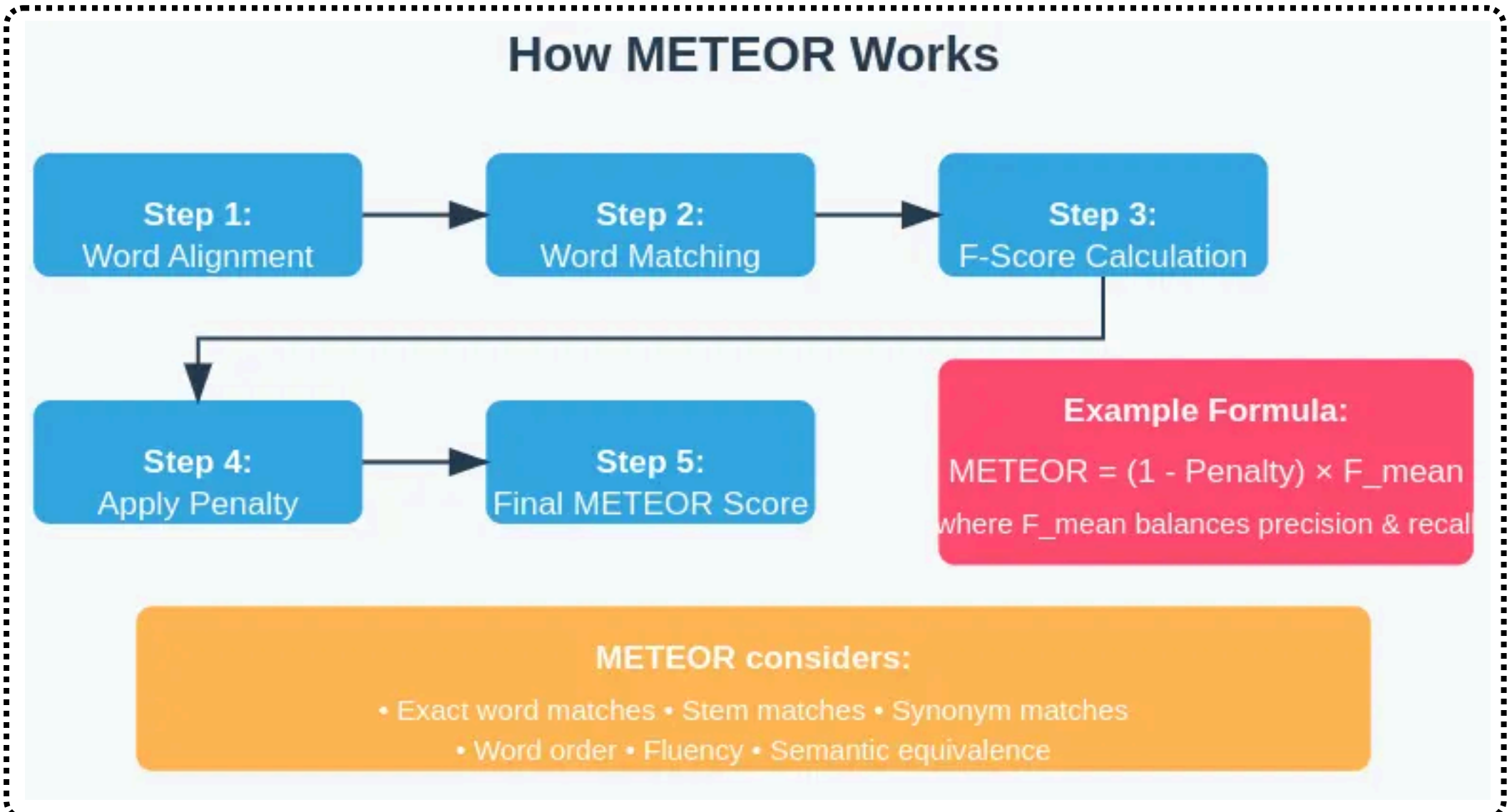
