

Applications of GPT in Political Science Research: Extracting Information from Unstructured Text

Kyuwon Lee, *University of Southern California, USA*

Simone Paci, *Stanford University, USA*

Jeongmin Park, *Oxford University, UK*

Hye Young You, *Princeton University, USA*

Sylvan Zheng, *New York University, USA*

ABSTRACT

This article explores the use of large language models (LLMs), specifically GPT, for enhancing information extraction from unstructured text in political science research. By automating the retrieval of explicit details from sources including historical documents, meeting minutes, news articles, and unstructured search results, GPT significantly reduces the time and resources required for data collection. The study highlights how GPT complements human research assistants, combining automated efficiency with human oversight to improve the reliability and depth of research. This integration not only makes comprehensive data collection more accessible; it also increases the overall research efficiency and scope of research. The article highlights GPT's unique capabilities in information extraction and its potential to advance empirical research in the field. Additionally, we discuss ethical concerns related to student employment, privacy, bias, and environmental impact associated with the use of LLMs.

In the expanding landscape of political science research, the integration of advanced artificial intelligence (AI) tools has opened novel avenues for data collection, annotation, and analysis. Among these tools, large language models (LLMs), such as OpenAI's Generative Pre-trained Transformer (GPT), have garnered attention for their potential to enhance research productivity and expand empirical research

capabilities (Ziems et al. 2024).¹ This study specifically examined the use of GPT for information extraction from unstructured text—an essential task that involves retrieving explicitly stated details that may be challenging to access manually. Unlike broader applications—such as generating text labels for classification (Chiu, Collins, and Alexander 2022; Wang 2023), simulating survey responses (Argyle et al. 2023b), generating stimulus for survey experiments (Velez and Liu 2024), and engaging in conversations with humans (Argyle et al. 2023a)—information extraction focuses on accurately identifying and retrieving explicit content within documents. Although GPT shows promise in various tasks, this study highlights their particular effectiveness in information extraction.

Our study is divided into detailed examinations of the utility of GPT for various data-collection tasks. In these examples, GPT's applications demonstrate its versatility in handling increasingly complex information tasks across two languages: English and

Kyuwon Lee is assistant professor of political science and international relations at the University of Southern California. She can be reached at kyuwonl@usc.edu.

Simone Paci is lecturer of political science at Stanford University. He can be reached at simone.paci@stanford.edu.

Jeongmin Park is a Postdoctoral Prize Research Fellow in Politics at Nuffield College at Oxford University. She can be reached at jp49@princeton.edu.

Corresponding author: Hye Young You is an associate professor of politics and public affairs at Princeton University. She can be reached at hyou@princeton.edu.

Sylvan Zheng is a PhD candidate in politics at New York University. He can be reached at saz310@nyu.edu.

Italian. In the first example, GPT is used to clean Optical Character Recognition (OCR) errors from scans of historical documents, demonstrating its basic ability to process textual data. In the more complex applications described in the second and third examples, GPT helps to extract participant information from semi-structured administrative-meeting-minutes data and detailed source information from lengthy news articles. In the fourth example, we show GPT's ability to perform an advanced task of synthesizing data from multiple Internet sources.

Our study is divided into detailed examinations of the utility of GPT for various data-collection tasks. In these examples, GPT's applications demonstrate its versatility in handling increasingly complex information tasks across two languages: English and Italian.

Each of these four applications demonstrates how GPT performs labor-intensive tasks not only with remarkable speed but also with accuracy that either matches or exceeds human efforts. Furthermore, the use of GPT in these contexts highlights its potential to manage large volumes of data—a capability that is particularly useful in political science when researchers often are faced with extensive but only partially structured datasets. The examples presented in this article highlight GPT's strengths in natural-language processing while mitigating its weaknesses in complex reasoning and “hallucination” (i.e., false information) (Ji et al. 2023; Wei et al. 2022) along with the reliability and consistency of synthetic survey data produced by LLMs (Bisbee et al., 2024).

