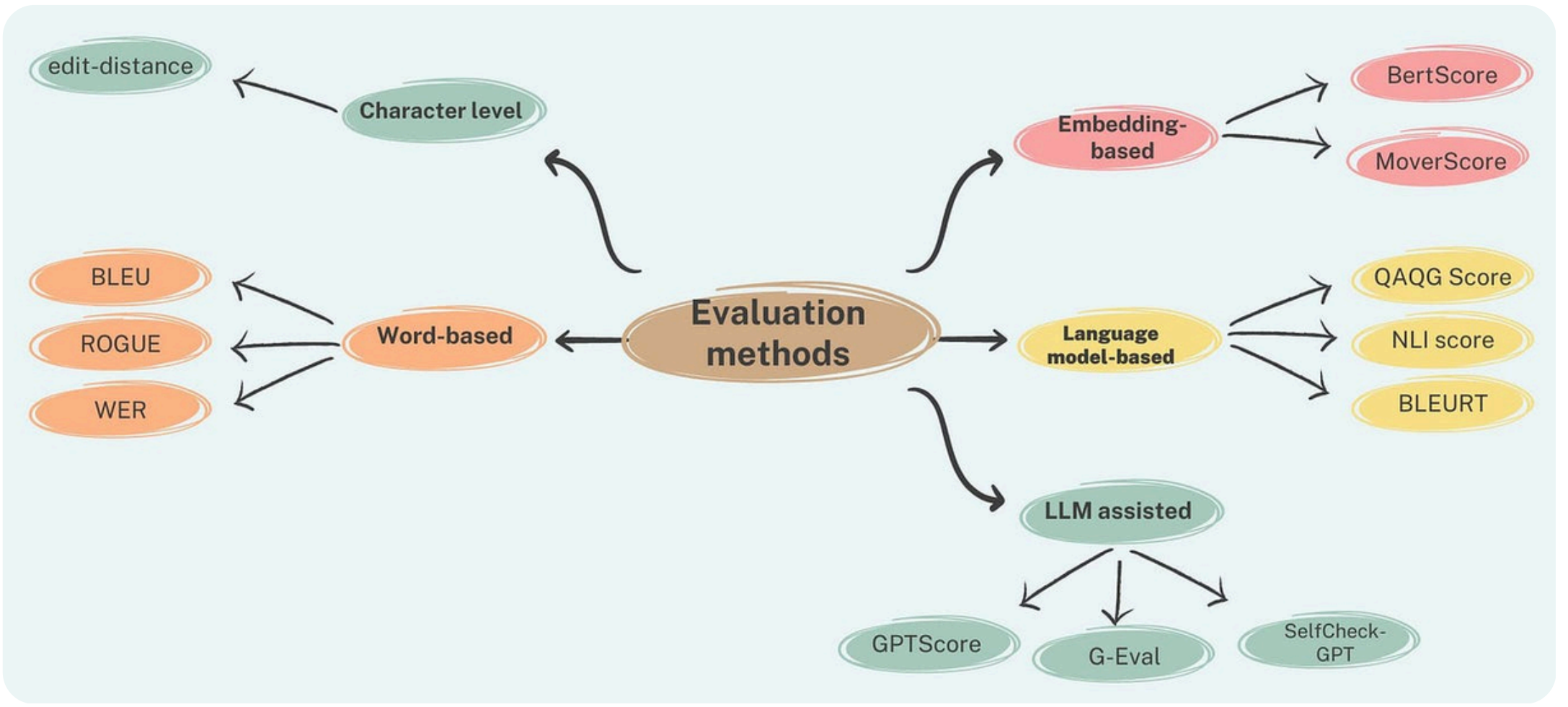
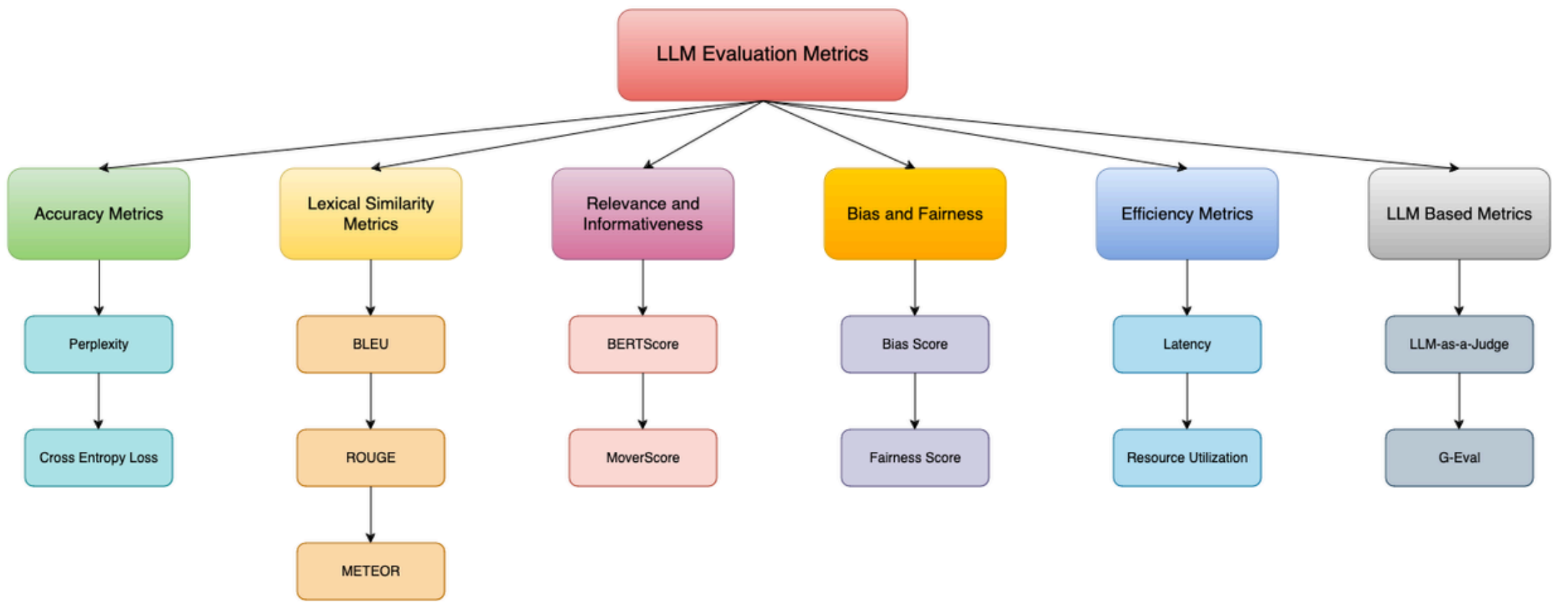
Analytics Vidhya