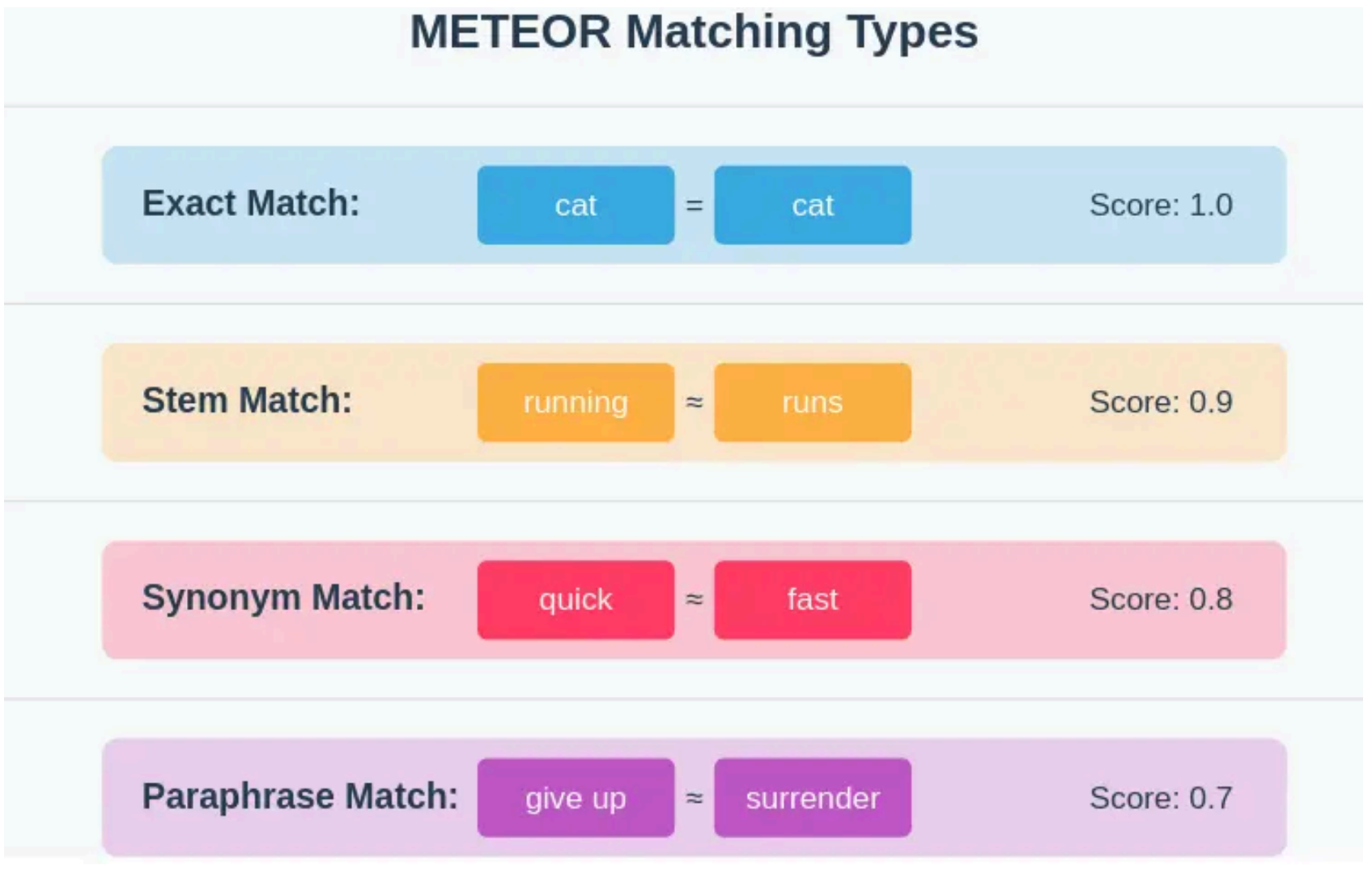
Analytics Vidhya

