Original Manuscript

Social Science Computer Review
2025, Vol. 0(0) 1–14
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/08944393251344865
journals.sagepub.com/home/ssc
Sage

Prompting the Machine: Introducing an LLM Data Extraction Method for Social Scientists

Laurence-Olivier M. Foisy[1], Étienne Proulx[1], Hubert Cadieux[1], Jérémy Gilbert[1], Jozef Rivest[2], Alexandre Bouillon[1], and Yannick Dufresne[1]

Abstract

This research note addresses a methodological gap in the study of large language models (LLMs) in social sciences: the absence of standardized data extraction procedures. While existing research has examined biases and the reliability of LLM-generated content, the establishment of transparent extraction protocols necessarily precedes substantive analysis. The paper introduces a replicable procedural framework for extracting structured political data from LLMs via API, designed to enhance transparency, accessibility, and reproducibility. Canadian federal and Quebec provincial politicians serve as an illustrative case to demonstrate the extraction methodology, encompassing prompt engineering, output processing, and error handling mechanisms. The procedure facilitates systematic data collection across multiple LLM versions, enabling inter-model comparisons while addressing extraction challenges such as response variability and malformed outputs. The contribution is primarily methodological—providing researchers with a foundational extraction protocol adaptable to diverse research contexts. This standardized approach constitutes an essential preliminary step for subsequent evaluation of LLM-generated content, establishing procedural clarity in this methodologically developing research domain.

Keywords

large language models, ChatGPT, methodology, data extraction, political science, social sciences, artificial intelligence

[1]Université Laval, Canada
[2]Université de Montréal, Canada

Corresponding Author:
Laurence-Olivier M. Foisy, Université Laval, 2325 Rue de l'Université, Québec, QC G1V 0A6, Canada.
Email: mail@mfoisy.com

Introducing a New Type of Dataset Collection Method

This research note explores an emerging field: the use of Large Language Models (LLMs) as data sources in the social sciences. The emergence of artificial intelligence and LLMs raises numerous questions, especially regarding the validity and accuracy of their information. Hartmann et al. (2023) identify three potential bias sources: majority biases from unbalanced training datasets, biases from human involvement in training procedures, and biases from safeguards preventing harmful content generation. Inherently, the training process embeds biases, creating a “black box” scenario that conceals their exact nature. Recent studies demonstrate that LLMs display biases against women and minority groups (Guo & Caliskan, 2021; Zack et al., 2023; Zhao et al., 2024). Others have added that ChatGPT often mirrors dominant American ideologies, reflecting a broader liberal and Western perspective (Johnson et al., 2022; Rozado, 2023; Rutinowski et al., 2023; Van den Broek, 2023), and this bias extends to LLMs’ evaluations of personalities, institutions, and political parties, favoring liberals and left-leaning entities (McGee, 2023; Motoki et al., 2023; M Foisy et al., 2024; Heyde et al., 2024).

These findings open a broad spectrum of research questions, such as the accuracy of political information generated by LLMs. However, before investigating these questions, a fundamental methodological challenge was encountered: generating the data to study. This research note seeks to initiate the discussion of this pivotal methodological issue and seeks to suggest a standardized practice to extract data from LLMs using their API.

Recent literature highlights the emerging methodological considerations surrounding LLM-generated datasets in the social sciences. As social scientists begin exploring LLMs as data sources, the focus has predominantly been on the outputs rather than the standardized procedures needed to extract this data. Pangakis & Wolken (2024) emphasize the need for human-centered approaches when using generative AI for research purposes, while Törnberg (2024) outlines best practices for structured data collection from language models. These methodological discussions have typically centered on evaluating biases and the reliability of the generated content, with Ziems et al. (2024) questioning whether LLMs can transform computational social science through their data generation capabilities. However, before researchers can effectively evaluate and utilize LLM-generated datasets, a fundamental prerequisite exists: establishing transparent and replicable procedures for how these datasets are created in the first place. Our research note specifically addresses this procedural gap, focusing not on validating LLM outputs, but rather on documenting a clear methodological approach for extracting data from these systems—an essential first step that must precede any substantive analysis of the content itself.

Data from LLMs is of significant interest in political science research, given its novelty and swift integration. However, as OpenAI’s GPT models are proprietary, their information sources and defining parameters remain largely undisclosed, creating uncertainties about the origin of their training data and their exact impact on outputs. This opacity hinders efforts to understand their mechanisms, limiting users’ ability to assess response reliability. Furthermore, LLMs’ outputs are prone to a certain degree of variability due to the inherent randomness of the models. This poses certain challenges for research reproducibility as it directly affects the validity and reliability of methods relying on these models, making it essential to acknowledge this limitation upfront. Given the widespread public use of LLMs, especially ChatGPT, for critical information like politics, conducting in-depth research to comprehend their impacts and identify potential biases is crucial (Fletcher & Nielsen, 2024). Thus, exploring new data collection methods is vital for advancing research in this domain.

