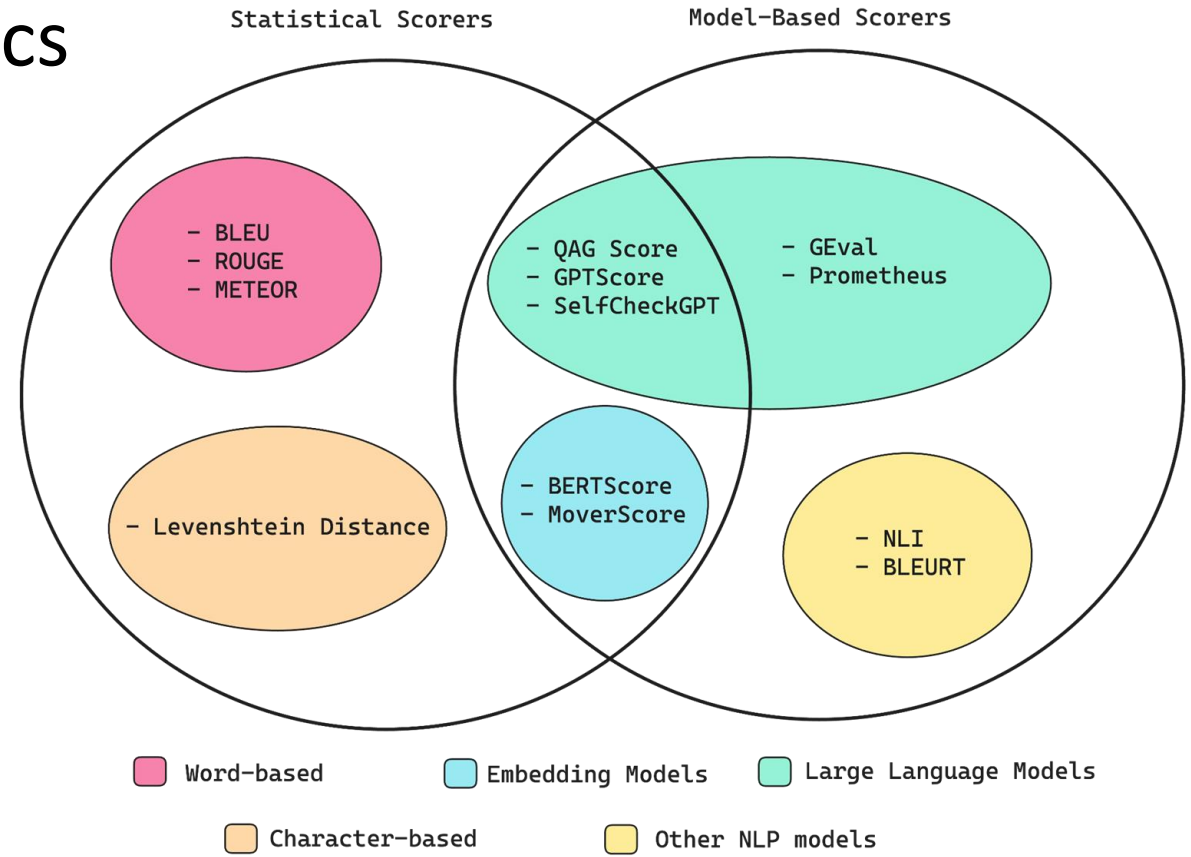
RIT

LLM Evaluation

Reza Fayyazi

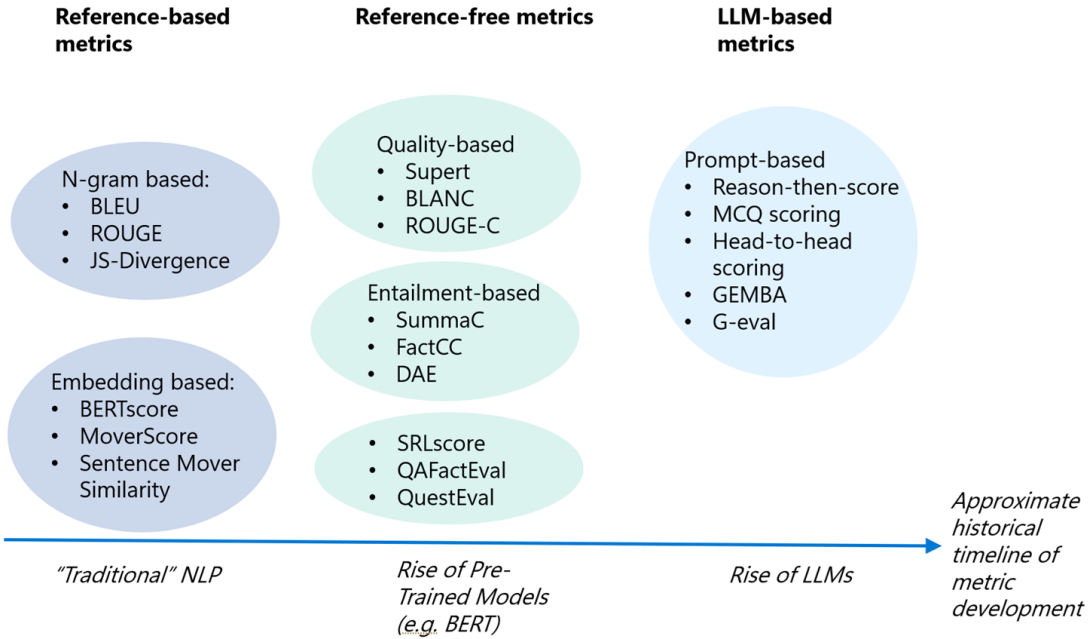
Evaluation Metrics

- An open research question



[1] https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation

Evaluation Metrics



[2] https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/generative-ai/working-with-llms/evaluation/list-of-eval-metrics

Benchmarks

Benchmarks	Description	Reference URL
GLUE Benchmark	GLUE (General Language Understanding Evaluation) benchmark provides a standardized set of diverse NLP tasks to evaluate the effectiveness of different language models	https://gluebenchmark.com/
SuperGLUE Benchmark	Compares more challenging and diverse tasks with GLUE, with comprehensive human baselines	https://super.gluebenchmark.com/
HellaSwag	Evaluates how well an LLM can complete a sentence	https://rowanzellers.com/hellaswag/
TruthfulQA	Measures truthfulness of model responses	https://github.com/sylinrl/TruthfulQA
MMLU	MMLU ((Massive Multitask Language Understanding) evaluates how well the LLM can multitask	https://github.com/hendrycks/test

[3] https://medium.com/data-science-at-microsoft/evaluating-llm-systems-metrics-challenges-and-best-practices-664ac25be7e5

BLEU Metric

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}.$$

[3] https://medium.com/data-science-at-microsoft/evaluating-llm-systems-metrics-challenges-and-best-practices-664ac25be7e5

Rouge Metric

$$\frac{\text{number of n-grams found in model and reference}}{\text{number of n-grams in reference}}$$

