53. POTENTIAL BIASES AND FAIRNESS IN LLM-BASED DATA IMPUTATION FOR SOCIOLOGICAL RESEARCH

Uchkov A.K.

Belarusian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus

Subbotkina I.G. – Associate Professor

This paper explores the risks of bias and fairness issues when using large language models to impute missing data in sociological surveys. Drawing on recent research, it highlights how LLMs can reproduce stereotypes and underrepresent minority groups. The article highlights the importance of fairness-aware approaches to ensure ethical and accurate use of LLMs in social science research.

Missing data is a persistent challenge in sociological surveys, often caused by non-response, incomplete entries, or participant dropout. If not addressed properly, it can reduce the quality of statistical analysis and lead to misleading or biased conclusions. In recent years, large language models (LLMs) such as GPT, Claude, LLaMA have gained attention as potential tools for imputing missing textual or categorical responses. These models can use the context of available answers to predict plausible values, often outperforming traditional statistical methods in terms of coherence and fluency. However, the use of LLMs in sociological research raises important methodological concerns: since LLMs are trained on large-scale internet data, they may reproduce and even amplify social biases present in their learning data. The goal of this article is to explore the risks of bias and fairness violations in the situations when LLMs are used for data imputation, using insights from recently published studies.

Large language models have become powerful tools for generating text and predicting likely language patterns, however, if we talk about sociological studies, it is important to understand that LLMs are trained on vast text corpora scraped from the internet, which often reflect dominant cultural narratives, gender roles, racial or socioeconomic stereotypes. These patterns can be unintentionally reproduced when models are used to fill in missing responses in survey data. For instance, research shows that LLMs tend to favor majority group norms in prediction tasks. In one example, the model was more likely to associate professional success with white male names than with names typically associated with other ethnic or gender groups [1]. When applied to missing values in surveys, this kind of bias could lead to overrepresentation of majority perspectives or normalization of certain lifestyles over others.

Study provide further evidence of such effects, noting that GPT-based models frequently associate women with caregiving roles and men with leadership or technical professions. This stereotype reinforcement becomes particularly problematic when imputing answers related to occupation, family roles, or social values. If survey responses from underrepresented groups are missing and filled in by such a model, the resulting data may reflect cultural assumptions rather than experiences of the real person [2].

In addition, article emphasize the importance of understanding the mechanism behind missing data – whether it is missing at random or not at random. They show that when missingness is patterned by group membership (for example, younger respondents skipping political questions), applying a one-size-fits-all imputation approach without adjusting for this can introduce serious distortions in group comparisons [3]. While LLMs offer flexibility in handling contextual data, they do not inherently model these underlying mechanisms unless specifically guided.

Research shows that even well-intentioned models can result in unequal outcomes. Their analysis demonstrates that imputation strategies without fairness constraints can disproportionately affect marginalized groups, even when the model performs well overall. For example, the same imputation algorithm applied across ethnic groups may yield more accurate results for majority populations, while increasing error rates among minorities—thus deepening existing disparities in data representation [4].

Taken together, these studies suggest that LLMs, if applied without caution, may reproduce stereotypes, underrepresent minority voices, and reinforce dominant cultural norms – all of which can distort the outcomes of sociological research. Such distortions are not just technical flaws; as research shows, they carry ethical and interpretive consequences. Inaccurate imputation can affect findings about inequality, social behavior, or policy preferences, leading to conclusions that reinforce existing disparities instead of trying to challenge them. Therefore, fairness must be considered from the outset when designing LLM-based imputation systems – it is important to apply suitable methodology depending on the context of the research.

References:

1. Chu, Z. Fairness in large language models: A taxonomic survey. / Z. Chu, Z. Wang, W. Zhang. – ACM SIGKDD explorations newsletter. – 2024. Vol. 26, iss. 1.

2. Gallegos, I. O. Bias and Fairness in Large Language Models: A Survey. / I. O. Galleos [et al.]. – Computational Linguistics. *– 2023. Vol. 50, No. 3.*

3. Fairness in Missing Data Imputation [Electronic resource]. – Mode of access: https://arxiv.org/pdf/2110.12002. – Date of access: 10.04.2025.

4. Caton, S. A Review of Missing Data Handling Methods in Education Research. / S. Caton, S. Malisetty, C. Haas. – Journal of Artificial Intelligence Research. – 2022. Vol. 74.