By proposing a structured way to extract output from LLMs, this study seeks to provide a clear and transparent framework, helping replicability and accessibility to the study of this topic (Agnew et al., 2024). In other words, the prompt strategy and design used to access these data must be clear and transparent so that even those who are not familiar with this procedure can potentially assess and evaluate how data was generated. Furthermore, this will also help those who seek to understand the capacities and limits of LLMs and are looking for an easy and accessible procedure to generate data. Finally, this method also offers a means to compare outputs across different LLMs while assessing intra-model variability (the variations in output generated by the same model).

Biases within these tools pose a significant concern as they could have perverse effects on democracy, which depends on free and diverse information sources for optimal functioning (Dahl, 2006). This concern grows with the rapid expansion of accessible AI systems and our increasing reliance on them (Rozado, 2023). However, this research note primarily aims to establish a methodology for extracting data from closed, opaque systems, making it an open-source resource for the scientific community to facilitate further research, rather than definitively ascertaining bias presence. The objective is to provide a standardized approach applicable across social science disciplines for extracting data to test and assess LLM biases.

To showcase this, this research note presents the generation of a novel dataset, created through prompts to OpenAI's GPT models. To do so, this research note uses a hypothetical research question that seeks to evaluate the biases GPT models have toward Canada and Quebec MPs. The rest of the note focuses on the data generation procedure and strategy. In this sense, the following content does not try to assess these biases. This question is only used as an example to facilitate the explanation about how to apply this procedure, and which outcome it provides.

Generating a Dataset

Prompting ChatGPT

A complete list of the 338 + 2 Members of Parliament (MPs) in the 43rd Canadian House of Commons and the 125 Members of the National Assembly (MNAs) in the 42nd National Assembly of Quebec was compiled from public sources.[1] The roster details MPs/MNAs and their electoral districts.

It is important to emphasize that while this research note uses Canadian federal MPs and Quebec provincial MNAs as its case study, the choice of this particular political context serves purely as a demonstration vehicle. Our primary contribution is methodological rather than substantive— establishing a standardized, replicable procedure for extracting political data from LLMs that can be applied to any political context globally. The Canadian case serves as a practical example to demonstrate the data extraction procedure, but the methodology presented here is deliberately designed to be transferable to any political context globally. The primary goal is to establish a standardized method for extracting LLM-generated political data that researchers can apply to their contexts of interest, whether studying political systems in Europe, Asia, Africa, or elsewhere. This approach allows for the broad applicability of the method while acknowledging that the specific findings about Canadian politicians are secondary to the methodological contribution. Researchers interested in studying LLM biases in other political contexts can readily adapt this procedure to their specific needs, substituting different political actors while maintaining the core extraction framework outlined in this note.

To simplify data generation from GPT models, an R script tailored for political scientists' use and customization was created and made available publicly on GitHub.[2] Researchers can easily tailor it to their specific research requirements. The script uses the *openai* package, an R wrapper for OpenAI's API, to initiate prompts (see Rudnytskyi, 2023).

The script seamlessly interacts with multiple GPT models, such as gpt-3.5-turbo, gpt-4, and gpt-4–0125-preview. However, its versatility enables researchers to customize the set of models used, adapting the tool to their study's specific needs[3] and ensuring the script's relevancy when newer models are released.

The option to adjust the temperature, a hyperparameter governing LLMs' output randomness, was omitted to ensure the data generated closely mirrors the model's default responses, akin to what a user would receive from ChatGPT (Ouyang et al., 2023). Nevertheless, researchers can readily modify the script to incorporate this feature, enabling temperature adjustments according to their research requirements.

A major benefit of this script is its capacity to craft standardized prompts that can be customized to various research objectives. This procedural flexibility allows researchers to extract data across diverse domains such as policy evaluation, sentiment analysis, or topic classification, while maintaining methodological consistency. The prompt_subject parameter provides a critical control mechanism for defining the thematic focus of LLM responses while preserving the core extraction procedure. This functionality is essential for formulating questions accurately and deriving detailed insights relevant to political discourses.

To address the methodological challenge of variability in LLM-generated content, our procedure incorporates a predefined iterative prompting protocol with configurable iteration parameters. This systematic approach serves two procedural purposes: first, it enables researchers to collect multiple responses to identical prompts, facilitating assessment of output consistency; second, it implements a structured retry mechanism to handle non-responsive or incorrectly formatted outputs. This procedural robustness enhances data collection reliability—a critical consideration when extracting large datasets from LLMs. Incorporating controlled retry logic and allowing for the specification of content iteration quantities ensures optimized data collection and limits the number of missing data. Process continuity is deliberately prioritized over immediate data quality, since continuous processing allows researchers to obtain a complete dataset with minimal manual intervention, which is crucial when working with hundreds of prompts.

Additionally, the script dynamically adds new columns to the dataset based on researchers' needs and specifications. This feature is especially useful for scholars seeking to collect diverse information from the GPT model, simplifying the integration of new data categories into the dataset.

The script uses the create_chat_completion() function to input entire dialogues between the user and the model as a prompt. This feature is crucial for setting the conversation's context and defining the expected output. Studies show that providing LLMs with examples of the expected output yields more accurate results by effectively guiding the model's responses (Brown et al., 2020; Zamfirescu-Pereira et al., 2023).

