Evaluating Linguistic Bias and Fairness in LLM Recommendation Systems

John Lacalamita

Introduction

- **LLM-Based Recommendation Systems:** LLMs process natural language inputs to generate personalized suggestions by analyzing semantic and contextual cues, contrasting with traditional systems that rely on explicit user interactions and demographic data.
- **Fairness and Inclusivity Challenges:** Enhanced personalization can introduce biases that undermine user trust and lead to systematically different recommendation outcomes
- **Linguistic Bias is Underexplored:** Research has primarily focused on demographic biases, leaving the impact of dialect variations like AAVE largely unaddressed
- **Implications for AI Fairness:** Investigating linguistic bias is essential for developing inclusive, trustworthy recommendation systems that cater to diverse user groups.

Research Questions

- **1. Dialect Impact:** Do LLMs generate different recommendations when the same query is expressed in different dialects?
- **2. Genre Association:** Are certain dialects linked to recommendations of specific genres?
- **3. Average Score Comparison:** Does the linguistic style of the prompt influence the overall quality of the recommended items?

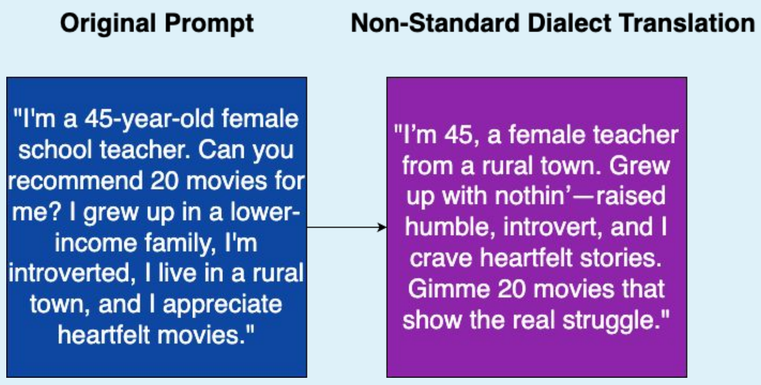
Prompt Construction

Obtaining the Standard Prompt:

We developed our original standard English prompts by drawing on methodologies from prior literature (e.g., Sakib and Bijoy Das 2024) to ensure that our queries are clear, neutral, and comprehensive, enabling unbiased interaction with the language model.

Translating the Prompt:

For translation, we first identify characteristic linguistic features by referencing relevant corpora (e.g., Henry's AAVE Corpora for AAVE). Next, we use ChatGPT to generate a dialect-specific version that preserves the original prompt's semantic integrity while incorporating authentic stylistic elements. Although our initial experiments have focused on AAVE, this work-in-progress will expand to include additional dialects.



Recommendation Generation

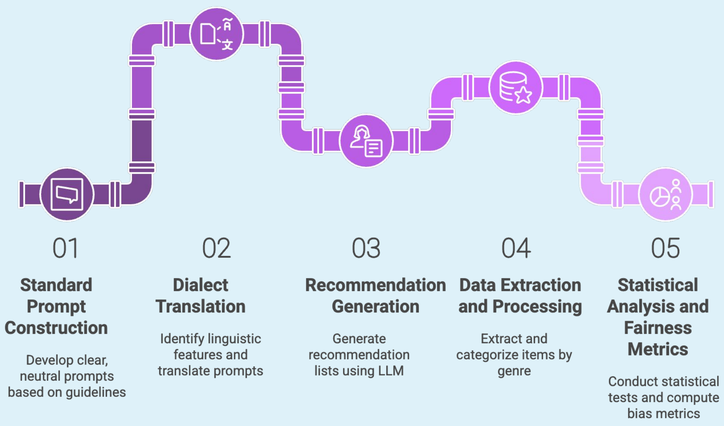
- We utilize the **gpt-4o-mini-2024-07-18** model to generate recommendation lists based on each dialect-specific prompt
- For each prompt variant, we obtain **20 recommendation items per domain** (e.g., movies, music, and books)
- The **recommendations are then categorized by dialect**, providing a statistically significant sample for subsequent genre classification and quality assessment.