Mastering LLMs

Day 36: How to Measure LLM Performance





Key Evaluation Metrics

Perplexity (PPL)

Perplexity measures how well a language model predicts a given dataset. A lower perplexity score indicates better performance.

$$PPL = e^{-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i)}$$

Where:

- *N* the number of words in the dataset,
- P(Wi) the probability assigned to the word by the model.

Use Case: Good for evaluating probabilistic language models but may not always reflect human-perceived quality.

BLEU (Bilingual Evaluation Understudy)

BLEU measures how closely a generated text matches a reference text by comparing n-grams.

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Where:

- *BP* the brevity penalty,
- *Pn* the precision of n-gram matches,
- *Wn* are the weights for different n-grams.

Use Case: Commonly used for translation and summarization tasks but does not account for semantics.

ROUGE

ROUGE is an abbreviation of Recall-Oriented Understudy for Gisting Evaluation).

ROUGE compares generated text with a reference by computing overlap in n-grams and sequences.

- **ROUGE-N**: Measures n-gram overlap.
- **ROUGE-L**: Considers longest common subsequence.
- **ROUGE-W**: Weighted version of ROUGE-L.

Use Case: Suitable for summarization tasks.

METEOR

ROUGE is an abbreviation of Metric for Evaluation of Translation with Explicit ORdering.

METEOR improves on BLEU by considering synonyms, stemming, and word order.

Use Case: Translation and summarization.

BERTScore

BERTScore uses contextual embeddings from transformer models (like BERT) to compare generated and reference texts.

Use Case: More effective than BLEU or ROUGE for capturing semantic similarity.

Exact Match (EM) and F1 Score

For tasks with exact answers, such as question-answering, EM checks if the output perfectly matches the reference, while F1 measures partial overlap.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Use Case: Useful for question-answering systems.

Human Evaluation

Since automated metrics may not fully capture language nuances, human evaluation is critical.

- **Fluency**: Is the text grammatically correct and natural?
- **Coherence**: Does the response make logical sense?
- **Relevance**: Does the model answer the query correctly?
- **Bias and Toxicity**: Does the model generate biased or harmful content?

Analytics Vidhya

Stay Tuned for **Day 37** of

Mastering LLMs