Crafting the Prompt

In the context of social science and LLMs, a prompt is a natural language instruction given to an LLM to elicit specific behaviors and outputs from the model (Brown et al., 2020; Meskó, 2023; Schulhoff et al., 2024; White et al., 2023). Prompt engineering refers to the crafting of these instructions to extract the most optimized, effective, and adequate outputs from the interaction with LLMs (Schulhoff et al., 2024; White et al., 2023). For example, Kojima et al. (2022) demonstrated that adding the now-famous "think step by step" at the end of a prompt increased the accuracy of the output by 61% on a mathematics benchmark (Gao, 2023; Kojima et al., 2022). Research has also shown that giving examples of what you want increases the LLM's ability to generate the desired output (Brown et al., 2020; Gao, 2023; Zhao et al., 2021), but it's also important to keep in mind that they can make mistakes.

The extraction procedure developed in this research note is designed to be content-agnostic, and capable of querying LLMs for various types of political information. While the procedure is demonstrated using politicians' policy characteristics as the example case, the same methodological framework could extract data on voting records, biographical information, or issue positions. However, it was decided to concentrate on politicians' policy characteristics, aiming to yield descriptive data akin to what voters might seek when using tools like ChatGPT to get more information about politicians. The example output was designed as a JSON object, encapsulating the MP/MNA's name and the 10 attributes provided by the model. JSON, known for its lightweight nature, facilitates human readability and writability and allows for straightforward machine production and interpretation (Tan & Motani, 2023). The structure of the JSON object was conceived as follows:

```
{
        "Name": "John Doe",
        "characteristic": [
        "Pro-environment",
        "Supports renewable energy",
        "Advocates for education reform",
        "Pro-healthcare reform",
        "Anti-corruption",
        "Economic growth focus",
        "Supports tax reform",
        "Pro-immigration reform",
        "National security emphasis",
        "Supports digital privacy"
        ]
}
```

The 10 policy attributes presented in our JSON template were deliberately selected to represent diverse policy domains (environment, economy, social issues, etc.) that are common across political systems. These same characteristics remained consistent across all extraction trials and models to ensure procedural standardization. Generic placeholders (e.g., characteristic1, characteristic2) were initially tested but found to be ineffective for eliciting varied model responses, often leading to repetitive attributes across different politicians. By providing concrete, diverse examples spanning multiple policy areas, established prompt engineering practices were followed where exemplar diversity helps prevent the model from fixating on a single domain (Brown et al., 2020; Gao, 2023; Zhao et al., 2021). It is important to emphasize that these example characteristics serve a purely procedural function within the extraction methodology—they demonstrate the expected output format and diversity to the model rather than representing substantive political categories aimed to be measured.

Politicians' party affiliations were deliberately omitted from prompts as part of our extraction procedure design. This methodological choice illustrates how researchers must consider potential sources of bias within the extraction process itself. The procedural challenge encountered was ensuring consistent data extraction across all subjects, including less prominent MPs/MNAs. The extraction protocol addressed this by incorporating specific prompt elements that maximize data completeness. For example, instructing models “to answer to the best of their knowledge” significantly reduces extraction failures, which ensures the generation of a complete dataset.

Refusals to generate outputs can be either attributed to cases of “cannot” (insufficient information) or “should not” (human-imposed guardrails on the model) (von Recum et al., 2024). Therefore, encouraging the model to answer does not necessarily mean that the information generated is invalid, as refusals might stem from guardrails rather than knowledge limitations. When using refusal-breaking mechanisms, researchers either bypass a “should not” restriction (accessing information the LLM actually possesses) or encourage a response in a “cannot” scenario (potentially leading to less certain information). Even with such mechanisms in place, systematic biases may still emerge in how models respond to different types of queries or subjects. The nature of these biases might simply shift from non-response patterns to patterns in the quality or certainty of responses generated under uncertainty. This represents a methodological tradeoff that researchers should explicitly consider when designing their extraction protocols. While this raises interesting questions about output reliability—questions that future research should address—it is critical to emphasize that our contribution focuses on establishing the extraction procedure itself, not validating the accuracy of extracted content.

The methodological priority was to develop a reliable extraction protocol that produces consistent, structured data across all subjects. This exemplifies the broader argument that standardized extraction procedures must be established before researchers can meaningfully evaluate output quality. Additionally, procedural elements from prompt engineering literature were incorporated, such as the “Take a deep breath and work on this problem step-by-step” instruction (Yang et al., 2023), demonstrating how extraction procedures can integrate established techniques to enhance procedural robustness.

The prompt was designed as follows:

```
prompt < - paste0(“Based on the previous example, Provide a list of 10 key characteristics
  describing,” paste0(data_mps$position[i]), “ ”, paste0(data_mps$name[i]), “‘s policies
  formatted in JSON. Make sure to output 10 characteristics. Please answer to the best of your
  knowledge. Take a deep breath and work on this problem step by step.”)
```

Managing the Output

The extraction procedure identified output management as the most significant methodological challenge in structured LLM data collection. This procedural obstacle requires systematic handling to ensure replicable data extraction. A three-stage approach to output processing was developed: (1) standardized extraction of structured content from potentially noisy responses, (2) robust parsing with error handling mechanisms, and (3) iterative verification and retry protocols. This methodological framework addresses a critical gap in the current literature on LLM data extraction.

