

International Journal of Population Data Science

Journal Website: www.ijpds.org



Content-Based Classification of URL Domains By Large Language Models

Marta Cantijoch¹, Conor Gaughan¹, Rachel Gibson¹, Riza Batista-Navarro¹, and Alex Cernat¹

¹University of Manchester, UK

Introduction & Background

Web tracking provides a valuable source of information across a variety of disciplines, including political science, health, psychology, marketing, economics, and communication. Usually, the digital footprint returned from web trackers includes a series of URLs accompanied by metadata like timestamps or time spent. These data are typically large, unstructured, and can easily become technically challenging to analyse. Traditionally, content-based URL classification required resource-intensive web-scraping and advanced technical expertise. The emergence of Large Language Models (LLMs) offers a potential solution for automating the classification of URL domains quickly and flexibly while reducing technical barriers.

Objectives & Approach

We audited the performance of nine state-of-the-art LLMs from OpenAI, Google, and Anthropic in classifying URL domains. The data we used for the analyses included all URLs visited on mobile devices by a sample of 599 participants based in the UK in 2020. The LLMs were tasked with categorising a total of 4,516 unique domains into 17 pre-defined categories, with an additional “unclassified” category. We used a zero-shot approach, i.e. we did not provide any definitions or examples to the LLMs about these categories. Performance was compared against a test set of 200 random domains annotated by a human.

Relevance to Digital Footprints

This research directly addresses the challenges of processing internet history records, a key form of digital footprint data. By demonstrating how LLMs can process and categorise large-scale web-tracking data, the study explores new methodologies for enriching linked datasets (e.g. web-tracking linked to survey data), thereby reducing the barriers for social scientists to derive meaningful insights from individual-level online behaviours.

Results

Our analyses show that LLMs can perform comparatively well at classifying URL domains into predefined categories. Results suggest a high level of agreement with human annotation (Mean accuracy score: 71%; Mean precision and recall: 74%). These are levels of inter-coder reliability considered more than satisfactory in social science research when using conventional methods (i.e. multiple human coders). We also found that LLM performance is positively correlated with domain popularity; more frequently visited websites are classified with significantly higher accuracy.

Conclusions & Implications

LLMs can drastically improve the speed and efficiency of URL domain classification at scale. While zero-shot performance is already impressive, further prompting engineering may enhance accuracy. Notwithstanding important privacy, transparency and environmental concerns, LLMs can become an equalising tool, making advanced web-tracking analysis more accessible and affordable for researchers.

Keywords

large language models, web tracking data, URL classification

