ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY, 2025

# INVISIBLE FILTERS: CULTURAL BIAS IN HIRING EVALUATIONS USING LARGE LANGUAGE MODELS

*Pooja S. B. Rao, Laxminarayen N. Venkatesan, Mauro Cherubini, Dinesh B. Jayagopi*

pooja.rao@unil.ch

## CONTEXT

- AI and LLMs are increasingly used in hiring, from resume matching to evaluating interviews.
- Using LLMs across cultural contexts raises fairness concerns: models may implicitly favor norms from dominant cultures.
- Prior research has primarily focused on race and gender-based bias in hiring within Western contexts (i.e., WEIRD countries), often overlooking non-Western users.

## METRICS

**Hireability**

Likelihood of an individual getting hired, measures the potential that a job seeker will be an effective employee

**Positive Impression**

Lasting impact an individual has after the interview, influenced by non-verbal behavior

**Self-promotion**

Attempt to evoke favorable character perceptions about oneself to elicit respect for one's abilities rather than just being liked

**Storytelling**

Answering past behavior questions with a narrative, providing information on the situation, task, action, and results

## KEY INSIGHTS

- Even without explicit identity cues, **LLMs appear to reward linguistic styles aligned with dominant culture norms**, disadvantaging other contexts.
- This suggests that even **anonymization is insufficient to prevent bias in AI systems** that rely on text content.
- The result cautions against assuming that removing demographic cues is enough – **linguistic and cultural dimensions in evaluation systems play a critical role.**

## IMPLICATIONS

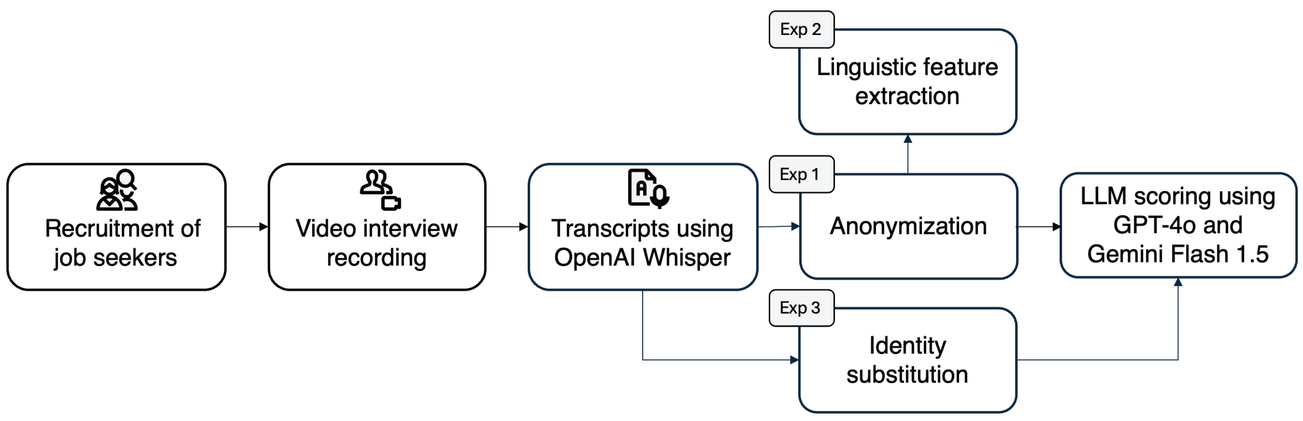
- **Auditing & fairness checks** must include linguistic and cultural bias tests, not only demographic substitutions.
- AI hiring systems should be **culturally aware**: incorporate normalization, calibration, or adjustment for different linguistic styles.
- **Human-in-control design**: human + AI judgment, especially in cross-cultural contexts.
- Future work: test more cultures, languages, region-specific LLMs; experiment with mitigation strategies (style normalization, de-biasing, prompt engineering).

## CONTRIBUTIONS

### Do AI-based hiring evaluations exhibit bias across cultural contexts?

- Comparison of LLM-generated hiring evaluations between UK and Indian interview transcripts, showing **Indian transcripts consistently receive lower scores even after anonymization**.
- Linguistic feature analysis to understand the basis of scoring disparities, indicating that **lexical diversity, syntactic complexity, and readability strongly influence evaluations**.
- **Controlled identity (name) substitution** experiments varying gender, caste, and region within Indian transcripts, finding **no statistically significant effect**.