Despite requests for JSON object outputs, both gpt-3.5-turbo and gpt-4 sometimes added textual explanations or context to the JSON, explaining the generated characteristics rationale. Fortunately, identifying a JSON object’s start is straightforward, marked by an opening “{”. Any text before the first “{” and after the last “}” was removed. When JSON delimiters were missing or misformatted, an empty string was used to indicate data extraction failure, preserving dataset integrity for analysis. This allowed the script to detect when it needed to prompt the model again with the same request.

JSON malformation was the main error type that prevented successful parsing. This issue occurred more frequently with GPT-3.5 than with GPT-4, which leads us to believe that while this issue may be mitigated with future model improvements, robust error handling remains a crucial component of any LLM data generation process.

Missing values can significantly impact dataset quality when conducting statistical analysis, as a single missing value in a row can render the entire row irrelevant for certain types of analysis. This concern extends beyond mere technical challenges—the distribution of missing or malformed responses may not be random across the dataset, potentially revealing systematic biases in how language models process different types of information. For instance, if extraction failures occur disproportionately for politicians with certain characteristics or ideological positions, this pattern itself could become a valuable data point for understanding model limitations. The processing approach described here, with its iterative retry mechanism, was designed specifically to address these challenges by minimizing missing values. However, researchers employing LLMs in political science should remain attentive to patterns in initially failed extractions, as these may reveal important insights about model biases, though a comprehensive analysis of such patterns falls beyond the scope of this methodological note.

After extraction, the JSON string was parsed with the jsonlite package, transforming it into a structured R object. This transformation enabled access to the policy characteristics within the JSON. To enhance robustness, error handling mechanisms were implemented to address parsing failures, mainly due to malformed JSON structures. This error handling was crucial for preserving the data processing workflow's integrity, ensuring the data collection and analysis process remained seamless, even with parsing errors.

After successful parsing, the presence of policy characteristics in the JSON object was verified. When the expected data were present and properly formatted, they were allocated to the appropriate columns in the dataset. Conversely, if valid policy characteristics were missing or parsing failed, the attempt was logged. If the maximum trial limit had not been reached, another trial was initiated after a short pause. This iterative method, along with structured error handling, guaranteed high data integrity in the dataset.

The repetition-until-completeness approach has proven to be effective in our data collection process. While it may not be the only solution to address missing or malformed data, it offers an implementation that is both straightforward and efficient. This iterative approach allows for multiple attempts to obtain the required information, systematically improving the completeness of the dataset with each retry. The effectiveness of this method is demonstrated by the completion rate achieved in our final dataset, where all targeted politicians were successfully processed.

This methodological framework addresses a critical gap in the current literature on LLM data extraction, offering a replicable approach to maintaining data integrity when working with potentially inconsistent model outputs.

Describing the Dataset

To illustrate the application of the extraction procedure, a brief demonstration of how the extracted data can be processed is presented. The following analysis is not intended as a substantive contribution, but rather as proof-of-concept of the types of data that can be systematically extracted using this methodology. This demonstration underscores the procedural value of having standardized extraction methods before conducting any substantive analysis of LLM-generated political data. Categorizing the characteristics reveals thematic trends and biases in the data, laying the groundwork for deeper analysis.

Initially, the most common words and bigrams from the characteristics generated by the three models were used to construct topic dictionaries. All words and bigrams with a cumulative frequency of over 100 were incorporated, resulting in 121 words/bigrams for categorization into primary categories. These 121 words/bigrams covered 94.5% of the characteristics, meaning at least one of these terms appeared in 94.5% of the characteristics, allowing for their categorization into defined categories.[4] Given the characteristics of the models are concise (up to 10 words, excluding stopwords), this method effectively captures the dataset's essence.

Figure 1 shows the categorical distribution from the three models' outputs. The analysis concentrates on the top six categories: economic development, social issues, environment, health, education, and fiscal policy, for visualization purposes. This focused examination helps gain a deeper understanding of the prevailing themes and the disparities in focus among various political parties. This approach enables a detailed exploration of key areas emphasized by the models, offering insights into potential biases in the generated data.

Figures 2 and 3 show the distribution of characteristics across the top six categories for federal and provincial parties, respectively. This research note primarily aims to describe the methodology for extracting political data from LLMs, not to evaluate the accuracy or integrity of the category rankings from different LLM models. Thus, these figures are presented to showcase the data type obtainable through ChatGPT model prompts. However, an interesting observation from Figures 2 and 3 is the clear variation in the model responses, as shown by the differing proportions of characteristics across categories. These variances suggest further research is needed into the precision and potential biases of LLM-generated content in political analyses. The discrepancies hint that some models may be more accurate or less biased than others, a hypothesis deserving future exploration.