How METEOR Improves AI Text Evaluation





What is a METEOR Score?

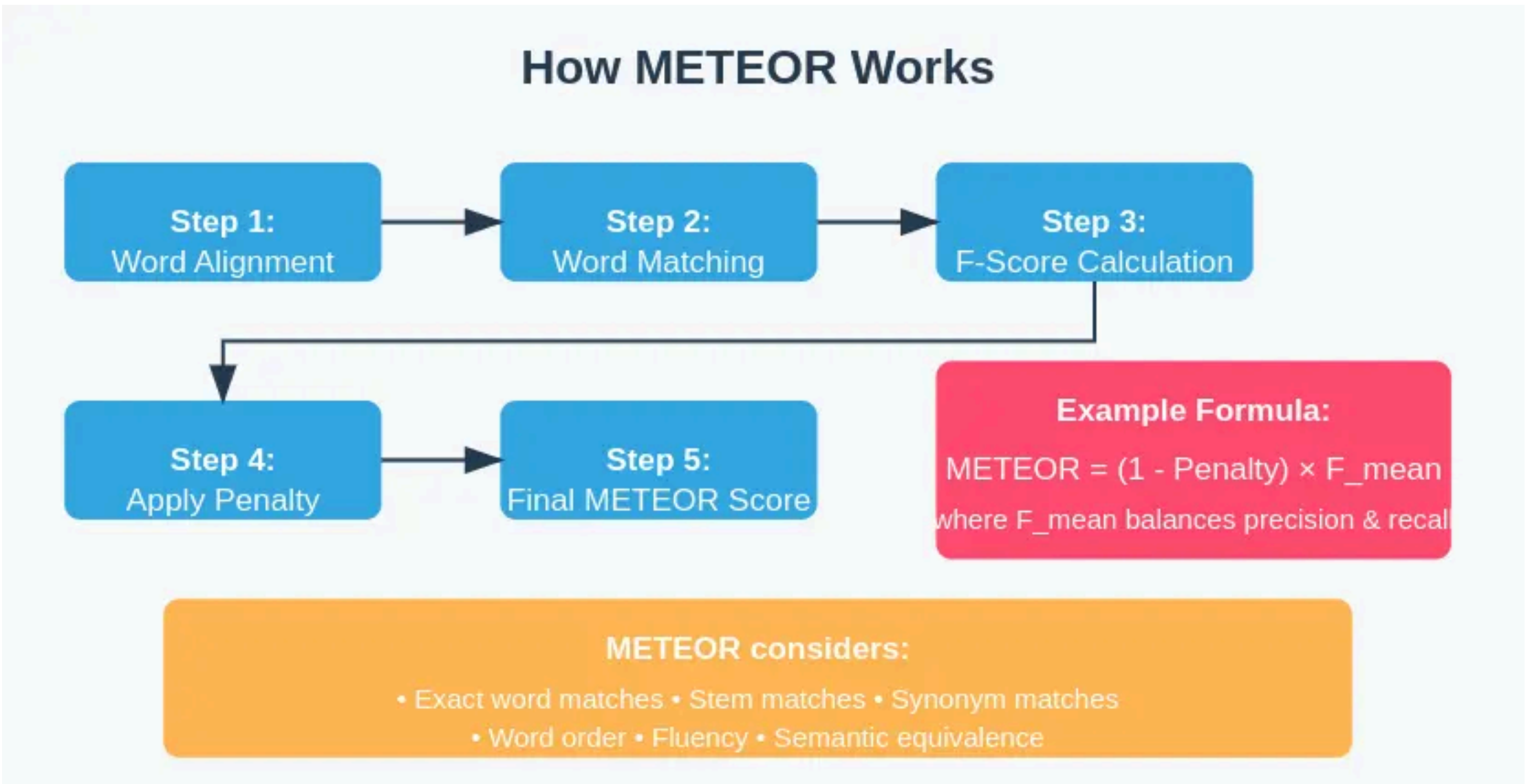
METEOR (Metric for Evaluation of Translation with Explicit Ordering) is an NLP evaluation metric originally designed for machine translation but now widely used for evaluating various natural language generation tasks, including those performed by **Large Language Models** (LLMs).

Unlike simpler metrics that focus solely on exact word matches, METEOR was developed to address the limitations of other metrics by incorporating semantic similarities and alignment between a machine-generated text and its reference text(s).

Quick Check: Think of METEOR as a sophisticated judge that doesn't just count matching words but understands when different words mean similar things!

How Does METEOR Work?

