https://www.sworldjournal.com/index.php/swj/article/view/swj33-01-078
DOI: 10.30888/2663-5712.2025-33-01-078

UDC 004.8:519.23

THE UNSEEN DATA: A STATISTICAL AND ENGINEERING PERSPECTIVE ON BIASES IN LARGE LANGUAGE MODELS

Tumanov O.O.
PhD
ORCID: 0000-0003-0674-0037
V.N.Karazina Kharkiv National University, Kharkiv, Maydan Svobody, 4, 61022

Abstract. *The paper argues that bias in large language models (LLMs) is a fundamentally statistical problem rooted in the nature of their training data. The unfiltered datasets used for training are not representative samples of human language, but rather deeply imperfect statistical corpora. This sampling bias, combined with historical inequality and demographic underrepresentation, results in biased parameter estimates during the model training process.*

The paper presents a clear taxonomy of bias, distinguishing between data-driven bias and model-based bias. Quantification using statistical tools such as the chi-square test is suggested. A medical LLM concept study illustrates how data skewed towards western medicine can lead to dangerous consequences. The article highlights the importance of moving from passive data collection to active data engineering, detailing techniques such as stratified sampling and synthetic data generation.

The paper also acknowledges the role of post-processing solutions, such as prompt engineering and fairness-aware algorithms, as a final layer of defense. In conclusion, the paper emphasizes that a more statistically sound approach to AI development is crucial. The paper also addresses the issues of determining fairness and the high computational cost of careful selection and quality control of data, and suggests future research on open, ethically prepared datasets and new statistical methods for quantifying intersectional biases.

Keywords: *LLM, AI Bias, Data Imbalance, Fairness, Machine Learning, Computational Linguistics, Statistical Bias, Ethical AI.*

Introduction

The rapid evolution of artificial intelligence technologies has led to the emergence of a new class of models that have become a new paradigm for data analysis: large-scale language models (LLMs). Trained on broad and diverse datasets, LLMs demonstrate an unprecedented ability to generate, generalize, and understand text that mimics human communication. In the context of the author's previous work on statistical analysis of social media, the emergence of LLMs represents a significant deviation from traditional approaches [10, 11]. These approaches have historically focused on structured multivariate data that has been the focus of prediction and cluster analysis. Today, LLMs are being actively implemented in systems ranging from healthcare, where they facilitate biomedical research, to software development, where

their use benefits productivity. Despite their undeniable achievements, the likely and often opaque nature of LLMs has called into question established statistical principles. Traditional statistical practice emphasizes tools such as p-values, confidence intervals, and regression coefficients, but determining the reliability and ethical validity of multivariate modeling (LLM) methods requires a new evaluative approach [12]. Therefore, this article addresses the problem of LLMs bias as a fundamentally statistical problem.

The aim of this paper is to explore the nature of bias in large language models, in particular its statistical roots. This includes a taxonomy of bias types, proposed methods for quantification, and approaches to correction. The paper also demonstrates how biases can lead to dangerous consequences in applied domains, and identifies future research directions.

1. The data-driven bias problem in large language models. The advent of large language models (LLMs) has revolutionized the field of artificial intelligence, with models like GPT-4 and Llama demonstrating unprecedented capabilities [2]. Their impact is already evident in sectors such as healthcare [7], finance [8], and education [1]. However, a significant problem remains: LLMs can reinforce societal biases. These biases – whether racial, gender, or cultural – are not a flaw, but a feature of the learning process itself. They are deeply embedded in the vast datasets of unfiltered Internet data from which these models are trained. This article argues that while biases can arise at various stages of a model's lifecycle, the root cause is the inherent lack of representativeness and balance within their massive training datasets. Training data, which reflect the content of the Internet, is a repository of human biases and systematic inequality. The statistical properties of these data – their significant heterogeneity and underrepresentation of certain demographics – are at the root of much of the bias we observe. To address this issue, we first establish a taxonomy of bias, then introduce quantitative statistical assessment methods, and finally propose practical, data-driven solutions.

2. Classification bias: taxonomy. Bias in LLM can be classified as a fundamentally statistical problem that is modified by selection bias, where the data set

used for training is not truly representative of the population. This leads to a distorted estimate of the model parameters. In addition, bias can be caused by algorithms that introduce systematic errors in the estimates [5]. In this article, LLMs' bias is divided into two main categories: data-driven bias and model-driven bias.