These preliminary observations highlight the methodological importance of the standardized extraction procedure, as it enables systematic comparison across different LLM models. By maintaining consistent extraction parameters while varying only the model source, researchers can use this procedure to investigate inter-model reliability—a critical methodological consideration before drawing substantive conclusions from LLM-generated data. The procedure outlined

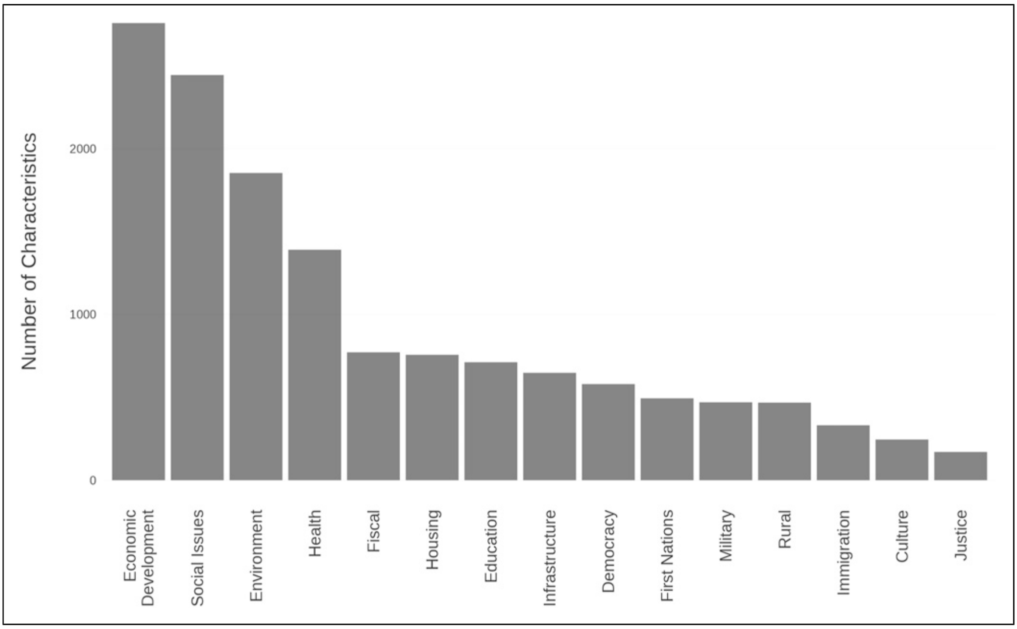

Figure 1. Distribution of the categories in the characteristics provided by the OpenAI models.

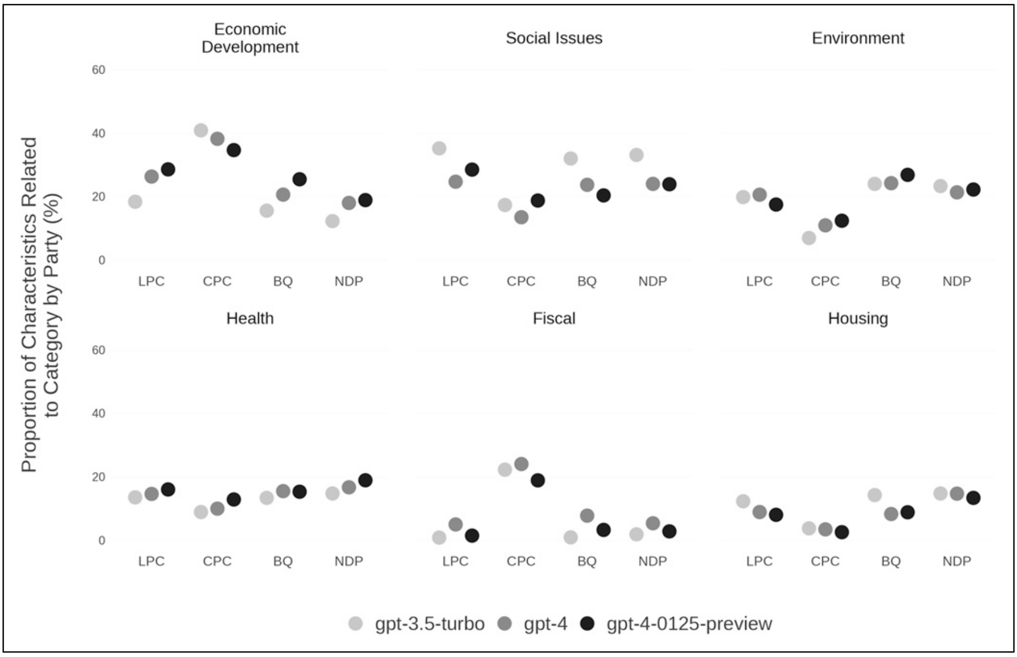

Figure 2. Proportion of characteristics related to top-6 categories, by federal party.

provides a structured framework for such methodological assessments that can be adapted to various political contexts and research questions. At the same time, the observed differences between models underscore the importance of assessing inter-model variability, as discussed in the introduction. The method proposed in this study provides a structured approach to evaluate these variations systematically.

