Content-Based Classification of URL Domains By Large Language Models

Conor Gaughan, Marta Cantijoch, Riza Batista-Navarro, Rachel Gibson, Alex Cernat

Digisurvor team, University of Manchester

Introduction

- Web-tracking data provide valuable information across a variety of disciplines (political science, health, psychology, marketing, communication), in particular when linked to surveys.
- Web-tracking data typically includes a series of URLs accompanied by metadata like timestamps and time spent: datasets are often large and unstructured.
- Classification of URLs becomes technically challenging.

→ **Can Large Language Models (LLMs) help reduce technical barriers for URL classification?**

In this paper:

- We aim to develop a method for the processing, filtration, and classification of URL domains leveraging LLMs.

- We audited nine state-of-the-art LLMs (OpenAI, Google, Anthropic) on their ability to classify URL domains.

Past research and gaps

- URL data filtered using pre-defined domains lists (Cardenal, Victoria-Mas, Majó-Vázquez, Lacasa-Mas, 2022)
  → A list is not always available. Analyses may exclude relevant domains that exist in the data but are not on the list.

- Automated methods used to identify malicious URLs to prevent phishing and malware (Kumi, Lim, & Lee, 2021)
  → While important, limited utility.

- Analyses of the text within the link itself for topical/content-based classification (Kan and Thi, 2005)
  → Information contained on the text of the link is very limited. Risk of lack of accuracy.

- Accessing web content directly (web-scraping) or using search engines/Wikipedia APIs to classify URLs automatically and dynamically
  → Time-consuming, technically challenging, expensive.

Emerging LLMs offer a quick, affordable, and flexible alternative
(without requiring human-generated training data).

Data

- Dataset from a Charlemagne Prize Academy Fellowship focusing on information consumption during the COVID pandemic

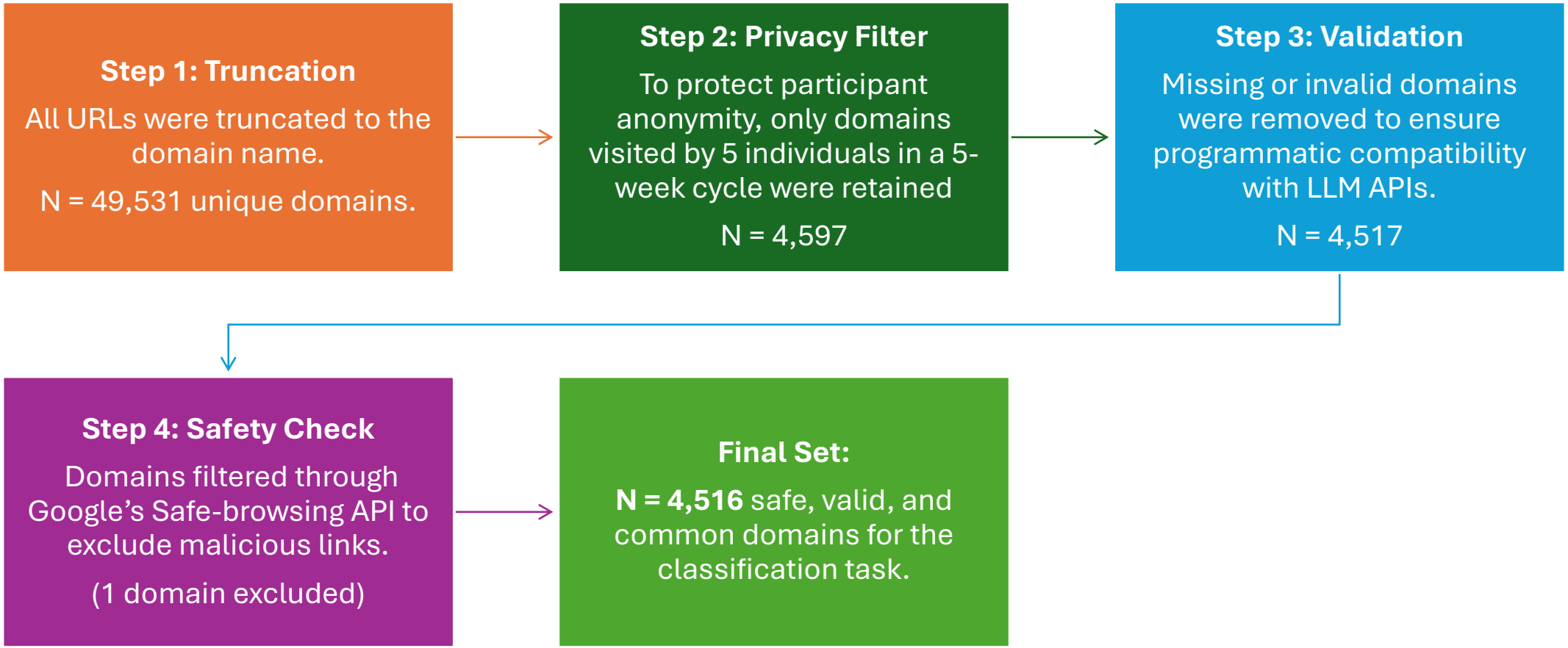
  With thanks to Maria Victoria-Mas (PI), Silvia Majó-Vázquez (CO-I) and team for sharing the data!

