GenAI metrics

you should know

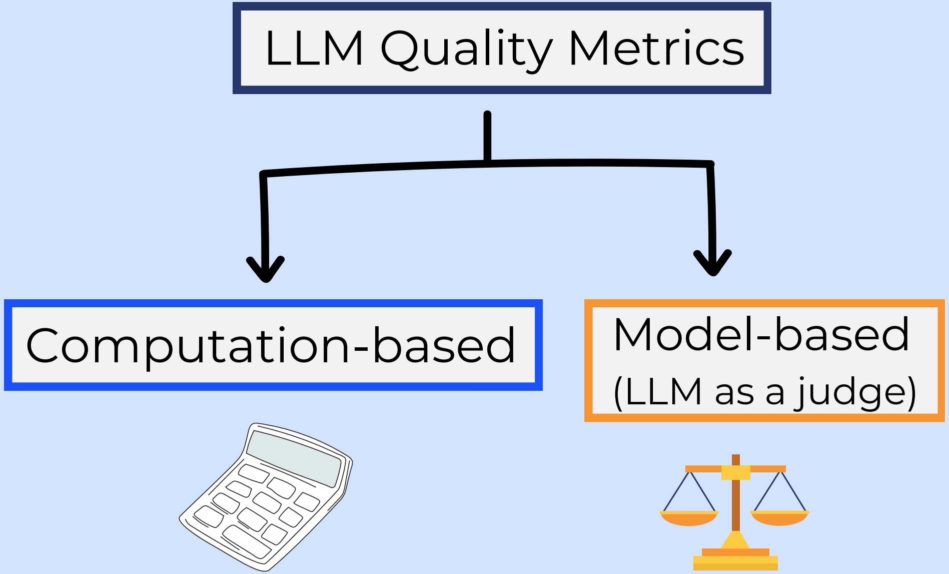
A quick guide by:

Paula Rodriguez

AI Engineer

1. Gen AI quality evaluations

- To automatically measure how well an **LLM performs a task with a non structured output**, there are two types of metrics we can use:



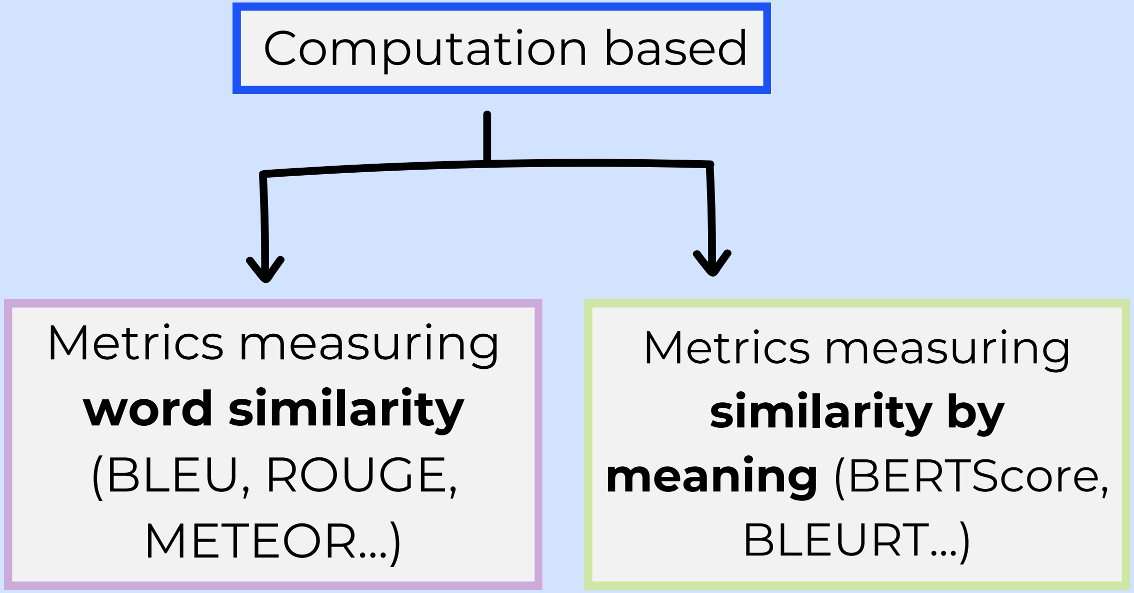
LLM Quality Metrics

Computation-based

Model-based (LLM as a judge)

2. Computation-based metrics

- Computation-based metrics are normally divided into two groups:



Computation based

Metrics measuring **word similarity** (BLEU, ROUGE, METEOR...)

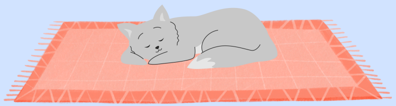
Metrics measuring **similarity by meaning** (BERTScore, BLEURT...)

2.1 Metrics focused on word similarity

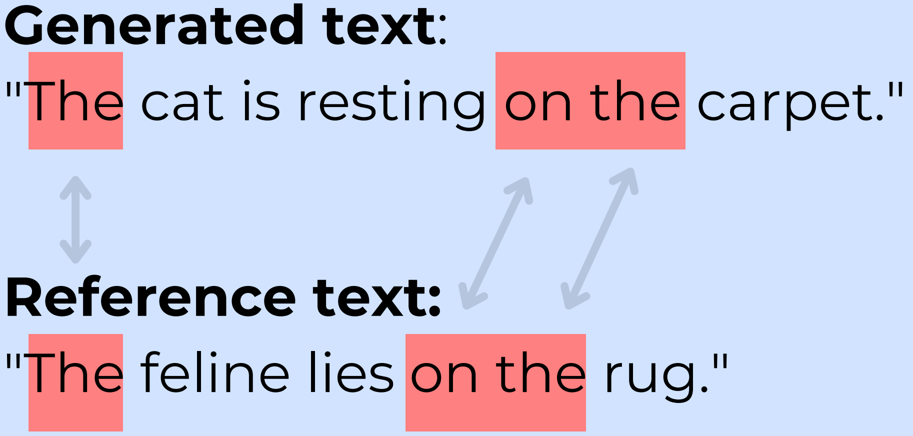
- These metrics **need a ground truth** to compare the generated text against. They calculate similarity by checking **matching words** or phrases.
- Used to automatically measure the outputs of a model for tasks like:
  - Summarization
  - Translation
  - Question answering
  - Content generation.
- Common metrics include **BLEU, ROUGE and METEOR**, but only METEOR includes synonyms and other techniques to **capture the actual meaning** of the words.

