Title: Realistic Evaluations of Bias in Large Language Models: Correlating Lexical and Semantic Metrics with Stereotype-Based Tasks

Research Question:
Can generalizable lexical and semantic language-based metrics effectively predict bias in large language models (LLMs) when compared to more specific stereotype-based tasks used in fairness and bias evaluations?

Scope of the Project:
This research aims to bridge the gap between specific stereotype-based bias evaluations, such as those conducted using the StereoSet dataset, and more general, easily computable lexical and semantic metrics. By investigating the correlations between these two types of evaluations, this project seeks to determine whether simpler, generalizable metrics can reliably indicate bias across different LLMs in a manner that aligns with realistic use-case evaluations.

Context of Prior Work:
Existing literature highlights the limitations of decontextualized bias benchmarks, often referred to as "trick tests," which fail to consistently predict real-world bias manifestations. A recent study on gender-occupation bias revealed that selecting the least biased model based on these trick tests correlated with real-world performance only as frequently as random chance (Lum et al., 2024). This raises concerns about the adequacy of such benchmarks in evaluating and mitigating bias in LLMs, emphasizing the necessity for evaluations grounded in realistic usage and tangible effects (RUTEd).

Building upon this foundation, the proposed research will explore whether there is a measurable correlation between general lexical and semantic metrics (e.g., Token Frequency Distribution, Syntactic Complexity, LIWC metrics) and the outcomes of stereotype-based bias evaluations (e.g., gender, race, religion, and profession biases) conducted on LLM outputs. While these metrics have been extensively utilized in natural language processing (NLP) for various purposes, they have not been systematically tested for their predictive power concerning LLM bias in realistic scenarios.

Plan to Answer the Research Question:

1. **Data Collection**:
    - Extract output responses from various LLMs using the StereoSet dataset, which evaluates bias across dimensions such as gender, profession, race, and religion.
    - Collect a diverse set of outputs from multiple LLMs to ensure broad applicability.
2. **Metric Calculation**:
    - Calculate a series of generalizable lexical and semantic metrics for each model output, including:

i. **Token Frequency Distribution**: Analyze the frequency distribution of tokens to identify skewed word usage indicative of bias.
ii. **LIWC Emotional Valence**: Measure sentiment polarity across different affective categories to assess emotional bias.
iii. **Syntactic Complexity Metrics**: Compute metrics such as average dependency length and parse tree depth to evaluate structural differences across biased outputs.
iv. **Lexical Formality Index**: Quantify the degree of formality in language use, particularly in contexts where informal language may signal bias.
v. **Readability Scores (Flesch-Kincaid, SMOG)**: Utilize multiple readability indices to determine whether certain biases correlate with more or less readable text.
vi. **HTML Tag Utilization**: Specifically measure the occurrence and significance of `<strong>` and other emphasis tags within text to observe if bias affects how content is emphasized.
vii. **Pronoun Usage Patterns**: Calculate the frequency and distribution of pronouns, with a particular focus on first-person plural ('we') to detect collective identity bias.
viii. **Type-Token Ratio (TTR)**: Measure lexical diversity by computing the ratio of unique words (types) to total words (tokens).
ix. **LIWC Domain-Specific Metrics**: Expand the analysis to include domain-specific LIWC categories such as social, cognitive, and perceptual processes to capture nuanced semantic biases.

3. **Evaluation and Analysis**:
    - Conduct statistical analyses to identify correlations between the general metrics and the performance of LLMs on stereotype-based tasks.
    - Assess whether these metrics can predict bias in outputs and whether they align with the more complex stereotype-based evaluations.
4. **Comparison and Validation**:
    - Compare the predictive power of the general metrics with existing stereotype-based evaluations to determine their efficacy in bias prediction.
    - Validate findings by applying the metrics to a different set of LLMs and bias evaluation datasets to check for consistency.

Contribution to the Field:
This research will contribute to the understanding of bias detection in LLMs by demonstrating the effectiveness of RUTEd evaluations, providing a more realistic approach to bias detection. It offers new strategies for more accurate and context-aware bias evaluation and mitigation in NLP models.