Proposal for the Joint Quantitative Political Science Conference for Asia and Australasia, 9-10 January 2025

Xavier Marquez, Victoria University of Wellington

Using Large Language Models to generate Political Regime Datasets

Large Language Models (LLMs) are trained on very large text corpora, including information relevant to the classification of political regimes. In principle this information could be elicited through appropriate prompting to generate quantitative political regime information – including country-year democracy indexes and indicators of military or personalistic rule – at relatively low cost, enabling researchers to create bespoke indexes of democracy or new regime typologies, or to extend existing ones. I report the here the results of using large language models (LLMs) to generate political regime data, using a variety of prompting strategies and models, ranging from the most powerful closed models (OpenAI’s GPT4o, Anthropic’s Claude Sonnet 3.5, and Google’s Gemini Pro) to smaller open weights models (e.g., the Llama series of models by Meta). I evaluate the generated data against existing datasets, and check the possibility of using LLMs to cheaply extend already-existing, well used, but not regularly updated datasets of political regime information (for example, Geddes, Wright, and Frantz’s “Autocratic breakdown and regime transitions: A new data set”, from 2014). Based on a smaller-scale experiment I performed in 2023 (Marquez 2023, <https://xmarquez.github.io/GPTDemocracyIndex/GPTDemocracyIndex.html>) I expect that the quality of LLM regime-generated political regime datasets is improving, and suggest various methods to increase their quality.