By presenting a range of unique examples, this article expands thinking in the discipline about the potential uses of LLMs rather than providing a specific how-to guide. We discuss the importance of creatively engineering prompts tailored to different tasks, illustrating that the first prompt may not always suffice and that careful refinement is crucial for optimal results. Through this approach, we hope to inspire further exploration and creative problem-solving using LLMs in political science research.

GPT's potential to reduce the gap in unequal research resources is another significant benefit of its inclusion in the political science toolbox. Traditionally, large-scale research projects often have been the purview of well-funded researchers who

APPLICATIONS

This section presents four examples in which LLMs streamline traditionally labor-intensive tasks and enable innovative approaches to data collection and analysis in political science.

Example 1: Cleaning and Analyzing Historical Data

This section explores the use of GPT in conjunction with OCR tools to clean and analyze historical documents. Although OCR

technology has advanced, the quality of output nevertheless depends on the quality of the scanned image and the choice of OCR tool, which often results in errors (e.g., misspellings and odd spacing). High-quality OCR tools such as Google Cloud Vision (GCV) produce cleaner text but often are impractical due to issues such as document accessibility and other resource constraints. To address these challenges, we used the GPT-4-1106-preview model to clean text produced by the open-source OCR tool, Tesseract.

We used previously unused archival materials concerning World War II-era race-related incidents and racial reform from the National Archives in College Park, Maryland. These materials, consisting of five boxes, contain the weekly intelligence reports of the Army Service Forces from August 1944 to January 1946. The reports provide a comprehensive description of race-related incidents involving military personnel, as well as the preventive or reactionary measures taken to mitigate racial strife (see the sample image in figure 1). The records contain important details about these incidents, such as the date and location, the people involved, and the actions taken by key players. Unfortunately, available OCR tools show varying levels of accuracy (see online appendix A, table A1).

This study proposes a time-saving approach that combines open-source tools (i.e., Tesseract) with GPT. We took the noisy text generated by Tesseract and used the GPT Application Programming Interface (API) to clean the noise, a process illustrated

GPT's ability to automate and streamline data extraction and analysis tasks could level the playing field, allowing researchers with limited budgets to undertake more extensive research efforts.

can afford large teams of research assistants (RAs) and expensive data-processing tools. However, GPT's ability to automate and streamline data extraction and analysis tasks could level the playing field, allowing researchers with limited budgets to undertake more extensive research efforts. However, the use of LLMs in research raises ethical concerns, including the potential loss of jobs for student RAs, privacy risks, social bias in output, and significant environmental impacts. The various ethical concerns of using GPT are discussed in detail.

in table 1. We then visualized the performance of this method compared to GCV-processed text for an entire box, consisting of 20 folders (997 images).² We used the GCV-processed text as gold-standard data because of its superior quality once images were obtained and preprocessed for accurate character recognition as well as the impracticality of generating human-typed gold-standard data for large archival materials. We measured the performance of the Tesseract-GPT combination using Character Error Rate (CER), a common metric used to evaluate OCR

Figure 1

Example of a Scanned Image from a Weekly Intelligence Report

CONFIDENTIAL

Weekly Intelligence Summary, 19 August 1944 through 25 August 1944

2- RACIAL ANNEX

citizens of Atlanta proposing that the Negroes be permitted to vote in the 4 July primary. The CIO and the Political Action Committee was held up as a means by which the Negro could win political freedom, by Ernest Wright, a CIO organizer and president of the People's Defense League, from New Orleans, the DAILY WORKER reported.

At 0445 hours on 19 August a young white woman and her male friend were sitting on a wall in front of her home when a Negro slipped up beside them and grabbed her purse and ordered the couple to follow him behind some bushes in the yard. The Negro made known his intention to attack the woman, so the couple broke and ran. The Negro overtook the woman and stabbed her in the back, inflicting minor wounds. He then fled and has not been apprehended.

