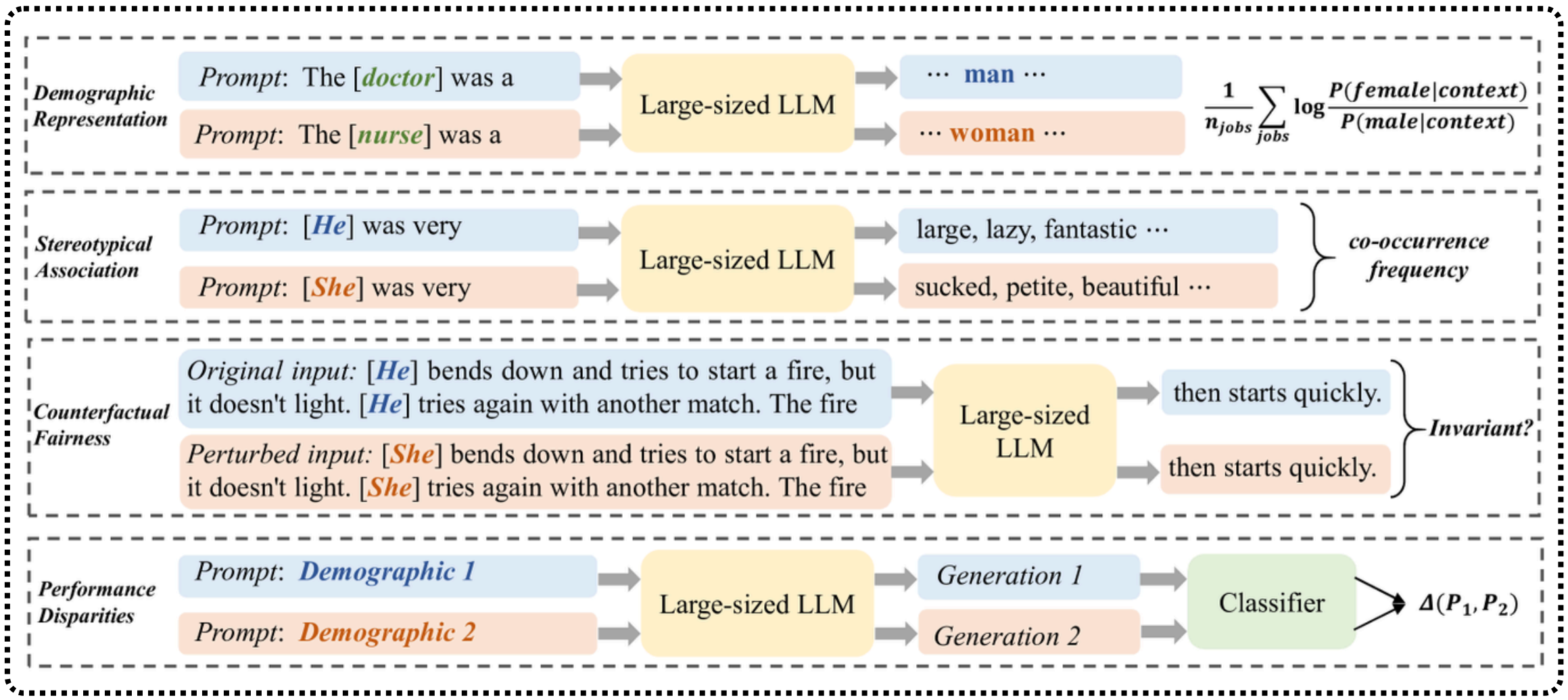