- Data collected by YouGov via a tracking device on participants' browsers (mobile devices only) between March 20 and May 21, 2020.

- Sample: N = 599 participants (representative of the UK online adult population).

- 3,016,282 total URL visits overall, averaging 5,036 links per person.

Pre-processing and filtering



Pre-processing and filtering

Step	Stage	N	N Per Participant	N Per Day	N Per Participant Per Day
1	Original URL Links	3,016,282	5,036	49,447	83
2	Unique Domains	49,531	83	812	1.4
3	Unique Domains (>=5)	4,597	8	75	0.13
4	Valid Domains	4,517	8	74	0.12
5	Safe Domains	4,516	8	74	0.12

Table 1: URL data after each stage of the filtration process.

Model selection

LLM provider	Flagship models	Characteristics
Open AI	GPT-5	Highest performance
	GPT-5 mini	Faster and cost-efficient
	GPT-5 nano	Even faster and more cost-efficient
Google	Gemini 2.5 Flash	Highest performance
	Gemini 2.5 Pro	Faster and cost-efficient
	Gemini 3 Pro (preview version)	Highest performing potential?
Anthropic	Claude 4.5 Opus	Highest performance
	Haiku	Fastest
	Sonnet	Balanced: speed/performance

(We set the temperature of the Google Gemini and Anthropic Claude models to 0 to restrict each model to more focused and deterministic outputs. GPT-5 models no longer support a temperature parameter)

Topic Categories

Categories		
News and Media	Technology and Software	Nonprofit and Advocacy
Health and Wellness	Banking, Business and Finance	Travel and Tourism
Retail and E-Commerce	Social Networks and Community	Gambling and Betting
Adult and Dating	Education and E-Learning	Search Engines
Government and Public Sector	Science and Nature	Other
Entertainment and Leisure	Food and Drink	-1 (Unclassified)

Table 2: Predefined list of possible domain categories.

Model Prompt

System Prompt

You are an assistant to determine the classification of a website domain.

Task Instructions

Classify the following website domains: {domain_list}, into one of the following categories: {category_list}. Where the assistant has no knowledge of the website domain, the assistant returns -1; otherwise, the assistant should provide the best estimation. Please return each domain alongside its assigned category all in a single JSON string only.

Results - Model Agreement

- To assess model performance, we create a "gold standard" test set by randomly sampling 200 domains from the 4,516 domains prior to model classification.
- These 200 domains were then classified into one of the 18 predefined categories by a human annotator
- For validation, these 200 domains will also be annotated by a second coder but this has not be done yet.
- Mean agreement between all 36 model pairs is high (Cramér's *V* = **0.78**)
- Anthropic and Google = (Cramér's *V* = **0.81**)
- OpenAI = (Cramér's *V* = **0.77)**