Two Negro soldiers and their wives complained at the Greyhound Bus Station in Atlanta that they had been mistreated by a bus driver while en route to Fort Benning, Ga. from Atlanta on 16 August. One of the Negro soldiers sat beside a white soldier with the latter's permission. When the bus driver observed the Negro's actions he cursed him, made him move and, when the Negroes disembarked from the bus in protest at a stop near Atlanta, refused to return their tickets. (C-3)

A Negro, 2nd Lt. JIMMIE D. WHEELER, stationed in Walterboro, S.C., reportedly was involved in two incidents on 15 August. While waiting for the bus at the Terminal Station, a white MP informed him that he would have to wait inside the colored waiting room instead of standing in front of the white main waiting room. Lt. WHEELER refused to go inside and demanded to see the duty officer. He took the MP's name and organization and stated that he would make a report of the incident. On the same morning two white enlisted men were reprimanded by Lt. WHEELER and their names and organizations taken, when they failed to salute the Negro while walking down the street. The white soldiers reported that they failed to realize the Negro was an officer as they did not see his insignia. (C-3)

The "leadership school" conducted by the Communist influenced Southern Negro Youth Congress and held in Atlanta 7-18 August was attended by about 35 individuals from the states of Georgia, Alabama, Louisiana, Florida and New York. The seven Negro directors of the school were reported to be active Communists. Such topics as "job security, civil liberties and social welfare" were discussed. On one occasion LUCY RANDOLPH MASON, a member of the C.I.O. Political Action Committee for the southeastern region, addressed the school. (B-2)

ROBERT A. CORLEY, of Atlanta, who was discharged from the Army recently as a psychoneurotic, is reported to be writing a book entitled, "Danger Lies Within". He is reported to be a "violently anti-Negro" agitator and to advocate in his book "complete elimination of Negroes in America". (F-0)

Table 1

OCR Results Using Tesseract and GPT

Description	Result
Prompt	"Do not change the content of the text itself, just take out noise. Also, when you see parentheses with uppercase alphabet and number inside, like (A-2), just leave it:"
OCR (Tesseract)	Two Negro. soldiers and their wives? complained. at the Greyhound ay ea . Bus Station in Atlanta that they had been mistreated by a bus driver i a F while en route to Fort Benning, Ge. from Atlanta on 16 August . One ee: Be } of the Negro soldiers, sat beside a white soldier with the latter's [i fo a cS permission. When the bus driver observed the. Negro'ts actions 16 as Ls. curséd him made him move. and, wén the Negroes disembarked f-m the ' fo Bae be bus in'protést at a stop near Atlanta, refused to return their i ie a tickets. _(C-3)
Cleaned with GPT	Two Negro soldiers and their wives complained at the Greyhound Bus Station in Atlanta that they had been mistreated by a bus driver while enroute to Fort Benning, Ga. from Atlanta on 16 August. One of the Negro soldiers sat beside a white soldier with the latter's permission. When the bus driver observed the Negro's actions cursed him, made him move and, when the Negroes disembarked from the bus in protest at a stop near Atlanta, refused to return their tickets (C-3).
GCV	Two Negro soldiers and their wives complained at the Greyhound Bus Station in Atlanta that they had been mistreated by a bus driver while enroute to Fort Benning, Ga. from Atlanta on 16 August. One of the Negro soldiers sat beside a white soldier with the latter's permission. When the bus driver observed the Negro's actions e cursed him, made him move and, when the Negroes disembarked from the bus in protest at a stop near Atlanta, refused to return their tickets. (C-3)

Note: Errors are highlighted in bold text.

performance. CER is defined as the ratio of the number of character-level errors to the total number of characters in the reference text. We also used GPT to extract critical details from each incident, including the location, main actors involved, and targets. Finally, we extracted a 10% sample from the cases and manually verified the accuracy of the information extracted by GPT.