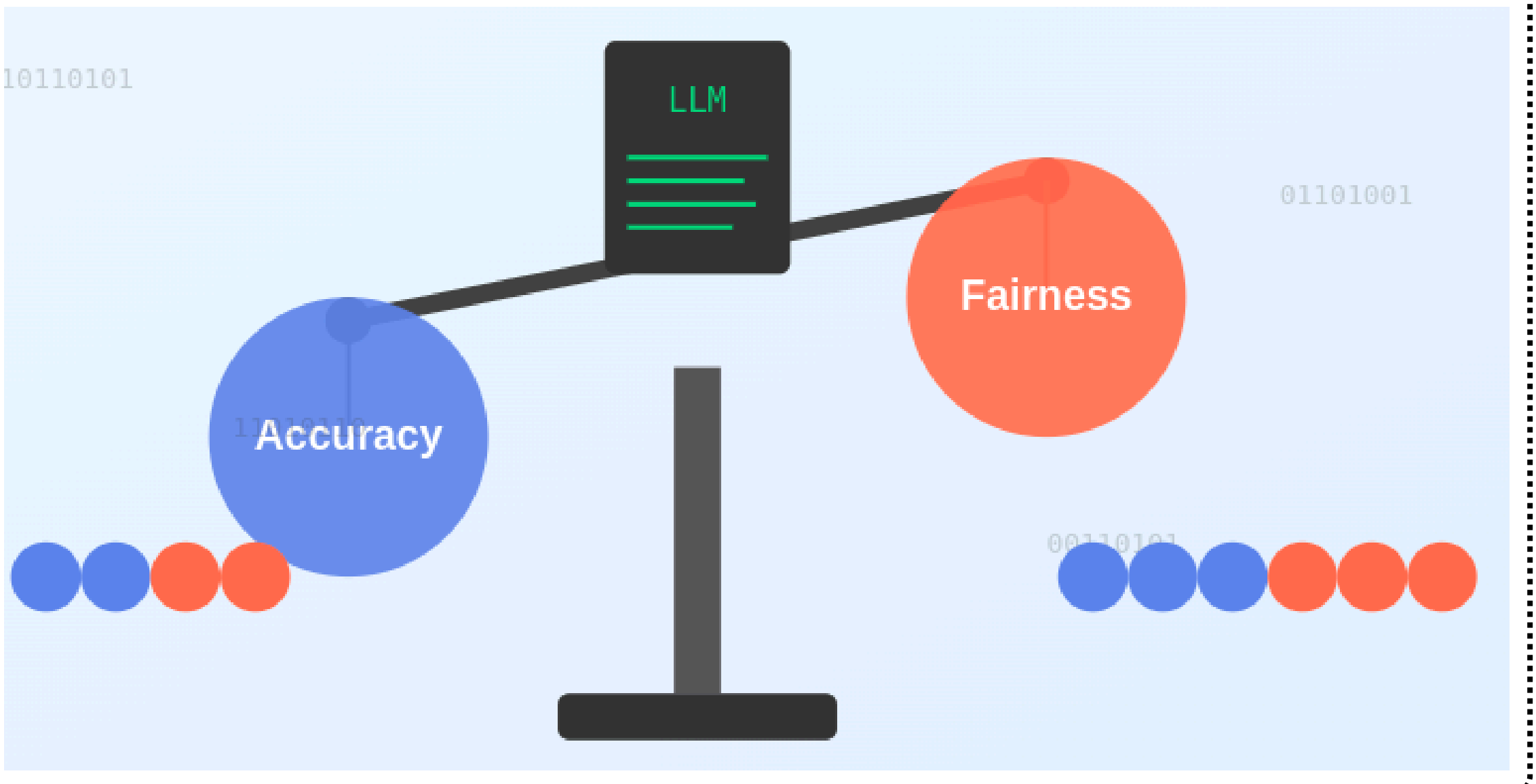
Analytics Vidhya

