

Information extraction from data with statistically enhanced LLMs

Abstract

Belgium and Europe produce vast administrative and survey datasets covering income, taxation, employment, housing, consumption, household finances, etc. Although accessible to researchers, these sources remain difficult to exploit: variables are hidden behind opaque codes, documented in codebooks running to several hundred pages, which hampers both their identification and the analysis of their interrelations. We present an approach combining large language models (LLMs) and statistical methods to make these data both accessible and reliable: LLMs open up access through natural language but lack rigour, whereas statistics are rigorous yet opaque to a non-specialist audience. The solution is built around three tools. The first identifies, from a question phrased in plain language, the relevant variables among more than 7,000 spread across six Belgian datasets, by means of semantic search (multilingual embeddings, retrieval-augmented generation, validation by a locally run LLM). The second detects and ranks associations that are both strong and unexpected, by combining statistical measures of significance and strength with a "surprise" score assigned by an LLM. The third determines which associations persist once potential confounders are accounted for, thereby distinguishing robust links from spurious correlations. Together, these tools form a pipeline that turns opaque microdata into reliable, shareable data stories, in the service of a shared reality grounded in common data.

Keywords: (Variational) Association Eplorer; Generative Adversarial Networks; Missing Data; Sampling Weights; Wasserstein Distance.