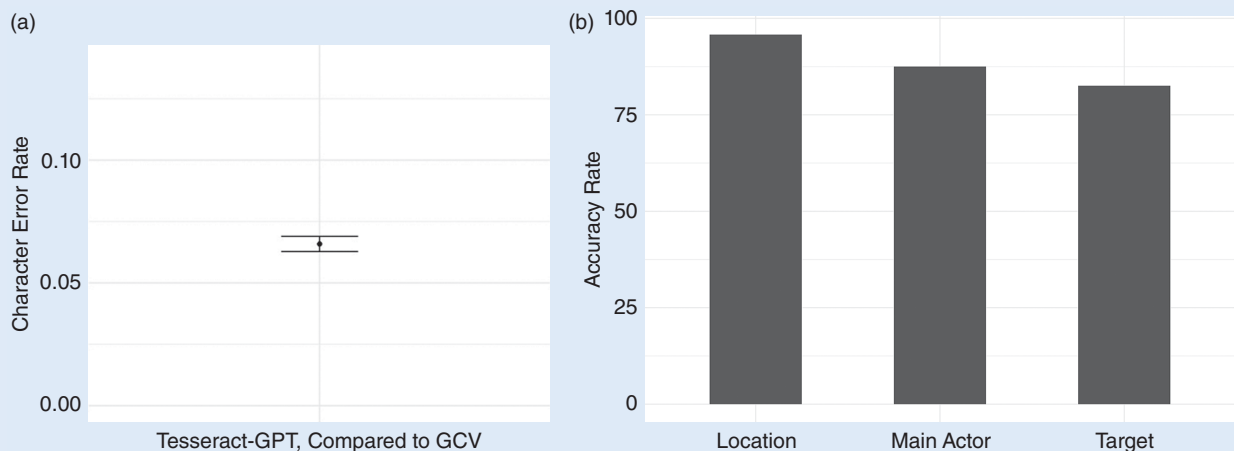
The results shown in figure 2 illustrate the effectiveness of GPT in cleaning and analyzing historical data. On average, about 6% of the characters in the OCR-generated text were incorrect, compared to the text generated by GCV, our reference text (figure 2a). Furthermore, the high accuracy rate for capturing relevant information—such as location, main actor, and target—illustrates the general effectiveness of GPT in information extraction, especially when it is related to objective, context-independent information (e.g., location) (figure 2b).³

Example 2: Extracting Unstructured Administrative Data

This section describes how GPT (specifically, the GPT-4-1106-preview model) can be used to collect and clean administrative data provided in a semi-structured format (often in PDFs). We focused on meeting minutes from federal advisory committees (FACs) within federal agencies in the United States. A significant number of FACs serve as independent advisors that make policy recommendations to federal agencies. These committees hold more than 5,000 public meetings annually, convening committee members, federal agency officials, and outside interest groups to discuss agency policy. As such, FAC meeting minutes provide a unique opportunity for scholars to examine the extent to which outside groups participate in bureaucratic policy making. Figure 3 presents two examples of committee-meeting minutes: from the Environmental Protection Agency (EPA) and the Centers for Disease Control and Prevention (CDC). Each example includes the name, position (e.g., chair, member, agency staff, or public attendee), and affiliation (e.g., Karmanos Cancer Institute) of each meeting participant.

Figure 2

Performance of GPT in Cleaning and Analyzing Archival Data



(a) Average Character Error Rate; (b) Accuracy Rate of Information Summarized

Our goal was to extract the participant information from 79 meeting minutes of two EPA FACs—the Clean Air Scientific Advisory Committee and the Science Advisory Boards—from 2017 to 2023. We used the GPT API and R to extract the name, affiliation, and position of each meeting participant from the FAC meeting minutes and generated structured comma-separated values (CSV) data. Table 2 lists the API prompt and R commands that we used. First, the prompt contains a phrase that asks GPT to create a delimited table of three columns. Second, the prompt contains sentences describing the information that GPT should fill in for each column based on the meeting minutes. Third, the prompt asks GPT to clean the participants' names and remove commas that are not delimiters.