Discussion

This research note introduces a procedure for extracting outputs from LLMs, specifically ChatGPT, using the OpenAI API. As mentioned earlier, the critical examination of inherent biases in LLMs is gaining momentum. Given LLMs' widespread accessibility[5] and potential societal impact, this emerging research agenda demands significant attention from social science researchers. The suggested procedure seeks to offer a "standardized" and easy way to extract data, which, in turn, fosters transparency, accessibility, and replicability to extract data from LLMs. This flexible framework serves as a methodological foundation that can support diverse research aims, including (but not limited to) future work scrutinizing ChatGPT's biases in electoral contexts, while maintaining the primary focus on the extraction procedure itself rather than the content analysis. A hypothetical research question was used to showcase how this procedure can be implemented. The goal was not to suggest an answer to it, nor to estimate the biases ChatGPT holds toward Canadian and Quebec MPs. It only served as an illustrative case. Yet a few things are worth noticing regarding the data generated. The descriptive statistics presented in this note reveal subtle differences in OpenAI models' interpretations of political parties, gender distinctions, and regional differences between Quebec MPs and other Canadian MPs. These variations imply that the information users receive about political parties can differ based on the chosen model. Furthermore, a paywall for accessing "state-of-the-art" models may create disparities in the

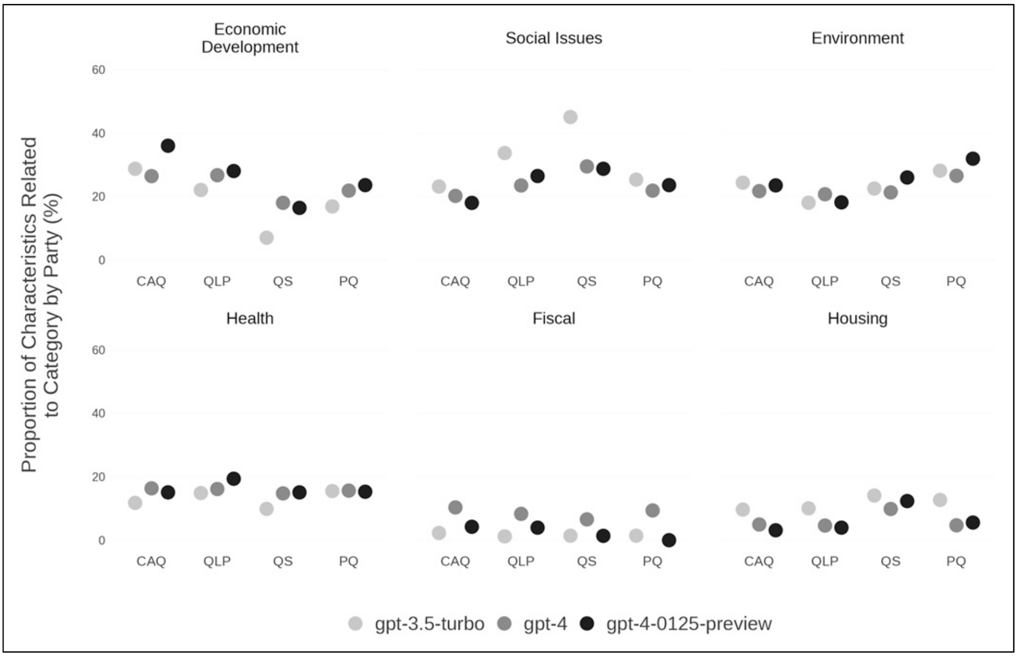

Figure 3. Proportion of characteristics related to top-6 categories, by Quebec provincial party.

chatbot's information quality and accuracy. Determining if these differences are biases and their potential sources—training corpus, model parameters, or human-defined limits—requires deeper investigation by future research.

While this research note focused exclusively on the presentation of the procedure to extract data, it fails to account for the validity of the output. This is mainly due to the goal of this note, which was to suggest a framework to foster transparency, accessibility, and replicability. However, this does not change the importance of validating the data generated. Recent studies discuss this topic, whether as the central focus of their paper or as evaluating the utility of LLMs in generating artificial samples, and suggest many ways to do so, depending on the data generated (e.g., Aldeen et al., 2023; Argyle et al., 2023; Gilardi et al., 2023; Hämäläinen et al., 2023; Pangakis and Wolken 2024; Törnberg 2024; Trott, 2024). Many metrics and strategies are suggested to assess the validity of the data generated by ChatGPT whether it is using F1-score, Cohen's Kappa, Precision, Recall, and by comparing these with human-generated data. Furthermore, Hämäläinen et al. (2023) suggest many requirements for satisfactory validation. Others, like Pangakis and Wolken (2024) propose a measure to identify an LLM's confidence in an annotation. All of these considerations are very important to assess and estimate the validity of LLMs and their capacities. Yet, as mentioned earlier, the important step to do before estimating these is to generate the data. Hence, the importance to have a clear and accessible procedure to do so, where anyone can easily assess how the data was generated.

To further enhance transparency and reproducibility, researchers using this procedure should systematically document several key elements: the exact model version used, temperature and sampling parameters, precise extraction dates, and variations in performance observed across multiple trials. This information is essential because, as recently highlighted by Barrie et al. (2024), LLMs present unique reproducibility challenges that differ from traditional human coding methods. Indeed, their non-deterministic nature, combined with the rapid updates of commercial models, can compromise the stability of results over time. The standardized approach presented in this research note specifically allows for the assessment of this variability by establishing a consistent methodological framework for future studies, thus promoting the comparability of results obtained by different researchers. This standardization constitutes a fundamental step toward a more rigorous use of LLMs in social sciences.