$$\frac{\text{count}_{\text{match}}(\text{gram}_n)}{\text{count}(\text{gram}_n)}$$

[3] https://medium.com/data-science-at-microsoft/evaluating-llm-systems-metrics-challenges-and-best-practices-664ac25be7e5

Levenshtein Similarity Ratio

The Levenshtein Similarity Ratio is a string metric for measuring the similarity between two sequences. This measure is based on Levenshtein Distance. Informally, the Levenshtein Distance between two strings is the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one string into the other. The Levenshtein Similarity Ratio can be calculated using Levenshtein Distance value and the total length of both sequences in the following definitions:

Levenshtein Similarity Ratio (Simple Ratio): $$Lev.ratio(a, b) = {(|a|+|b|)-Lev.dist(a,b) \over |a|+|b|}$$ where `|a|` and `|b|` are the lengths of `a` and `b`.

A few different methods are derived from **Simple Levenshtein Similarity Ratio:**

- **Partial Ratio:** Calculates the similarity by taking the shortest string, and compares it against the sub-strings of the same length in the longer string.
- **Token-sort Ratio:** Calculates the similarity by first splitting the strings into individual words or tokens, sorts the tokens alphabetically, and then recombines them into a new string. This new string is then compared using the simple ratio method.
- **Token-set Ratio:** Calculates the similarity by first splitting the strings into individual words or tokens, and then matches the intersection and union of the token sets between the two strings.