Whereas GPT easily extracted individuals' names and affiliations, it often had difficulty extracting participants' position labels from meeting minutes because the labels were so diverse and broad. For example, "invited speaker" was not included as an example of a participant position in the prompt; as a result, GPT often would label those individuals as something else, such as "registered speaker." This could have been problematic if accurately identifying individuals' positions was critical to understanding their roles in FAC meetings. To address this, researchers can include in the prompt the extensive set of position labels that appear in meeting minutes. However, we also found that simply

adding "etc." at the end of a list of example positions quickly solved the problem by giving GPT the latitude to determine which information in the meeting minutes concerned the participants' positions.

Although researchers may be concerned about data fabrication by GPT, we found that it rarely occurs in tasks like this, in which GPT constructed datasets based on given information. After GPT created datasets from the meeting minutes, undergraduate RAs validated each dataset to ensure that all meeting-attendee information was included in terms of names, affiliations, and positions. In our example, GPT failed to extract complete information from four of the 79 meeting minutes because our prompt did not include a complete list of participant positions. In this case, RAs filled in the position information for those participants that GPT was unable to retrieve from the meeting minutes.⁴

The collected data allowed us to examine who attended these FAC meetings (see the list provided in online appendix B). The data showed that a substantial number of interest groups voluntarily participated in FAC meetings and that their participation rate varied over time. This has not been documented by existing studies of FACs that focus primarily on FAC members appointed by agency heads (Feinstein and Hemel 2020).

Our example shows that the data collection and cleaning process for FAC meetings still requires human validation.

Figure 3

Examples of Federal Advisory Committee Meeting Minutes

(a) Participants:

CASAC Air Monitoring and Methods Subcommittee (See Roster with affiliations, Attachment A):

Mr. George A. Allen
Dr. David T. Allen
Dr. Linda J. Bonanno
Dr. Doug Burns
Dr. Judith C. Chow
Dr. Kenneth Demerjian
Mr. Eric Edgerton
Mr. Henry (Dirk) Felton
Dr. Philip Fine
Dr. Philip Hopke
Dr. Rudolf Husar
Dr. Daniel Jacob
Dr. Peter H. McMurry
Dr. Allen Robinson
Dr. Armistead (Ted) Russell
Dr. James Jay Schauer
Dr. Jay Turner
Dr. Yousheng Zeng

Drs. David Allen, Linda Bonanno, Doug Burns, Phil Hopke, Daniel Jacob, Peter McMurry, James Schauer and Yousheng Zeng could not participate during the June 12, 2014 public teleconference.

EPA SAB Staff:

Mr. Edward Hanlon, Designated Federal Officer

Other Attendees:

A list of persons who requested information on accessing the public teleconference line is provided in Attachment B.

(a) EPA Meeting Minutes; (b) CDC Meeting Minutes

Figure 3
Continued.

(b) ATTACHMENT 2: ROSTER OF THE ACBCYW MEMBERSHIP

CHAIR

Elana Silber, MBA
Executive Director
Sharsheret
1086 Teaneck Road
Site 2G
Teaneck, NJ 07666
PHONE: 201-833-2341
FAX: 201-833-25025
E-mail: esilber@sharsheret.org
Term: 1/17/2017 - 11/30/2020

MEMBERS

Lindsay Avner
Founder and Chairman of the Board
Bright Pink
670 N. Clark Street
Chicago, IL 60654
Phone: 312-787-4412
E-mail: LINDSAY@BeBRIGHTPINK.ORG
Term: 1/18/2017 - 11/30/2020