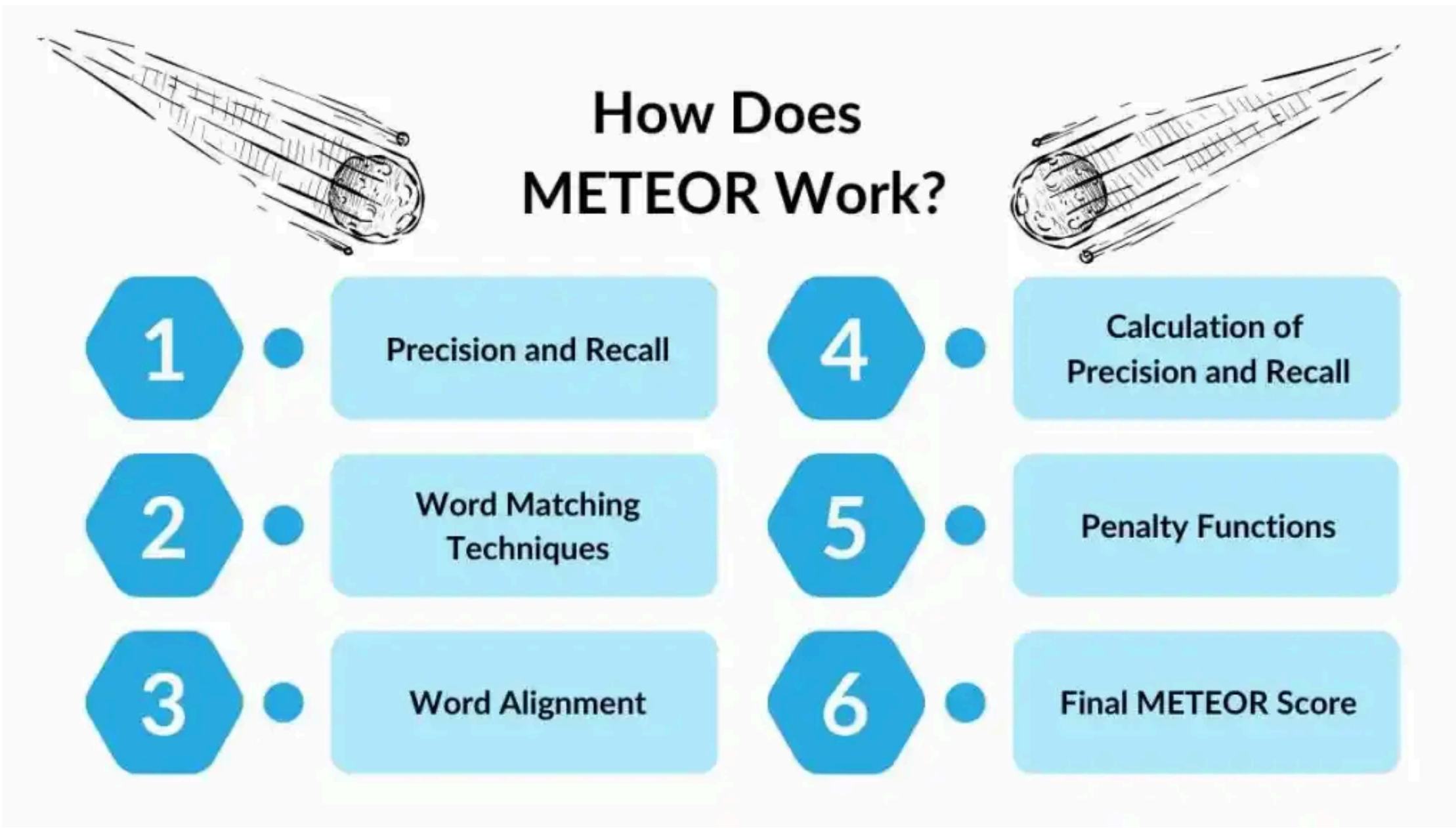
METEOR evaluates text quality through a step-by-step process:



- Alignment: First, METEOR creates an alignment between the words in the generated text and reference text(s).

- Matching: It identifies matches based on:
  - Exact matches (identical words)
  - Stem matches (words with the same root)
  - Synonym matches (words with similar meanings)
  - Paraphrase matches (phrases with similar meanings)

- **Scoring**: METEOR calculates precision, recall, and a weighted F-score.
- **Penalty**: It applies a fragmentation penalty to account for word order and fluency.
- **Final Score**: The final METEOR score combines the F-score and penalty.



It improves upon older methods by incorporating:

- **Precision & Recall**: Ensures a balance between correctness and coverage.
- **Synonyms Matching**: Identifies words with similar meanings.
- **Stemming**: Recognizes words in different forms (e.g., “run” vs. “running”).
- **Word Order Penalty**: Penalizes incorrect word sequence while allowing slight flexibility.