[3] https://medium.com/data-science-at-microsoft/evaluating-llm-systems-metrics-challenges-and-best-practices-664ac25be7e5

Semantic Similarity metrics

BERTScore, MoverScore, and Sentence Mover Similarity (SMS) metrics all rely on contextualized embeddings to measure the similarity between two texts. While these metrics are relatively simple, fast, and inexpensive to compute compared to LLM-based metrics, studies have shown that they can have poor correlation with human evaluators, lack of interpretability, inherent bias, poor adaptability to a wider variety of tasks and inability to capture subtle nuances in language.

The semantic similarity between two sentences refers to how closely related their meanings are. To do that, each string is first represented as a feature vector that captures its semantics/meanings. One commonly used approach is generating embeddings of the strings (for example, using an LLM) and then using **cosine similarity** to measure the similarity between the two embedding vectors. More specifically, given an embedding vector (A) representing a target string, and an embedding vector (B) representing a reference one, the cosine similarity is computed as follows:

$$ \text{cosine similarity} = {A \cdot B \over ||A|| ||B||}$$

As shown above, this metric measures the cosine of the angle between two non-zero vectors and ranges from -1 to 1. 1 means the two vectors are identical and -1 means they are dissimilar.

[3] https://medium.com/data-science-at-microsoft/evaluating-llm-systems-metrics-challenges-and-best-practices-664ac25be7e5

Reference-free Metrics

Reference-free (context-based) metrics produce a score for the generated text and **do not rely on ground truth**. Evaluation is based on the context or source document. Many of these metrics were developed in response to the challenge of creating ground truth data. These methods tend to be newer than reference-based techniques, reflecting the growing demand for scalable text evaluation as PTMs became increasingly powerful. These include quality-based, entailment-based, factuality-based, question-answering (QA) and question-generation (QG) based metrics.

- Quality-based metrics for summarization. These methods detect if the summary contains pertinent information. SUPERT quality measures the similarity of a summary with a BERT-based pseudo-reference, and BLANC quality measures the difference in accuracy of two reconstructions of masked-tokens. ROUGE-C is a modification of ROUGE without the need for references and uses the source text as the context for comparison.

- Entailment-based metrics. Entailment-based metrics are based on the Natural Language Inference (NLI) task, where for a given text (premise), it determines whether the output text (hypothesis) entails, contradicts or undermines the premise [24]. This can help to detect factual inconsistency.The SummaC (Summary Consistency) benchmark, FactCC, and DAE (Dependency Arc Entailment) metrics serve as an approach to detect factual inconsistencies with the source text. Entailment-based metrics are designed as a classification task with labels “consistent” or “inconsistent”.

- Factuality, QA and QG-based metrics. Factuality-based metrics like SRLScore (Semantic Role Labeling) and QAFactEval evaluate whether generated text contains incorrect information that does not hold true to the source text. QA-based, like QuestEval, and QG-based metrics are used as another approach to measure factual consistency and relevance.

[3] https://medium.com/data-science-at-microsoft/evaluating-llm-systems-metrics-challenges-and-best-practices-664ac25be7e5

Prompt-based evaluators

LLM-based evaluators prompt an LLM to be the judge of some text. The judgement can be based on (i) the text alone (reference-free), where the LLM is judging qualities like fluency, and coherence; (ii) the generated text, the original text, and potentially a topic or question (reference-free), where the LLM is judging qualities like consistency, and relevancy (iii) a comparison between the generated text and the ground truth (reference-based), where the LLM is judging quality, and similarity. Some frameworks for these evaluation prompts include Reason-then-Score (RTS), Multiple Choice Question Scoring (MCQ), Head-to-head scoring (H2H), and G-Eval (see the page on Evaluating the performance of LLM summarization prompts with G-Eval). GEMBA is a metric for assessing translation quality.

LLM-evaluation is an emerging area of research and has not yet been systematically studied. Already, researchers have identified issues with reliability in LLM evaluators such as positional bias, verbosity bias, self-enhancement bias, limited mathematical and reasoning capabilities, and issues with LLM success at assigning numerical scores. Strategies that have been proposed to mitigate positional bias include Multiple Evidence Calibration (MEC), Balanced Position Calibration (BPC), and Human In The Loop Calibration (HITLC).

[3] https://medium.com/data-science-at-microsoft/evaluating-llm-systems-metrics-challenges-and-best-practices-664ac25be7e5

Thank you!

Reza Fayyazi
rf1679@rit.edu