User Engagement Metrics

June 13, 2025

The user engagement metrics are computed from interaction data across users, sessions, and time periods. Below are the mathematically defined metrics based on the logic in the code.

Daily Active Users (DAU)

Let D be the set of all days in the dataset, and U_d the number of unique users on day $d \in D$. Then:

$$\text{Average DAU} = \frac{1}{|D|} \sum_{d \in D} |U_d|$$

Monthly Active Users (MAU)

Let M be the set of months, and U_m the number of unique users in month $m \in M$. Then:

$$\text{Average MAU} = \frac{1}{|M|} \sum_{m \in M} |U_m|$$

Session Duration

For each session $s \in S$, define:

$$\text{Duration}(s) = \frac{t_{\text{end}}(s) - t_{\text{start}}(s)}{60} \quad \text{(in minutes)}$$

Then:

$$\text{Average Session Duration} = \frac{1}{|S'|} \sum_{s \in S'} \text{Duration}(s) \quad \text{where } |S'| > 1$$

Sessions Per User

Let S_u be the number of unique sessions by user u. Then:

$$\text{Average Sessions Per User} = \frac{1}{|U|} \sum_{u \in U} S_u$$

Queries Per Session

Let Q_s be the number of user-generated (human) queries in session s. Then:

$$\text{Average Queries Per Session} = \frac{1}{|S|} \sum_{s \in S} Q_s$$

Feature Usage Frequency

For each feature f, define its keyword set K_f. Let C be the set of all human message contents.

$$\text{Usage}(f) = \sum_{c \in C} \mathbf{1}\left[\exists k \in K_f : k \in c\right]$$

Retention Rate

Let T_u be the first activity date of user u, and A_u the set of all dates user u returned. Then:

$$\text{Retention}_{d\text{-day}} = \frac{1}{|U|} \sum_{u \in U} \mathbf{1}\left[\exists t \in A_u \text{ such that } t \geq T_u + d\right]$$

for $d \in \{1, 7, 30\}$.

Churn Rate

Let L_u be the last active date for user u, and T the latest date in the dataset.

$$\text{Churn Rate} = \frac{1}{|U|} \sum_{u \in U} \mathbf{1}[L_u < T - 30]$$

This quantifies the proportion of users who have not returned in the last 30 days.

LLM Evaluation Metrics

These metrics evaluate how closely an LLM-generated response matches the intent, content, and semantic expectations of the task.

Alignment Score

The alignment score measures how well a generated response satisfies predefined instructions or aligns with reference expectations. Let R be a set of reference rules or expected behaviors, and let $\alpha_i \in [0, 1]$ represent satisfaction of rule i. Then:

$$\text{Alignment Score} = \frac{1}{|R|} \sum_{i=1}^{|R|} \alpha_i$$

Each α_i can be derived from a rule match, a rubric, or a classifier output indicating correct alignment.

In practice, alignment is implemented by splitting the generated and reference texts into sentences, and calculating the average of maximum cosine similarities between each generated sentence and its best-matching reference sentence:

$$\text{Alignment} = \frac{1}{n}\sum_{i=1}^{n} \max_j \cos(\vec{t}_i, \vec{r}_j)$$

Coverage Score

Coverage score quantifies how much of the reference content is covered by the generated response. It is computed by combining:

- **Lexical coverage**: Fraction of important reference keywords present in the generated text.
- **Semantic coverage**: Cosine similarity between full embeddings of reference and generated texts.

Let K_{ref} be the set of key reference tokens (excluding stopwords), and K_{text} the set of tokens in the generated text.

$$\text{Coverage} = \frac{1}{2}\left(\frac{|K_{\text{ref}} \cap K_{\text{text}}|}{|K_{\text{ref}}|} + \cos(\vec{g}, \vec{r})\right)$$

Semantic Similarity

Semantic similarity measures overall meaning alignment between two texts by comparing their vector embeddings.

$$\text{Semantic Similarity} = \frac{\vec{g} \cdot \vec{r}}{||\vec{g}|| \cdot ||\vec{r}||}$$

Here $\vec{g}$ and $\vec{r}$ are sentence embeddings (e.g., from Sentence-BERT) of the generated and reference texts, respectively.

Metric Comparison Summary

Metric	**Level**	**Uses Cosine?**	**What it Measures**
Semantic Similarity	Full text	Yes	Overall semantic closeness between generated and reference content.
Alignment Score	Sentence-level	Yes	Average of best sentence-to-sentence matches across texts.
Coverage Score	Word-level + Text-level	Yes (partially)	Key concept recall (via keyword overlap) and semantic coverage.

LLM-as-a-Judge Metrics

When using an LLM as a judge, the model is asked to evaluate a generated answer in terms of the following criteria, each rated on a scale from 1 to 10:

- **Relevance (1–10):** Measures how closely the answer addresses the specific question asked, based on the provided context. A high score indicates direct and meaningful alignment with the question.
- **Coherence (1–10):** Reflects how logically and grammatically well-structured the answer is. A coherent answer should be internally consistent, easy to follow, and fluent.
- **Accuracy (1–10):** Assesses the factual correctness of the answer with respect to the available context. A high score indicates that the answer does not contain hallucinated or misleading information.

These scores are typically extracted from a structured response format like:

```
Relevance: X, Coherence: Y, Accuracy: Z
```