How Well Do Large Language Models Mimic Human Sentiment? A Comparative Study of AI and Human Survey Responses in the Context of Indian Sports Culture

Abir Bordia

UWCSEA East, Singapore

Abstract

Background/Objective: Large language models (LLMs) have grown rapidly in capability, yet questions remain about whether they can accurately replicate the culturally situated, opinion-driven responses of specific human demographic groups. This study investigates the degree to which six widely deployed LLMs, ChatGPT, Grok, Gemini, Perplexity, Claude, and DeepSeek, are capable of mimicking the sentiment and behavioral patterns of an Indian demographic when responding to an opinion-based sports survey.

Methods: A structured survey comprising eight questions about cricket and football preferences was administered to 105 human participants drawn from an urban Indian population (Udaipur, Rajasthan), with a gender distribution of 65% male and 35% female and an age range of 13 to 76 years. Each of the six LLMs was then prompted to generate 105 responses replicating this demographic profile under identical conditions. All responses were compiled into structured datasets and visualized using bar charts, pie charts, and comparative histograms across five evaluation dimensions: demographic adherence, cultural weighting, temporal recency, recall realism, and linguistic nuance.

Results: Grok achieved the highest aggregate similarity score of 20 out of 25, most closely matching the human baseline across demographic and cultural dimensions. Grok and DeepSeek both reproduced the gender distribution with approximately 65.7% male responses, matching the human 65.4% split. ChatGPT and Gemini each scored 18 out of 25. Perplexity scored 13 out of 25, underperforming notably on cultural weighting by presenting a more globally balanced, rather than India-specific, view of cricket and football popularity. Claude scored lowest at 11 out of 25, primarily because it over-diversified its

responses to open-ended questions, naming 101 distinct footballers, compared to the human tendency to converge on two or three dominant names.

Conclusions: The findings indicate that while modern LLMs can approximate broad demographic trends, they consistently struggle to replicate the culturally specific, recall-limited, and linguistically colloquial character of real human responses. Cultural context-awareness and cognitive constraint simulation represent significant ongoing challenges for LLM development. Future research should extend this framework to other cultural settings and employ larger, more stratified human samples.

Keywords: Large Language Models, Artificial Intelligence, Human Sentiment, Survey Mimicry, Cultural Context, India, Sports Preferences, AI Evaluation