2.1. Data bias. This is the most common and deepest form of bias in LLM. Data bias comes from the training data itself. It includes:

- **Historical bias:** This bias arises when data reflects historical and social injustice. For example, a model trained on old texts might learn that "doctors" are predominantly male, perpetuating stereotypes. This bias is a direct mirror of real-world inequality.
- **Insufficient bias:** This occurs when certain groups or topics are underrepresented in the dataset. If there is little data about a particular cultural group, the LLM will perform poorly when interacting with users from that background. The model's knowledge is a statistical reflection of the composition of its training data.
- **Displacement of the sample:** Also known as selection bias, this occurs when the data collection method is not random or representative of the target population. If the data set is taken from a narrow subset of the Internet, the model will only be an expert on the opinions and language of a specific narrow community.

2.2. Model-driven bias. Although data bias is the primary cause, biases can also appear during model development and training. They can be observed in specific mechanisms, for example:

- **Algorithmic shift:** This applies to biases introduced by the learning algorithm itself, for example if the optimization algorithm favors dominant data patterns, unintentionally suppressing minority information [4].
- **Interaction bias:** This bias is not explicitly present in the data or the algorithm, but arises from the complex interactions between them. A model may not exhibit a direct bias on a single feature, but may exhibit a strong bias when multiple features are combined. This is often difficult to predict and diagnose.

3. Data set as a statistical sample. From a statistical perspective, the data used to train LLM is a massive, complex, and fundamentally imperfect statistical sample of human language. This perspective is crucial because it allows us to apply rigorous principles of statistical inference to understand the underlying causes of model bias.

The problem of sample bias: The fundamental assumption of statistical inference – that the sample is drawn randomly and independently of the target population – is violated in LLM training data. The data is not a random sample; it is collected based on the availability of certain platforms. This leads to non-random selection, where certain populations are overrepresented, and stratification and skew, where certain themes and perspectives appear much more frequently than others.

Data imbalance and biased parameter estimation: The non-random nature of the training data translates directly into data imbalance. During training, LLM parameters are estimated to minimize a loss function. With unbalanced data, the model optimization process is heavily influenced by the most common patterns. It allocates more learning opportunities to these dominant features, which leads to biased parameter estimates. For example, if a dataset contains a high frequency of stereotypical associations, the model will learn to assign a higher probability to these distorted associations, even if they are not representative of the wider population. This phenomenon is a direct statistical consequence of sampling bias.

4. Quantification of data sets. A qualitative understanding of bias is not enough; a data-driven approach requires quantitative measurement. This section describes basic statistical and computational techniques for this purpose. First, *diversity and inclusion indicators* can be analyzed by examining the frequency of tokens, pronouns, and names to assess the representation of different demographic groups. Using named object recognition (NER), we can tag objects and analyze their distribution by attributes such as gender or nationality. Furthermore, *correlation analysis and statistical tests* can reveal hidden biases. The chi-square ($\chi 2$) test is ideal for determining whether there is a statistically significant relationship between two categorical variables, such as "gender-related pronouns" and "occupation." A significant p-value signals data bias. Similarly, sentiment analysis can be used to perform correlation analysis between

sentiment scores and demographic identifiers. Finally, the AI community has developed

bias benchmarks and audits as standardized tests for known biases. The Winograd Schema Challenge [9] and the Bias in Bios [3] dataset are used to test whether stereotypes affect model coreferences or occupational classifications. By applying these quantitative methods, we can move from subjective claims to objective, data-driven conclusions about the magnitude of bias.

5. Conceptual Case Study: The Biased Medical LLM. Let's look at a conceptual example of an LLM designed for medical consulting. The software development team collects the data set mainly from online sources in the USA and UK. This creates a significant sampling bias, as 95% of the data comes from sources focused on Western medicine, with little or no representation of other health systems such as Traditional Chinese Medicine (TCM) or Ayurveda. This distorted body leads to:

- ***Distorted statistical associations:*** The model learns that medical terms are predominantly associated with Western biomedical concepts. He does not associate "treatment" with practices such as acupuncture or herbal remedies because there is no statistical basis for this in the data.
- ***Performance difference:*** A user from a non-Western background will experience the low performance of the model. A query about "acupuncture for migraine relief" may be met with an unhelpful or dismissive response.
- ***Harmful recommendations*:** LLM may misunderstand culturally specific descriptions of symptoms (such as "imbalance of bodily humors") and provide irrelevant or dangerous advice.

