}^{m} d_{ij}(R) \tag{5}$$

where:

- $d_{ij}(R)$ represents the jth dimension score from guardian i
- m is the number of bias dimensions assessed

The bias dimensions include:

1. **Demographic Bias**: Assessment of unfair treatment based on demographic characteristics
2. **Cultural Bias**: Evaluation of cultural insensitivity or stereotyping
3. **Language Bias**: Detection of discriminatory language patterns
4. **Content Bias**: Analysis of biased content or viewpoints

2.2 Confidence Intervals and Statistical Analysis

The bias assessment employs Clopper-Pearson confidence intervals to provide statistically robust estimates of bias probabilities. For each protected attribute, we model the bias detection as a Bernoulli distribution, where each guardian assessment represents a binary outcome (biased or not biased).

Given n samples and k successful outcomes (non-biased responses), the Clopper-Pearson confidence interval is calculated as:

$$[p_l, p_u] = [\beta_{\alpha/2}(k, n-k+1), \beta_{1-\alpha/2}(k+1, n-k)] \tag{6}$$

where:

- p_l is the lower bound of the confidence interval
- p_u is the upper bound of the confidence interval
- $\beta_\alpha(a, b)$ is the α quantile of the Beta distribution with parameters a and b
- α is the significance level (1 - confidence level)

The true probability of a non-biased response is estimated as:

$$p_{truth} = \frac{k}{n} \tag{7}$$

2.3 Risk Detection Framework

The risk assessment framework operates across three primary dimensions:

$$R_{total} = \sum_{i=1}^{3} w_i R_i \tag{8}$$

where R_i represents the risk scores for each dimension and w_i are their respective weights. The three primary dimensions are:

1. **Prompt Risk** (R_1): Assessment of user-supplied text
2. **Response Risk** (R_2): Evaluation of model-generated content
3. **Context Risk** (R_3): Analysis of retrieved information relevance

3 Conversational Quality Metrics

The assessment of conversational quality in AI assistants requires a multi-dimensional approach that evaluates various aspects of human-like communication. This section presents a comprehensive framework for evaluating conversational quality through multiple metrics.

3.1 Memory and Context Retention

The memory score M is evaluated on a scale of 0 to 10, measuring the assistant's ability to maintain context and recall previous interactions:

$$M = \frac{1}{n} \sum_{i=1}^{n} m_i \tag{9}$$

where m_i represents individual memory assessments for n previous interactions, evaluated by an LLM judge.

3.2 Language Adaptation

The language score L measures the assistant's ability to adapt to the user's preferred language:

$$L = \sum_{i=1}^{k} w_i l_i \tag{10}$$

where:

- l_i represents different aspects of language adaptation
- w_i are weighting factors for each aspect
- k is the number of language adaptation criteria

3.3 Grice's Maxims Compliance

Following Grice's Cooperative Principle [4], we evaluate the assistant's adherence to four fundamental maxims:

1. **Maxim of Quantity**: Information should be as informative as required
2. **Maxim of Quality**: Information should be true and supported by evidence
3. **Maxim of Relation**: Information should be relevant to the conversation
4. **Maxim of Manner**: Information should be clear and unambiguous

The Gricean compliance score G is calculated as:

$$G = \frac{1}{4} \sum_{i=1}^{4} g_i \tag{11}$$

where g_i represents the compliance score for each maxim, evaluated on a scale of 0 to 1.

3.4 Sensibleness and Specificity

Based on the Sensibleness and Specificity Average (SSA) metric [5], we define a composite score:

$$SSA = \frac{S + Sp}{2} \tag{12}$$

where:

- S is the sensibleness score
- Sp is the specificity score

The sensibleness score S evaluates whether the response makes sense in the given context, while the specificity score Sp measures how specific and detailed the response is.

and γ_i are the respective weights that sum to 1.

4 Context Adherence Metrics

The evaluation of context adherence in AI assistant responses is crucial for ensuring relevant and appropriate interactions. This section presents a framework for assessing how well an assistant's responses align with the provided context and expected outcomes.

4.1 Context Evaluation Framework

The context evaluation process employs an LLM-as-a-judge approach, where a specialized language model (specifically deepseek-r1) evaluates the following components:

1. **Context**: The provided background information and conversation history
2. **Human Question**: The user's query or input
3. **Assistant Answer**: The actual response generated by the AI assistant
4. **Ground Truth/Observation**: The expected or ideal response

4.2 Evaluation Process

The evaluation process follows a structured approach:

1. **Context Analysis**: The judge model analyzes the provided context and its relevance to the conversation
2. **Response Assessment**: The assistant's answer is evaluated against the ground truth
3. **Scoring**: A numerical score is assigned based on context adherence
4. **Insight Generation**: The judge provides detailed reasoning for the assigned score

4.3 Chain-of-Thought Evaluation

The evaluation process is enhanced by the judge model's ability to provide its reasoning through Chain-of-Thought (CoT) analysis. This includes:

- Step-by-step reasoning about context relevance
- Analysis of response alignment with ground truth
- Identification of potential context mismatches
- Suggestions for improvement

4.4 Storage and Analysis

The evaluation results are stored in an Elasticsearch database, containing:

- Context adherence scores
- Generated insights
- Complete thinking process
- Ground truth comparisons
- Timestamps and metadata

This structured storage enables:

- Longitudinal analysis of context adherence
- Pattern identification in context mismatches
- Performance tracking over time
- Quality improvement opportunities

4.5 Integration with Other Metrics

The context adherence evaluation complements other metrics by providing:

- Additional validation of response quality
- Insights into context-aware performance
- Ground truth alignment verification
- Continuous improvement feedback

5 Toxicity Detection Through Clustering Analysis

The toxicity metric provides automated detection and quantification of toxic language patterns in AI assistant responses. Unlike the bias metric which focuses on fairness across protected attributes, the toxicity metric specifically identifies harmful, offensive, or inappropriate language through advanced clustering techniques and lexicon-based analysis.