Try It Yourself: Consider these two translations of a French sentence:

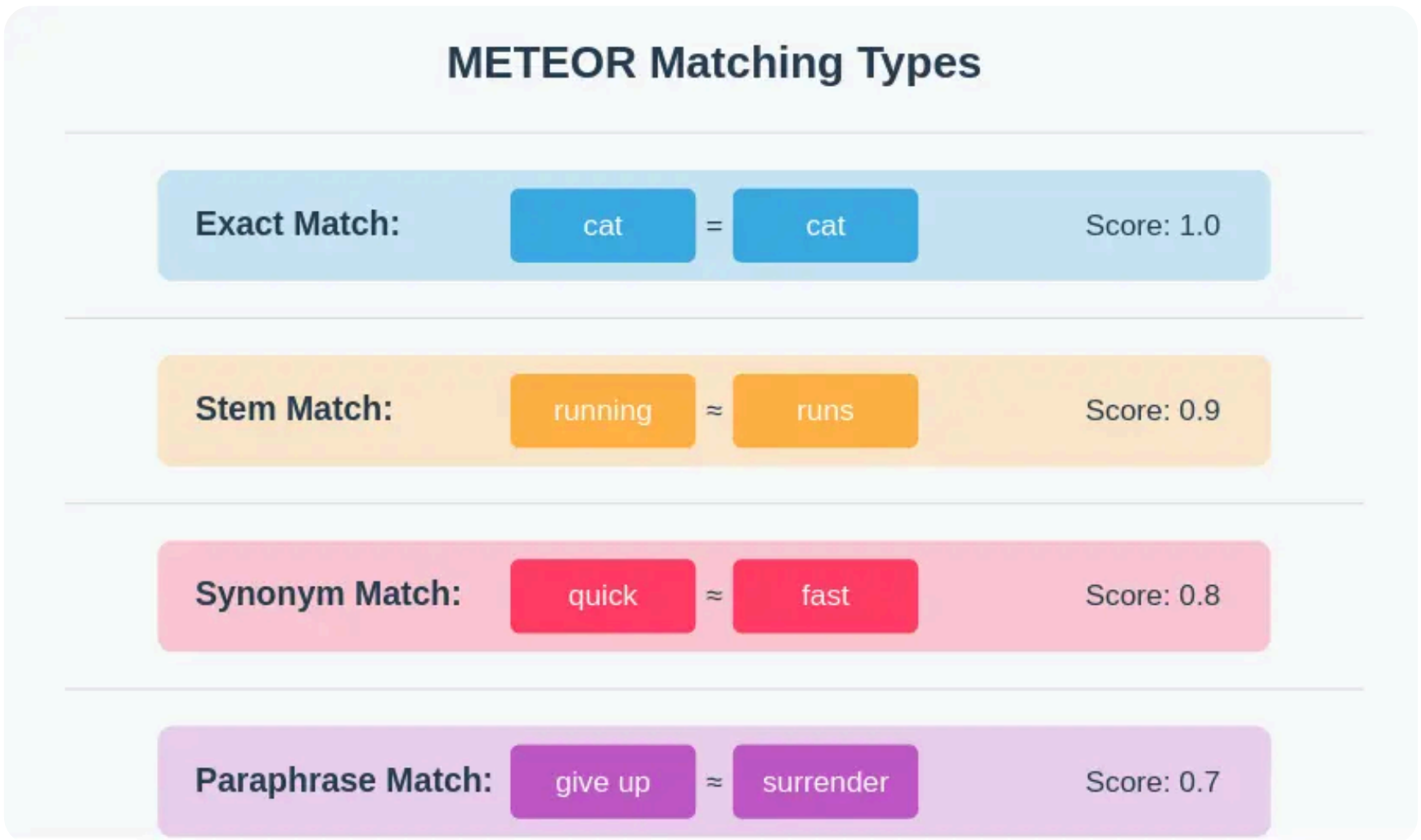
- **Reference**: “The cat is sitting on the mat.”
- **Translation A**: “The feline is sitting on the mat.”
- **Translation B**: “Mat the one sitting is cat the.”

Which do you think would get a higher METEOR score? (Translation A would score higher because while it uses a synonym, the order is preserved. Translation B has all the right words but in a completely jumbled order, triggering a high fragmentation penalty.)

Key Features of METEOR

METEOR stands out from other evaluation metrics with these distinctive characteristics:

- **Semantic Matching**: Goes beyond exact matches to recognize synonyms and paraphrases
- **Word Order Consideration**: Penalizes incorrect word ordering
- **Weighted Harmonic Mean**: Balances precision and recall with adjustable weights
- **Language Adaptability**: Can be configured for different languages
- **Multiple References**: Can evaluate against multiple reference texts



For more information, kindly visit this article

Intermediate LLMs NLP

How METEOR Improves AI Text Evaluation?

Discover how METEOR improves AI text evaluation by considering word order, semantics, and paraphrasing for accurate assessment.