The procedure proposed here can be used in several other situations. For example, scholars could compare results with human-generated data, such as parliamentary speeches, survey responses, tweets, and Wikipedia entries. Comparing ChatGPT's outputs with these datasets could help researchers formulate hypotheses about the biases' nature and extent in the model's responses. This comparison might reveal how LLMs like ChatGPT process human language and thought, highlighting how training data or algorithms influence output significantly. Furthermore, as the illustrative cases suggest, this procedure can also be used to estimate the variability in models and between models. This can be an interesting avenue to explore, especially to understand the capacities of these models to give reliable answers. AI and LLMs research in political science is in its early stages, offering a broad array of unexplored research opportunities. Mastering data retrieval and analysis methodologies is crucial for advancing in this field. Researchers are encouraged to delve deeper into these subjects due to their significant ethical implications and their impact on public discourse. This emerging field allows for examining AI's impact on democratic processes and contributing to the creation of transparent, accountable, and unbiased AI systems.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Laurence-Olivier M. Foisy https://orcid.org/0009-0004-7505-9477
Étienne Proulx https://orcid.org/0009-0005-8671-1018
Jérémy Gilbert https://orcid.org/0009-0001-8915-2865
Jozef Rivest https://orcid.org/0000-0002-3195-3838
Alexandre Bouillon https://orcid.org/0009-0007-0241-5778
Yannick Dufresne https://orcid.org/0000-0002-6211-2193

Notes

1. The House of Commons of Canada officially comprises 338 seats; however, the dataset includes 340 MP entries due to changes during the legislature.
2. Detailed instructions and annotations within the GitHub repository assist social scientists in customizing the script for their needs.
3. The threshold frequency of 100 was initially chosen arbitrarily but was deemed appropriate upon determining that it covered 94.5% of the characteristics. Given that each word and bigram was manually coded, setting a threshold was necessary to ensure feasibility while maintaining comprehensive coverage of the dataset.
4. ChatGPT is readily accessible to anyone with a computer or smartphone, further facilitated by the integration of AI into Microsoft's Bing search engine, and Google's Gemini.
5. https://github.com/clessn/prompting_the_machine.

References

Agnew, W., Bergman, A. S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., & McKee, K. R. (2024). The illusion of artificial inclusion. In Proceedings of the CHI conference on human factors in computing systems, Honolulu, HI, 11–16 May 2024 (pp. 1–12). https://doi.org/10.1145/3613904.3642703

Aldeen, M., Luo, J., Lian, A., Zheng, V., Hong, A., Yetukuri, P., & Cheng, L. (2023). ChatGPT vs. Human annotators: A comprehensive analysis of ChatGPT for text annotation. In 2023 International conference on machine learning and applications (ICMLA), Florida, USA, 15–17 December 2023 (pp. 602–609). https://doi.org/10.1109/ICMLA58977.2023.00089

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Barrie, C., Palmer, A., & Spirling, A. (2024). Replication for language models problems, principles, and best practice for political science. https://arthurspirling.org/documents/BarriePalmerSpirlingTrustMeBro.pdf

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020, July 22). Language models are few-shot learners. https://doi.org/10.48550/arXiv.2005.14165

Dahl, R. A. (2006). *On political equality*. Yale University Press.

Fletcher, R., & Nielsen, R. (2024). *What does the public in six countries think of generative AI in news?* Reuters Institute for the Study of Journalism.

M Foisy, L. O., Drouin, J., Pelletier, C., Rivest, J., Cadieux, H., & Dufresne, Y. (2024). Ain't no party like a GPT party: Assessing OpenAI's GPT political alignment classification capabilities. *Journal of Information Technology & Politics*, 1–13. https://doi.org/10.1080/19331681.2024.2444587

Gao, A. (2023). Prompt engineering for large language models. Available at SSRN 4504303.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(30), Article e2305016120. https://doi.org/10.1073/pnas.2305016120

Guo, W., & Caliskan, A. (2021). Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society, USA, 19–21 May 2021 (pp. 122–133). https://doi.org/10.1145/3461702.3462536

Hämäläinen, P., Tavast, M., & Kunnari, A. (2023). Evaluating large language models in generating synthetic HCI research data: A case study. In Proceedings of the 2023 CHI conference on human factors in computing systems, Hamburg Germany, 23–28 April 2023 (pp. 1–19). https://doi.org/10.1145/3544548.3580688

Hartmann, J., Schwenzow, J., & Witte, M. (2023). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *SSRN Electronic Journal*.

Heyde, L. V. D., Haensch, A.-C., & Wenz, A. (2024, January). Assessing bias in LLM-Generated synthetic datasets: The case of German voter behavior. SocArXiv 97r8s. Center for Open Science. https://doi.org/10.31235/osf.io/97r8s

Johnson, R. L., Pistilli, G., Menédez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J., & Bertulfo, D. J. (2022, March 15). The ghost in the machine has an American accent: Value conflict in GPT-3. https://arxiv.org/abs/2203.07785

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 22199–22213, 1613. https://dl.acm.org/doi/10.5555/3600270.3601883.