5.1 Clustering Analysis and Latent Space Representation

The toxicity analysis incorporates a sophisticated clustering approach to identify patterns of toxic language in assistant responses. The process involves several key steps:

5.1.1 Embedding Generation

Responses are first converted into dense vector representations using a sentence transformer model:

$$E = \text{Transformer}(R) \tag{13}$$

where R represents the set of assistant responses and E is the resulting embedding matrix.

5.1.2 Dimensionality Reduction

The high-dimensional embeddings are reduced to a lower-dimensional space using UMAP (Uniform Manifold Approximation and Projection) [**?**]:

$$L = \text{UMAP}(E, n_{components}, n_{neighbors}, min_{dist}) \tag{14}$$

where:

- L is the reduced latent space representation
- $n_{components}$ is the target dimensionality (default: 2)
- $n_{neighbors}$ is the number of neighbors for local structure preservation
- min_{dist} is the minimum distance between points in the embedding

5.1.3 Cluster Detection

The latent space is then analyzed using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) [**?**]:

$$C = \text{HDBSCAN}(L, min_{cluster\_size}, \epsilon) \tag{15}$$

where:

- C represents the cluster assignments
- $min_{cluster\_size}$ is the minimum number of samples in a cluster
- ϵ is the cluster selection epsilon parameter

5.2 Toxicity Scoring

Each cluster is evaluated for potential toxicity using lexicon-based analysis combined with a Laplace-smoothed scoring mechanism. The toxicity score for cluster c is calculated as:

$$\text{ToxicityScore}(c) = \frac{n_{toxic}}{n_{total}} \tag{16}$$

where:

- n_{toxic} is the count of toxic words from the lexicon found in cluster c
- n_{total} is the total word count in cluster c

The toxicity lexicon used is the HurtLex database [**?**], which categorizes offensive terms across multiple dimensions including:

1. Negative stereotypes ethnic slurs
2. Profanity and vulgarity
3. Animal metaphors
4. Physical disabilities and diversity
5. Cognitive disabilities and diversity
6. Moral and behavioral defects

5.3 Clustering Validation

The quality of the clustering is assessed using silhouette scores:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{17}$$

where:

- $a(i)$ is the mean distance between sample i and all other points in the same cluster
- $b(i)$ is the mean distance between sample i and all points in the nearest cluster

6 Best-of-N Evaluation Through Tournament Selection

The BestOf metric implements a tournament-style evaluation framework to determine the optimal response among multiple candidates. This approach mirrors human evaluation processes and provides robust, comparative assessment of response quality.

6.1 Tournament Structure

For a set of n candidate responses $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$, the tournament is organized as a single-elimination bracket:

$$\text{Rounds} = \lceil \log_2(n) \rceil \tag{18}$$

Each match in round k is defined as:

$$M_k^{(i,j)} = \text{Judge}(r_i, r_j, q, c) \tag{19}$$

where:

- q is the user query
- c is the conversation context
- Judge is an LLM-based evaluation function

6.2 Pairwise Comparison

For each match, the judge evaluates both responses according to multiple criteria:

$$\text{Score}(r_i, r_j) = \sum_{k=1}^{m} w_k \cdot f_k(r_i, r_j) \tag{20}$$

where:

- f_k represents individual evaluation criteria (relevance, accuracy, coherence, helpfulness)
- w_k are the weights assigned to each criterion
- m is the total number of criteria

6.3 Winner Selection

The winner of each match is determined by:

$$\text{Winner}(r_i, r_j) = \begin{cases} r_i & \text{if Score}(r_i, r_j) > \text{Score}(r_j, r_i) \\ r_j & \text{otherwise} \end{cases} \tag{21}$$

Each evaluation includes:

- Confidence score $\in [0, 1]$

- Detailed verdict explanation
- Structured reasoning chain
- Comparative analysis of strengths and weaknesses

6.4 Metrics Output

The final BestOf metric produces:

$$\text{BestOfMetric} = \{w_{final}, \mathcal{C}, \theta\} \tag{22}$$

where:

- w_{final} is the tournament winner
- $\mathcal{C}$ is the complete set of contests with verdicts
- θ is the aggregate confidence score

References

[1] Plutchik, R. (2001). The nature of emotions. *American Scientist*, 89(4), 344-350.

[2] Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3), 436-465.

[3] IBM. (2024). Granite Guardian: Enterprise-grade risk detection model. *Hugging Face Model Hub*.

[4] Grice, H. P. (1975). Logic and conversation. *Syntax and Semantics*, 3, 41-58.

[5] Adolphs, L., et al. (2020). Evaluation of neural response generation models. *arXiv preprint arXiv:2001.09977*.

[6] DeepSeek. (2024). DeepSeek-R1: Advanced reasoning model. *DeepSeek AI Documentation*.

[7] McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*.

[8] McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11), 205.

[9] Bassignana, E., Basile, V., & Patti, V. (2018). Hurtlex: A multilingual lexicon of words to hurt. *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*.