This example demonstrates that bias is not only an ethical issue; it is a statistical and functional failure. The LLM, despite its size, is a statistical misrepresentation of its educational data.

6. Solutions: Mitigation Strategies. Diagnosing data bias points to a clear, two pronged approach for mitigation: focusing on data-driven strategies and implementing model-agnostic post-processing solutions. The ultimate goal is to move from passive data collection to active, statistically driven data engineering.

One such data-driven approach is **stratified sampling** [6]. This basic statistical method involves dividing a data set into homogeneous subgroups (strata) and sampling from each to ensure adequate representation (Figure 1).

In the context of LLM, this means actively balancing the representation of different languages, geographic regions, or ethnic groups to prevent the model from underperforming for certain groups.

When there is a complete lack of data, **data augmentation and synthetic data generation** can be used to create new, realistic data points to fill in "data deserts". Data augmentation involves making minor changes to existing data to increase diversity, while synthetic data generation creates new data from scratch based on statistical models of underrepresented groups. However, these methods must be carefully controlled to avoid inadvertently introducing new biases.

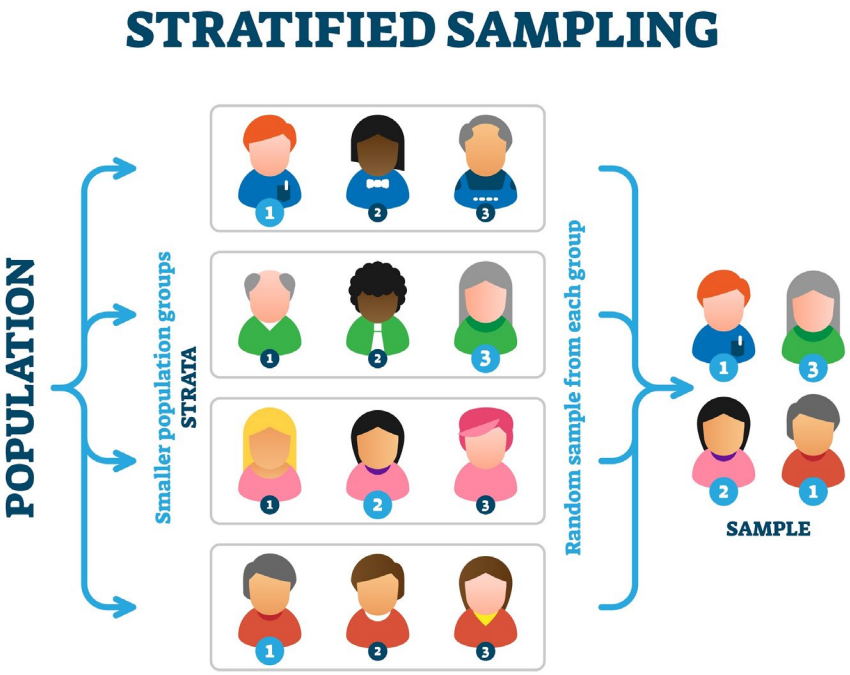

Figure 1 - The scheme of the Stratified sampling

source: Authoring

The ultimate solution lies in a fundamental shift in philosophy from "mindless collection" to intentional and ethical data curation. This approach, although more resource-intensive, ensures that the final LLM is built on reliable, representative data,

reducing the need for significant post-training correction.

Even with careful data preprocessing, some biases may persist. A comprehensive strategy should include post-processing and model-agnostic methods that drive the model's behavior.

- **Operational equipment:** This practical, software-centric solution involves developing cues to distract the LLM from its learned biases. Cues can provide contextual framing or explicit constraints to guide the model toward fair and balanced responses.
- **Fairness post-processing:** This involves applying algorithms to the model's output to correct for residual bias. Eliminating layers can change the ranking of results to penalize stereotypical associations; the output census can automatically replace biased language with more neutral alternatives.

