Instructable LLMs for scaling data-driven language and culture research

Andres Karjus

ERA Chair for Cultural Data Analytics, Tallinn University
Estonian Business School

The increasing capacities of instructable large language models (LLMs) presents an unprecedented opportunity to scale up data analytics in sciences dealing with language and text, and to automate qualitative tasks previously typically allocated to human labor. Of particular interest to the humanities and social sciences is the capacity to use them as zero-shot classifiers and inference engines. While classifying texts or images for various properties has been long available in the form of supervised learning, the necessity to train (or tune pretrained) models on sufficiently large sets of labeled examples has arguably hampered their adoption for research beyond generic tasks. Approaches like word or sentence embeddings and topic modeling allow for explorative approaches but are typically laborious to interpret and difficult to use for confirmatory inference. This presentation reports on a set of experiments applying LLMs in zero-shot classification and reasoning scenarios, derived from or replicating existing research. These cover various practical, exploratory and confirmatory tasks, including linguistic feature analysis, stance and text reuse detection, literary genre inference, historical news topics prediction, event cause inference, social network inference from texts, novel word sense inference, and semantic change and divergence quantification. Given the prevalence of English focus in the emerging LMM literature, most of these examples deal with harder scenarios, including data in languages other than English, or historical texts prone to OCR errors. It is shown that in all but the most difficult tasks requiring expert knowledge, LMMs can serve as a viable analytic alternative to human annotations. LLM outputs naturally contain errors (as does human annotation), but the error rate can and should be included in subsequent statistical modeling; a bootstrapping approach is discussed. Finally, a quantitizing mixed methods research design, borrowed from corpus linguistics, is argued as an optimal framework for leveraging this scalability while fostering replicability and transparency.