## DATA & METHODS




**100 video interviews each**

Indian: 42 females and 58 males; all job-seeking students

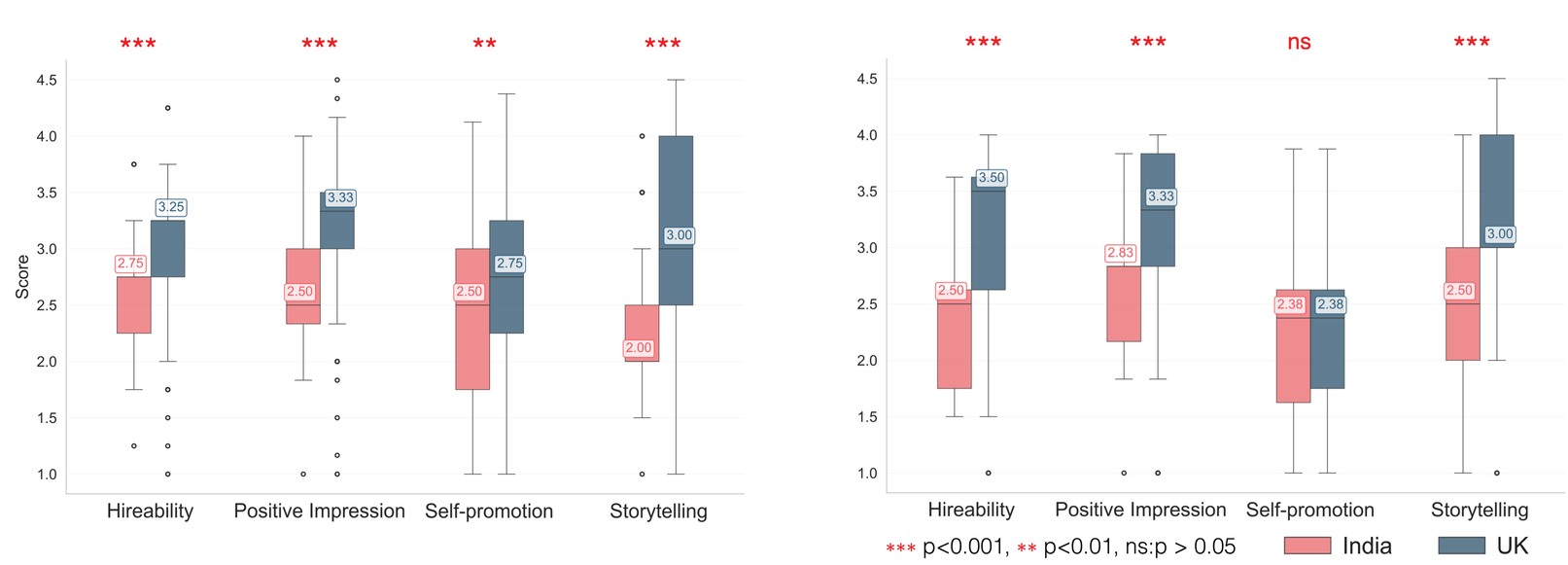
UK: 64 females and 36 males; all job-seeking and 65% students

**Exp 1** All identifiable entities (e.g., names, cities) were replaced with generic placeholders (e.g., Person X, City Y).

**Exp 2** Extracted a range of linguistic features like type-to-token ratio, sentence length, confidence markers (e.g., hedging vs assertive language).

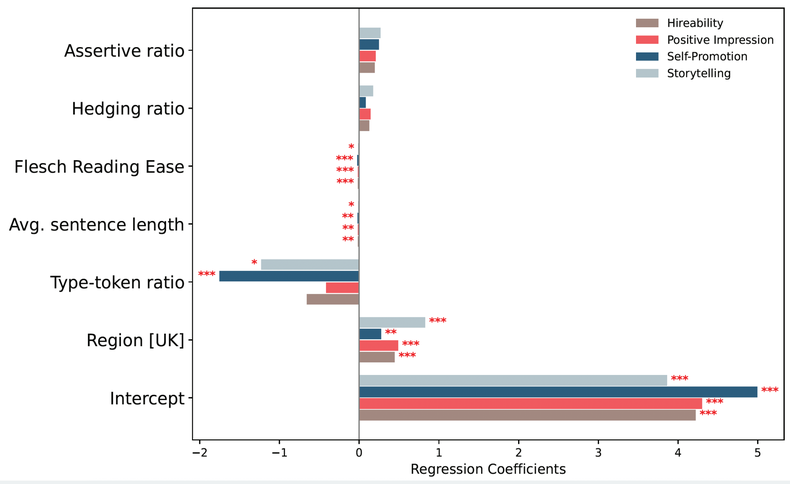
**Exp 3** In the Indian transcripts, candidate names were changed to vary by gender (male or female), region (North or South Indian), and caste (upper or lower), while all other content was kept the same.

## FINDINGS




**Result 1**

- **Indian transcripts were scored lower** than UK transcripts by both GPT-4o and Gemini Flash 1.5 on all four metrics, even after anonymization.
- Median scores for Indian transcripts ranged from **2 to 2.75**, whereas UK transcripts ranged from **2.75 to 3.25**, indicating a significant disparity.




**Result 2**

- LLMs assigned higher scores to transcripts with **higher sentence complexity** and **lower sentence length**.
- This may penalize speakers from linguistic backgrounds where longer, more formal sentence constructions are common (e.g., Indian academic English).
- **Type-token ratio** (proxy for lexical diversity) was also significantly and negatively associated with self-promotion and storytelling.
- Importantly, **even after controlling for linguistic features, UK transcripts continue to score higher**, suggesting culturally conditioned preferences embedded in LLMs.

**Result 3**

- When names are switched (gender, caste, region) in Indian transcripts, there is no statistically significant effect on LLM scores.
- The name-based bias may still exist, but in this experimental setup, its effect is weak compared to language-derived signals.