McGee, R. W. (2023, February 15). Is chat gpt biased against conservatives? An empirical study. (SSRN Scholarly Paper 4359405). https://doi.org/10.2139/ssrn.4359405

Meskó, B. (2023). Prompt engineering as an important emerging skill for medical professionals: Tutorial. *Journal of Medical Internet Research*, *25*, Article e50638.

Motoki, F., Pinho Neto, V., & Rodrigues, V. (2023). More human than human: Measuring chatgpt political bias. Available at SSRN 4372349. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4372349

Ouyang, S., Zhang, J. M., Harman, M., & Wang, M. (2023, August 5). LLM is like a box of chocolates: The non-determinism of ChatGPT in code generation. https://doi.org/10.48550/arXiv.2308.02828

Pangakis, N., & Wolken, S. (2024). Keeping humans in the loop: Human-centered automated annotation with generative AI. arXiv.org. https://doi.org/10.48550/ARXIV.2409.09467

Rozado, D. (2023). The political biases of ChatGPT. *Social Sciences*, *12*(3), 148. https://doi.org/10.3390/socsci12030148

Rudnytskyi, I. (2023). Openai: R wrapper for OpenAI API (Version 0.4.1) [Computer software]. https://cran.r-project.org/web/packages/openai/index.html

Rutinowski, J., Franke, S., Endendyk, J., Dormuth, I., & Pauly, M. (2023). The self-perception and political biases of ChatGPT.

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., … Resnik, P. (2024). The prompt report: A systematic survey of prompting techniques. arXiv preprint arXiv:2406.06608.

Tan, J. C. M., & Motani, M. (2023, October 8). Large Language model (LLM) as a system of multiple expert agents: An approach to solve the abstraction and reasoning corpus (ARC) challenge. https://doi.org/10.48550/arXiv.2310.05146

Törnberg, P. (2024). Best practices for text annotation with large language models. arXiv.org. https://doi.org/10.48550/ARXIV.2402.05129

Trott, S. (2024). Can large language models help augment English psycholinguistic datasets? *Behavior Research Methods*, *56*(6), 6082–6100. https://doi.org/10.3758/s13428-024-02337-z

Van den Broek, M. (2023). *ChatGPT's left-leaning liberal bias*. University of Leiden. https://www.universiteitleiden.nl/binaries/content/assets/algemeen/bb-scm/nieuws/political_bias_in_chatgpt.pdf

von Recum, A., Schnabl, C., Hollbeck, G., Alberti, S., Blinde, P., & von Hagen, M. (2024). Cannot or should not? Automatic analysis of refusal composition in IFT/RLHF datasets and refusal behavior of black-box LLMs. arXiv preprint arXiv:2412.16974.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:2302.11382.

Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., & Chen, X. (2023, December 7). Large Language models as optimizers. https://arxiv.org/abs/2309.03409

Zack, T., Lehman, E., Suzgun, M., Rodriguez, J. A., Celi, L. A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D. W., & Abdulnour, R.-E. E. (2023). Coding inequity: Assessing GPT-4's potential for perpetuating racial and gender biases in healthcare. medRxiv, 2023–2007. https://www.medrxiv.org/content/10.1101/2023.07.13.23292577.abstract

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In: Proceedings of the 2023 CHI conference on human factors in computing systems, Hamburg Germany, 23–28 April 2023 (pp. 1–21). https://doi.org/10.1145/3544548.3581388

Zhao, J., Ding, Y., Jia, C., Wang, Y., & Qian, Z. (2024). Gender bias in large language models across multiple languages. arXiv preprint arXiv:2403.00277.

Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021, July). Calibrate before use: Improving few-shot performance of language models. In International conference on machine learning (pp. 12697–12706). PMLR.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, *50*(1), 237–291. https://doi.org/10.1162/coli_a_00502

Author Biographies

Laurence-Olivier M. Foisy is a PhD student in Political Science at Université Laval. His research agenda centers on the application of advanced digital methods to social science research, with particular emphasis on the integration and assessment of artificial intelligence tools in research methodology.

Étienne Proulx is a Master's student in Political Science whose research focuses on textual analysis, parliamentary discourse, and ethical issues related to artificial intelligence. His work explores the intersection of political communication, computational methods, and emerging technologies.

Hubert Cadieux is a master's student in Political Science at Université Laval. His research focuses on advanced quantitative methods, including Bayesian, frequentist, and machine learning approaches, to analyze public opinion and voting behavior, with emphasis on operationalizing growth potential in multi-party systems.

Jérémy Gilbert is a master's student in Political Science at Université Laval. His research primarily focuses on public opinion, local politics and quantitative methods of textual analysis, notably by integrating AI in text annotation procedures.

Jozef Rivest is a master's student in Political Science at the University of Montreal. His research primarily focuses on comparative political behavior in East Asia, particularly Japan, as well as quantitative research methods, with an emphasis on measurement and scaling techniques.

Alexandre Bouillon is a Master's student in Political Science at Université Laval. His research explores public opinion and political parties, with a special emphasis on political marketing and quantitative analysis using clustering algorithms.

Yannick Dufresne is an Associate Professor in Political Science at Université Laval. His research focuses on public opinion analysis using digital methods and the measurement of latent concepts and complex issues through large-scale survey research.