Large Language Models for Statistical Inference: Context Augmentation with Applications to the Two-Sample Problem and Regression

Marc Ratkovic

(University of Mannheim)

04/02/2026, 4:15 pm

Department of Statistics, Ludwigstr. 33, Room 144
and online via Zoom (Link)
(Meeting-ID: 631 1190 7291; Password: StatsCol)

We introduce context augmentation, a data-augmentation approach that uses large language models (LLMs) to generate contexts around observed strings as a means of facilitating valid frequentist inference. These generated contexts serve to reintroduce uncertainty, incorporate auxiliary information, and facilitate interpretability. For example, in the two-sample test, we compare the log-probability of strings under contexts from its own versus the other group. We show on synthetic data that the method's t-statistics exhibit the expected null behaviour while maintaining power and, through a replication, that the method is powerful and interpretable. We next introduce text-on-text regression. Contexts generated around the predictor string are treated as mediating variables between the predictor and outcome strings. Using negative controls, we then distinguish between semantic and syntactic dimensions of prediction. Analysis of real-world dialogic data illustrates behaviour predicted from a psycholinguistic framework. Theoretically, we provide identification conditions, derive an influence-function decomposition, and show that repeated cross-fitting of a pivotal statistic yields higher-order efficiency. We derive bounds linking estimation error, context count, and number of cross-fits. Taken together, context augmentation offers the ability to connect LLMs with longstanding statistical practice.

About the Speaker:
Marc Ratkovic is Professor and Chair of Social Data Science at the Department of Political Science, University of Mannheim. He is working to integrate machine learning and deep learning methods into practical social science methodology. Prior to joining Mannheim, he received his PhD from the University of Wisconsin-Madison under David Weimer, was mentored by Kosuke Imai at Princeton University as a post-doctoral fellow, and spent a decade teaching in the Department of Politics with an affiliation at the Center for Statistics and Machine Learning at Princeton University.

References:

Ratkovic, M. (2025). Large Language Models for Statistical Inference: Context Augmentation with Applications to the Two-Sample Problem and Regression. *arXiv preprint arXiv:2506.23862* [https://arxiv.org/abs/2506.23862]