Michele Maria Cerullo, JD
Assistant General Counsel
DaVita Medical Group, Legal Department
10051 5th Street North, Suite 200
St. Petersburg, FL 33702
PHONE: 727-828-8910
FAX: 727-258-0536
E-mail: MMCERULLO@ICLOUD.COM
TERM: 4/01/2019 - 11/30/2020

EXECUTIVE SECRETARY

Temeika L. Fairley, PhD
Designated Federal Official (DFO)
Office of Program Development
Division of Cancer Prevention and Control
Centers for Disease Control and Prevention
4770 Buford Highway NE., Mailstop F-76
Atlanta, GA 30341
Phone : 770-488-4518
Fax : 770-488-4760
E-mail : tff9@cdc.gov

Lisa Astalos Chism, DNP, APRN, NCMP, FAANP
Clinical Director, Women's Wellness Clinic
Nurse Practitioner
Sexual Health Counselor and Educator
Karmanos Cancer Institute
4100 John R Street
Detroit, MI, 48201
Phone: 313-576-9326
Fax: 313-576-8379
E-mail: chism1@karmanos.org
Term: 4/04/2019 - 11/30/2021

Anna Crollman
Breast Cancer Advocate
5021 Holly Brook Drive
Apex, NC 27539
Phone: 828-712-3706
E-mail: mycancerchic@gmail.com
Term: 4/08/2019 - 11/30/2020

However, having RAs review the GPT-generated data is much less resource-intensive and time-consuming than hiring RAs to build data based on meeting minutes. If the minutes of a meeting contain 50,000 characters (i.e., five to six pages), it would cost 30 cents to run the GPT code on the transcript.

Example 3: Extracting Primary Sources from News Articles

This section describes our approach to using GPT to extract semi-structured data from the extensive, unstructured text of news articles, focusing on identifying the diverse sources cited by journalists. Newspaper articles typically reference a wide range of sources—from politicians and bureaucrats to private citizens and business owners—which significantly influences the information conveyed to the public. Although we focused on newspapers, our approach could be applied to similar tasks, such as extracting witness information from court records and guest appearances in news transcripts.

Identifying sources was particularly challenging due to the length of the input documents and the nuanced integration of source information within the article text, including variations in name and context. In the initial phases of prompt development, we found that GPT had difficulty aggregating sources that were

mentioned by multiple similar names and often failed to extract all sources, especially for longer articles. We suspected that this task was hampered by performance degradation because input text length increases the relatively complex level of reasoning required to identify and aggregate sources (Wei et al. 2023). Based on the common errors that we observed, we divided the source-extraction task into subtasks and used a separate prompt to solve each separately, with the output of one subtask prompt feeding directly into the next. This made the logic of each subtask explicit, which also made debugging easier.

The details of the method are shown in figure 4. First, we identified all quotes and information attributed to third parties in the news article. Second, we aggregated the quotes and information at the speaker or organization level. Third, we transformed the data into structured JSON (i.e., a format for organizing and managing data in a hierarchical structure) that can be processed with any data tool of choice. The full set of prompts and sample output are provided in online appendix C.

To validate our approach, we used the described method to extract 214 sources for 50 articles and employed crowd workers to identify errors in the extracted sources. To ensure worker quality, we included results from only those workers who successfully identified

Table 2
GPT Prompt and API Command in R

Description	Command in R
Prompt	'Return me a csv delimiter table of three columns, "name," "affiliation," and "position." Do not return anything else except for the table. The first column "name" has the names of meeting participants and people, if any, who made public comments. When writing down names, remove any prefix, suffix such as Ph.D. or MPH, and texts within parentheses. The second column "affiliation" should have the information on people's affiliation. The third column should be labeled as "position" and specify whether people are "chair," "members," "Designated Federal Officer," "epa staff," "public participants," or "registered speakers," etc. Fill in all values for the "position" column. Remove all commas for values in columns. Use the following text to create the table:'
Read PDF into R	minutes=pdf_text("minutes.pdf") %>% str_split("_n")
Run GPT API	response=POST(url="https://api.openai.com/v1/chat/ completions", add_headers(Authorization=paste("Bearer", apiKey)), content_type_json() , encode="json", body=list(model="gpt-4-1106- preview", temperature=1, messages=list(list(role="user", content=paste(c(prompt,unlist(pdf_2022_4 [eval(parse(text=print(meeting\$page[1])))])), collapse=" "))))))
Extract API Response	capture.output(cat(content(response)\$choices[[1]] \$message\$content))