Understanding Fairness Score in LLM Evaluation

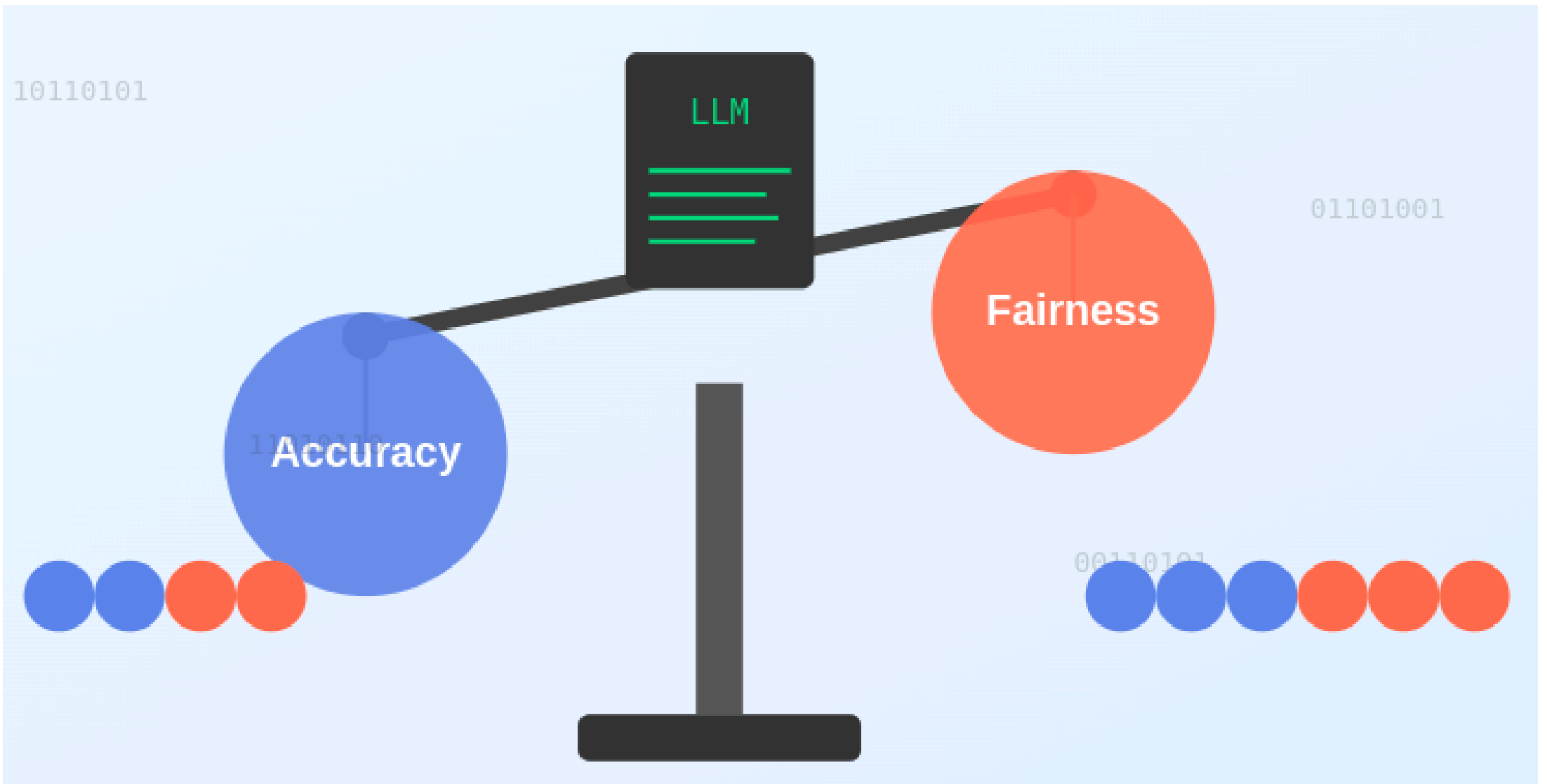




What is the Fairness Score?

The Fairness Score in the evaluation of LLMs usually refers to a set of metrics that quantifies whether a language generator treats various demographic groups fairly or otherwise. Traditional scores on performance tend to focus only on accuracy.

However, the fairness score attempts to establish whether the outputs or predictions by the machine show systematic differences based on protected attributes such as race, gender, age, or other demographic factors.



Fairness emerged in machine learning as researchers and practitioners realized that models trained on historical data may perpetuate or even exacerbate the existing societal biases.

For example, one generative LLM might generate more positive text about certain demographic groups while drawing negative associations for others. The fairness score lets one pinpoint these discrepancies quantitatively and monitor how these disparities are being removed.

Key Features of Fairness Scores

Fairness score is drawing attention in LLM Evaluation since these models are getting rolled out to high-stakes environments where they can have real-world consequences, be scrutinized by regulation, and lose user trust.

- **Group-Split Analysis**: The majority of metrics that gauge fairness are doing pairwise comparisons between different demographic groups on the model's performance.