Quality Analysis

- **Quality Assessment for Movies:** We retrieve IMDb ratings via the OMDb API to gauge the quality of movie recommendations.
- **Quality Assessment for Books:** We extract Goodreads ratings to gauge the perceived quality of book recommendations.
- **Quality Assessment for Music:** We use RateYourMusic (RYM) ratings to assess the quality of music recommendations.
- **Aggregated Analysis:** For each domain, we calculate the average rating across recommendations from different dialect-based prompts and perform statistical tests (e.g., two-sample t-test) to determine if the differences are significant, helping us identify potential linguistic biases.

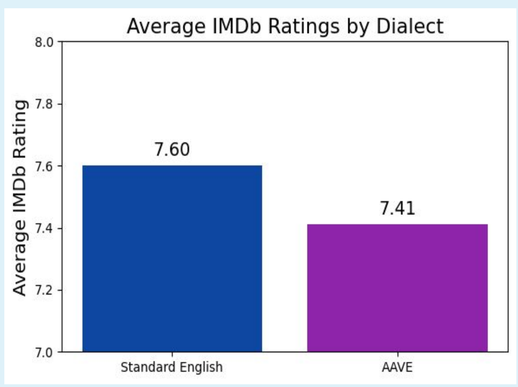
Genre Analysis

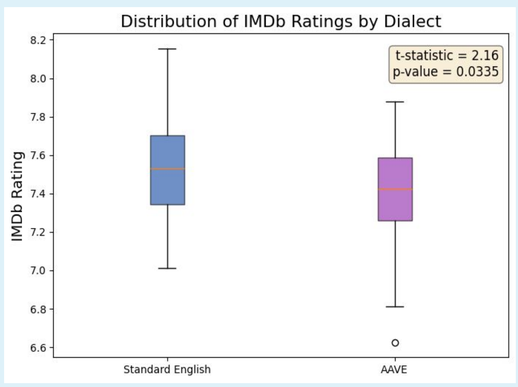
- **Predefined Genre Mapping:** We assign each recommended item to a standard set of genres (e.g., action, drama, comedy) to facilitate direct comparisons.
- **Distribution Comparison:** We measure how frequently each genre appears under different dialect-based prompts, identifying potential biases or disparities.
- **Statistical Metrics:** We apply Kullback-Leibler divergence to quantify distributional differences across dialect conditions, highlighting systematic variations.
- **Interpretation:** These findings help reveal whether certain dialects are disproportionately steered toward specific genres, shedding light on possible cultural or linguistic biases.



Preliminary Results

- **IMDb Rating Comparison:** Preliminary data indicates that movies recommended from Standard English prompts average an IMDb rating of approximately 7.60, while those from AAVE prompts average around 7.41.
- **Statistical Significance:** A two-sample t-test yielded a t-statistic of 3.787 and a p-value of 0.0002, suggesting that the observed differences in ratings are statistically significant.
- **Genre Distribution Insights:** Early analysis hints at shifts in genre frequency between dialects, with certain genres appearing more often in one dialect over the other.
- **Implications for Bias:** These initial findings support the hypothesis that linguistic style impacts both the quality and composition of recommendations, underscoring the need for further investigation into linguistic bias in recommendation systems.





Future Work

- **Expand Dialect Coverage:** Extend analysis to include additional dialects beyond the current focus on AAVE and Standard English.
- **Enhance Multi Domain Analysis:** Incorporate additional domains such as books and music, using corresponding quality metrics (Goodreads, RYM) to provide a more comprehensive evaluation.
- **Dialect Translation:** Hire dialect experts or regional translators to verify dialect translations

Limitations

- **Translation Nuances:** The two-step translation process may not perfectly capture all cultural and linguistic nuances, potentially affecting recommendation outcomes.
- **Dependence on External Metrics:** Quality assessments rely on external rating systems (IMDb, Goodreads, RYM) that may have inherent biases.
- **Sample Size Constraints:** Preliminary data and limited sample sizes may affect the statistical power and generalizability of the findings.