	BLEU	ROUGE	METEOR
Compares **matching words** or phrases	✓	✓	✓
Original **use case**	Translation	Summarization	Translation
Needs **ground truth**	✓	✓	✓
Considers **meaning**	✗	✗	✓

A practical example:



- Let’s say we want to measure the quality of a text generated by a LLM.
- So, we use a ground truth (a perfect response) and metrics like ROUGE, BLEU and METEOR.



Generated text:

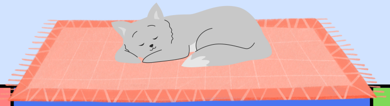
"The cat is resting on the carpet."

Reference text:

"The feline lies on the rug."

🟥 **Matching words** in generated and reference text

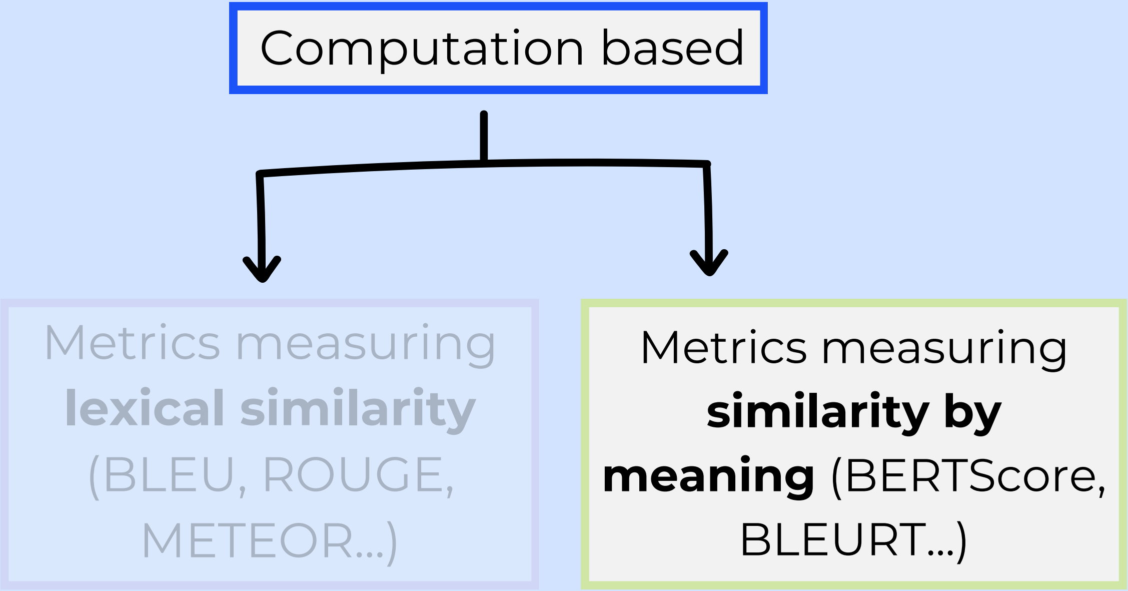
- This would be the results (0 indicates very different, 1 indicates exactly the same):
- As you can see, BLEU and ROUGE fail to capture the similarity between the two sentences since there are **few matching words**.



ROUGE score	BLEU score	METEOR score
0.3683	≈0.0000	0.7934

METEOR metric **better captures** the similarity in these two sentences.

- Other ways to capture meaning when evaluating the similarity of two texts, is by using metrics based on meaning, like BERTScore and its derivatives.



Computation based

Metrics measuring **lexical similarity** (BLEU, ROUGE, METEOR...)

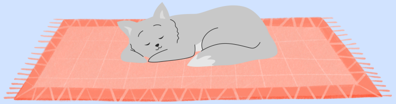
Metrics measuring **similarity by meaning** (BERTScore, BLEURT...)

2.2 Metrics focused on similarity by meaning

- These metrics also **need a ground truth** to compare the generated text against.
- They calculate similarity by **measuring the distance between vectors**. These vectors represent, using numbers, the meaning a text.
- Used to automatically measure the outputs of a model for tasks like:
    - Summarization
    - Translation
    - Question answering
    - Content generation.
- Common metrics include **BERTScore** and its derivatives like BLEURT, Sentence-BERT...

	BERTScore
Compares **text by meaning**	✓
Original **use case**	Text generation in general
Needs **ground truth**	✓
Can be used for every **language**	✗

Let’s compare the results:



- We’ve now included BERTScore metric for the past example:
- As you can see, **BERTScore better captures the similarity** between texts although the wording is different.
- It is always a good idea to **test different metrics** to see which one better fits your use case.

ROUGE score	BLEU score	METEOR score	BERTScore
0.3683	≈0.0000	0.7934	**0.8001**

Voilá!

Now you better understand GenAI metrics!

Paula Rodriguez

AI Engineer

Follow for easy & visual weekly AI posts