Although these techniques treat symptoms rather than the root cause, they are an important part of a multi-layered defense strategy, especially in real-world applications where user prompts are unpredictable.

7. Conclusions and future directions

The paper provides a comprehensive analysis of the problem of bias in large language models, establishing it as a fundamentally statistical problem arising from biased data selection for training. We describe in detail how unfiltered training corpora that reflect societal stereotypes and inequality lead to distorted estimates of model parameters.

We propose a taxonomy of bias, distinguishing between data-driven bias and model-induced bias, and argue for its quantification using statistical tools such as the chi-square test. We also provide examples of the potentially dangerous consequences of bias in applied domains, particularly in medical LLM. Therefore, a holistic approach to addressing bias must start with the data itself, which will ensure more reliable and ethical results.

Limitations and challenges: Several important issues remain. “Fairness” is not a universally accepted measure; different statistical definitions may conflict. Furthermore, the computational cost of training on large, diverse datasets is enormous.

Future research: We suggest several avenues for future research: the creation of standardized, open-source datasets that are ethically managed; new statistical methods to quantify intersectional biases (e.g., racial and gender bias); and the development of a normative framework for the development of LLMs.

References

1. Alqahtani, T., Badreldin, H. A., Alrashed, M., Alshaya, A. I., Alghamdi, S. S., bin Saleh, K., Alowais, S. A., Alshaya, O. A., Rahman, I., Al Yami, M. S. and Albekairy, A. M. (2023) 'The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research', *Research in Social and Administrative Pharmacy*, 19(8), pp. 1236–1242. doi: https://doi.org/10.1016/j.sapharm.2023.05.016.

2. Chiarello, F., Giordano, V., Spada, I., Barandoni, S. and Fantoni, G. (2024) 'Future applications of generative large language models: A data-driven case study on ChatGPT', *Technovation*, 133, p. 103002. doi: https://doi.org/10.1016/j.technovation.2024.103002.

3. De-Arteaga, M. et al. (2019) 'Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting', in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM. doi: https://doi.org/10.1145/3287560.3287572.

4. Gallegos, I. O. et al. (2024) 'Bias and Fairness in Large Language Models: A Survey', *Computational Linguistics*, 50(3), pp. 1097–1179. doi: https://doi.org/10.1162/coli_a_00524.

5. Guo, Y. et al. (2024) 'Bias in Large Language Models: Origin, Evaluation, and Mitigation', *arXiv*. doi: https://doi.org/10.48550/arXiv.2411.10915.

6. Makwana, D., Engineer, P., Dabhi, A. and Chudasama, H. (2023) 'Sampling Methods in Research: A Review', 7, pp. 762-768.

7. Mesko, B. (2023) 'The ChatGPT (Generative Artificial Intelligence) Revolution Has Made Artificial Intelligence Approachable for Medical Professionals', *J Med Internet Res*, 25, p. e48392. doi: https://doi.org/10.2196/48392.

8. Noguer I Alonso, M. (2024) 'Large Language Models in Finance: Reasoning', *SSRN*. doi: http://dx.doi.org/10.2139/ssrn.5048316.

9. Sakaguchi, K., Le Bras, R., Bhagavatula, C. and Choi, Y. (2020) 'WinoGrande: An Adversarial Winograd Schema Challenge at Scale', in *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05), pp. 8732–8740. doi: https://doi.org/10.1609/aaai.v34i05.6399.

10. Tumanov, O. O. (2019) 'Aspects of Using Social Media in Research', *Scientific Bulletin of the National Academy of Statistics, Accounting and Audit*, (4), pp. 24–29. doi: https://doi.org/10.31767/nasoa.4.2019.03.

11. Tumanov, O.O. (2019) 'Social media as an object of statistical research', *Business Inform*, (12), pp. 8–14. DOI: https://doi.org/10.32983/2222-4459-2019-12-8-14.

12. Tumanov, O. O. (2020) 'Statistical methods for analyzing social media data', *Business Inform*, 2, pp. 266–272. DOI: https://doi.org/10.32983/2222-4459-2020-2-266-272 .

Article sent: 19.09.2025