intentional errors that we embedded in the worker task (see online [appendix C](#) for details about crowd-worker sourcing and screening).

We identified three types of errors: minor details (i.e., incorrect title, name, or organization); false sources (Type I), in which the extracted source was not cited in the article; and missing sources (Type II), in which a source present in the article was not extracted. We manually reviewed each error identified by the crowd workers and estimated the overall error rates. Our results show that the GPT-based system was highly accurate in extracting source details and rarely made Type I or Type II errors (i.e., all error rates were less than 5%). [Figure 5](#) lists the error rates with 95% confidence intervals. Furthermore, a manual inspection revealed that the majority of errors were edge cases, for which it is difficult to determine with certainty the difference between a source citation and a mere mention of a particular entity (e.g., "President Xi Jinping of China has vowed repeatedly to move ahead with steps in his country to curb climate-altering pollution..."). In particular, when crowd workers noticed that source entities extracted by GPT were not cited in the article (i.e., Type I errors), these entities were always mentioned at least in the text. In other words, these errors were exclusively mistakes in judging whether a mentioned entity (i.e., Xi Jinping in the example) should be considered a cited source as opposed to outright hallucination of source entities. The remaining true missing-source (i.e., Type II) errors tended to occur in longer articles with six or more sources.

We used this set of prompts to extract 31,431 sources from 5,795 *New York Times* articles about climate change during the period

2012–2022 using the "GPT-4 Turbo" model. [Figure C1](#) in online [appendix C](#) shows the distribution of sources and articles per year. The total cost of the extraction and validation was \$1,300.

Example 4: Extracting Elite Biographies from Online Sources

This section leverages GPT to extract specific information from an unstructured corpus of sources obtained through systematic Google searches. This exercise reflected a broad category of data-collection tasks for which researchers could not rely on a specific set of source materials or a corpus of structured text. In these cases, data collection involved searching for sources as well as extracting the relevant information. As a result, data collection drew from various sources, including websites, news articles, and academic and expert texts.

We replicated a large human-coded data-collection effort by Montano, Paci, and Superti (2024), which examined whether having a daughter influenced the pro-women policies of Italian mayors. The original study reflected a growing interest in political science in the role of elite biographical characteristics (Krcmaric, Nelson, and Roberts 2020). However, this approach faced a significant challenge because systematic biographical data rarely are readily available. As a result, researchers must resort to time-consuming and expensive data collection. The original effort by Montano, Paci, and Superti (2024) leveraged systematic Google searches for 1,800 mayors. It was conducted by three RAs from July 2023 to February 2024. For each mayor, the RAs reviewed up to the first 20 available search results for a total of more than 7,300 Italian webpages.⁵ Each link was checked for three pieces of information: whether it contained any information about the mayor's children, the number of kids, and the number of daughters.