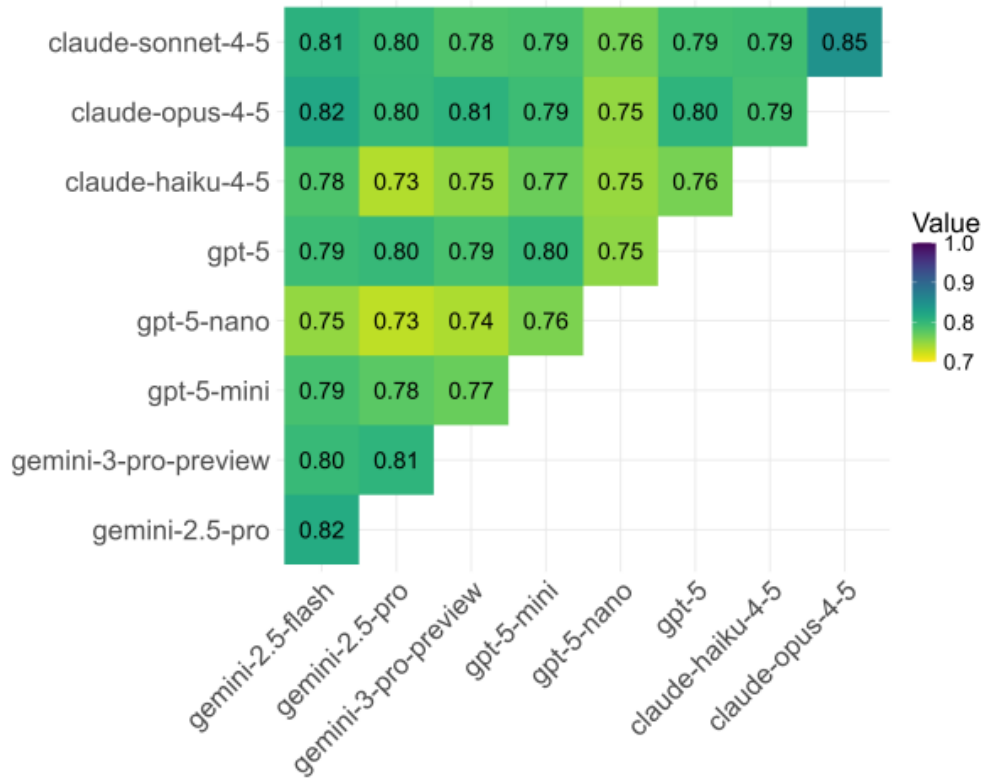


Figure 1: Cramér's *V* Correlation Matrix of Model Agreement. *N* = 3,896.

Results - Model Accuracy

- Model performance assessed against our random sample of 200 domains coded by a human also suggests a relatively high level of agreement with human annotation
- Mean overall accuracy across all nine models is **71%** against a chance baseline of only **5.56%.**
- gemini-3-pro-preview model demonstrates a slightly lower degree of accuracy at **67%**
- Mean macro precision across all nine models is higher at 77% but mean macro recall is lower at 68% indicating that the models are moderately conservative in their classification of domains.
- Mean macro F1 scores across all nine models is **74%.**

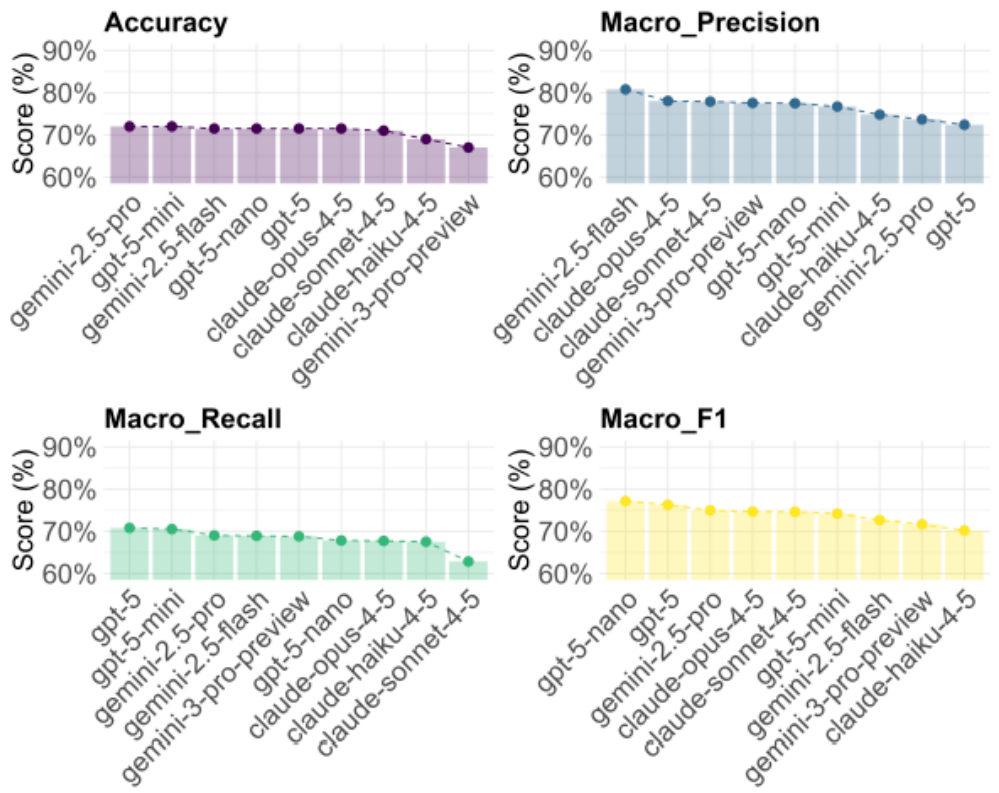


Figure 2: Model Classification Metrics Against Human Annotation. $N = 200$

Results – Model Accuracy By Domain Popularity

- We assume that model performance is likely to be contingent on the relative popularity of each domain website.
- We compute a domain's popularity by the log of the number of times it appears across each participant in each 5-week cycle.
- We then fit local regressions against two dependent variables:

1) whether a domain was given a classification [0: No; 1: Yes];
2) whether a domain classification matches the human annotation [0: No; 1: Yes]

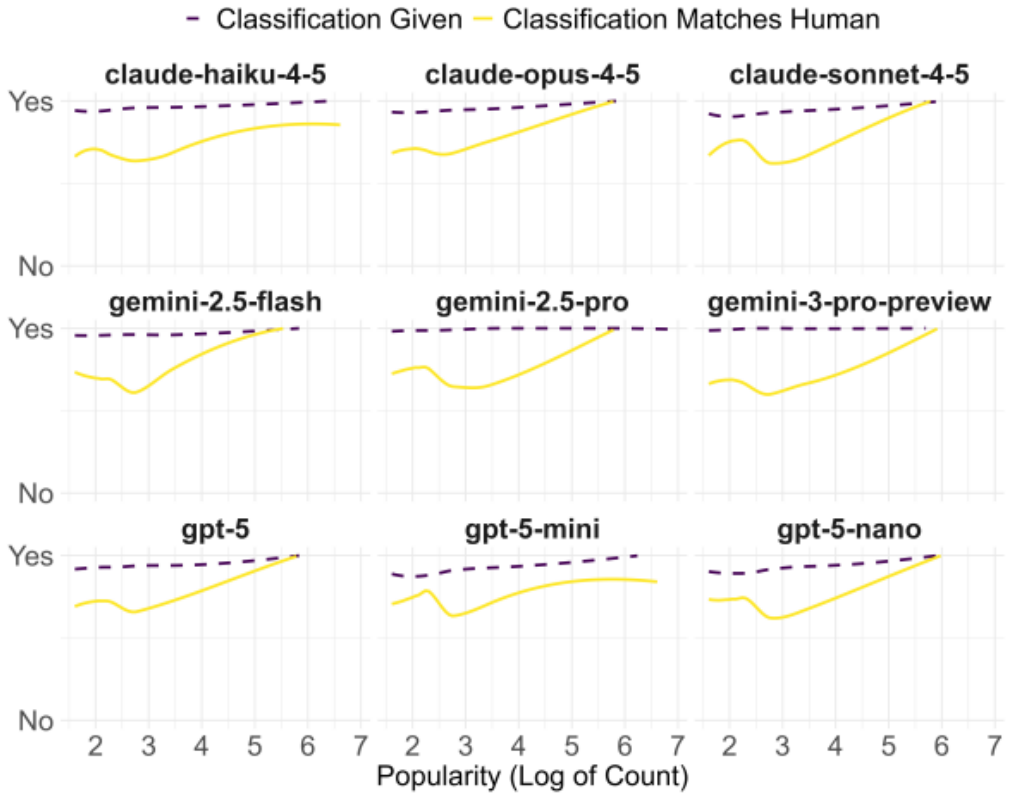


Figure 3: Proportion of domains classified by each model (N = 3,896) and model accuracy (N = 200) by popularity of domain. Smoothed regression lines are fit using LOESS.

Figure 4: Wordcloud of domains sized by count and coloured by classification. Based on classifications by gemini-2.5-pro. N = 4,516. Maximum word count for plotting is 1,000.

Conclusions & Future work

- LLMs demonstrated impressive "out-of-the-box" capability, classifying over two-thirds of domains (71% mean accuracy) using only prior knowledge. Models perform best on most popular sites.

- **Overall: LLMs can drastically improve the speed and flexibility of large-scale URL classification for social science research.**

- Room for improvement:
  - Using few-shot prompting (as opposed to zero-shot) to provide context.
  - Reducing number of categories (if useful for the research project).
  - Removing single-category restrictions which may have suppressed higher accuracy scores.

- Future work:
  - Re-assess including rarer domains: does performance drop significantly?
  - Expand beyond UK-centric data to test cultural and linguistic generalisability.