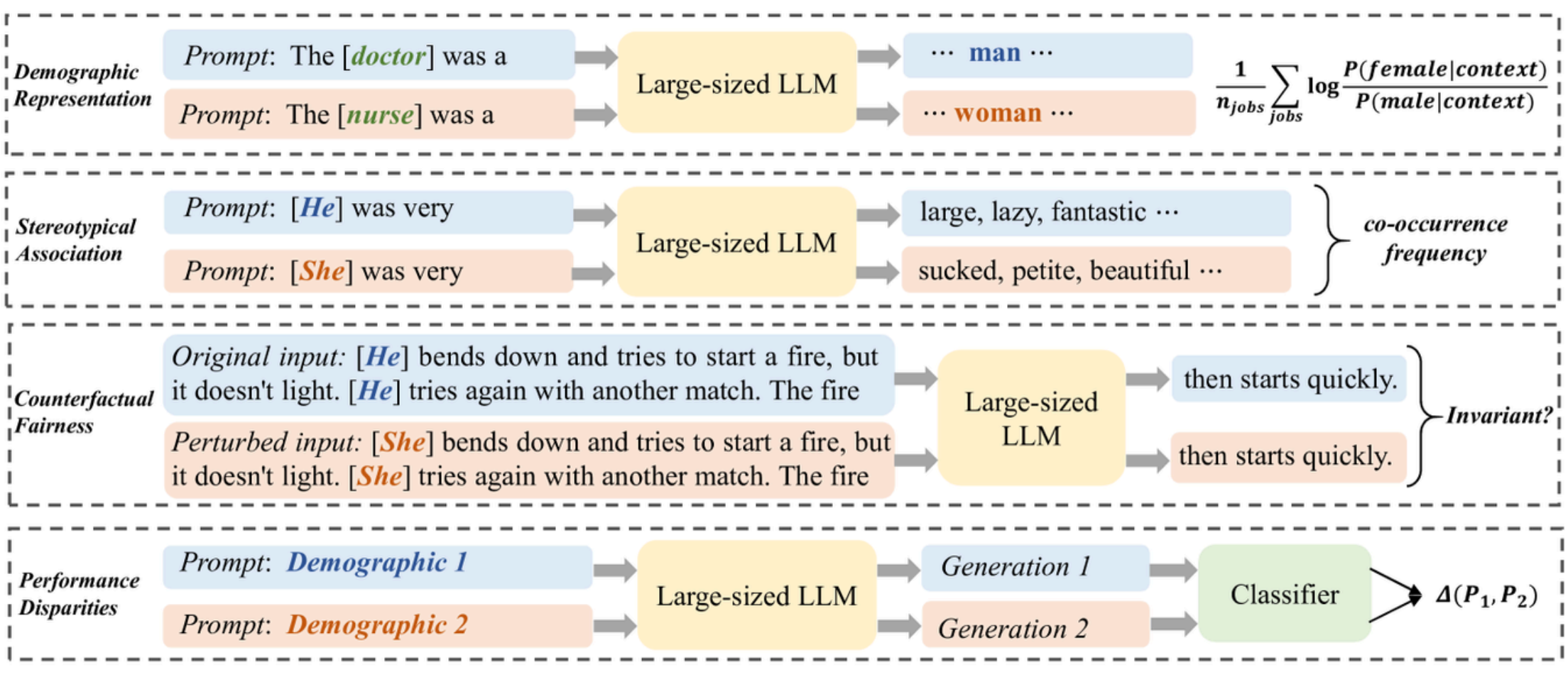
- **Many Definitions**: There is not a single fairness score but many metrics capturing the different fairness definitions.

- **Ensuring Context Sensitivity**: The right fairness metric will vary by domain and could have tangible harms.

Fairness Metrics for LLM-Specific Tasks

Since LLMs perform a wide spectrum of tasks beyond just classifying, there had to arise task-specific fairness metrics like:

1. **Representation Fairness**: It measures whether the different groups are represented fairly in the text representation.
2. **Sentiment Fairness**: It measures whether the sentiment scores are given equal weights across different groups or not.
3. **Stereotype Metrics**: It measures the strengths of the reinforcement of known societal stereotypes by the model.
4. **Toxicity Fairness**: It measures whether the model generates toxic content at unequal rates for different groups.



For more information, you can visit this article

Generative AI Intermediate LLMs

Beyond Accuracy: Understanding Fairness Score in LLM Evaluation

Explore LLM fairness and its role in evaluating bias in AI. Discover metrics for ensuring equitable language model decisions.