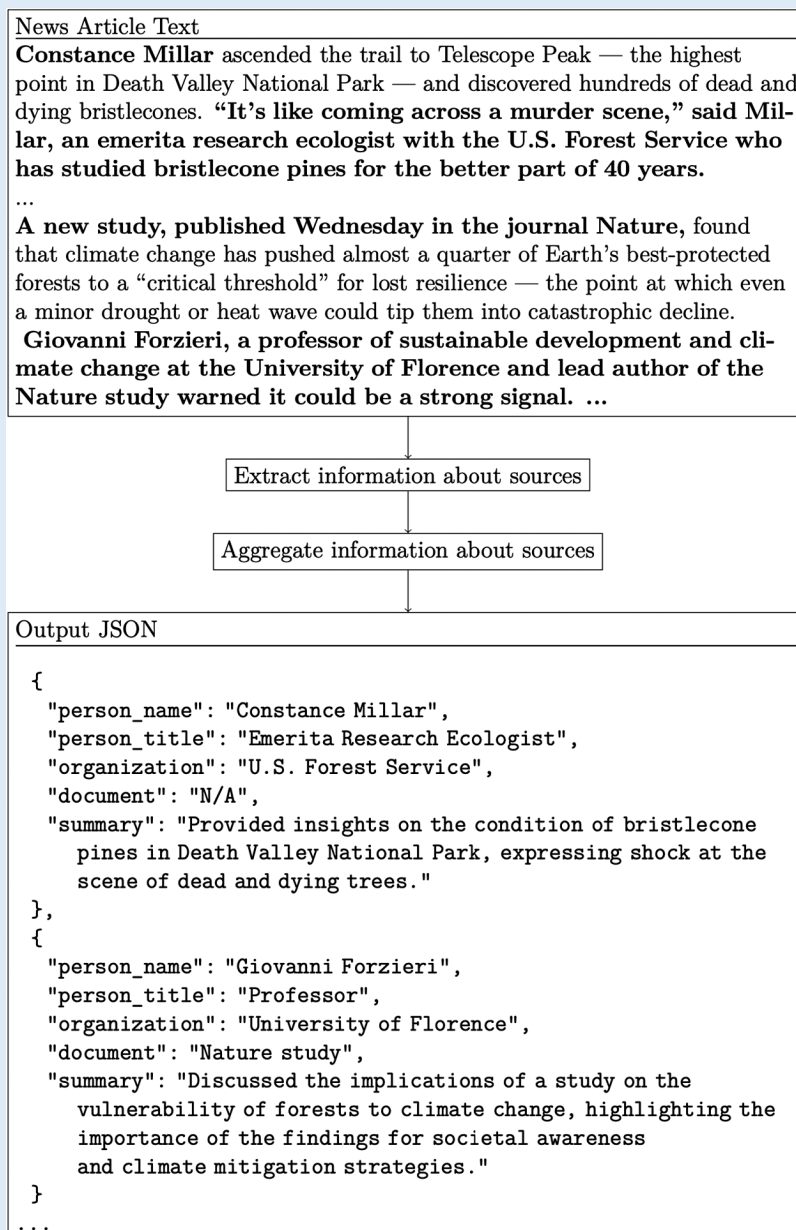
We automated this process by scraping the original links and feeding the text into the GPT-4 Turbo API along with a carefully engineered prompt (see [table D1](#) in online [appendix D](#)). We developed the prompt through an iterative trial-and-error procedure that sampled random draws from the list of webpages and manually checked the model output. The final prompt included instructions to make informational extraction more efficient, especially in edge cases. For instance, we directed GPT to infer gender from names and to assume that the mayor had at least one child if it was mentioned that he had grandchildren. Furthermore, because each webpage came from search results about a specific mayor, we could develop mayor-specific prompts, specifying their name and municipality.

This task tested GPT-4's ability to parse through ambiguous and heterogeneous data. Most sources (about 90%) did not contain relevant information. The relevant information was encoded in myriad ways and the nuance of textual clues could be misleading. [Table 3](#) presents illustrative examples of GPT-4 output. In three cases, GPT-4 correctly recovered the source information. The fourth case was an example of an error in which the information was encoded in a complex way. The text mentioned the mayor's "only son" and his two daughters. GPT-4 understood this as the mayor having three children whereas, in truth, the two were the mayor's son's daughters and thus the mayor's granddaughters—not to be counted as his direct offspring.

Given the same set of search-result links, we estimated the error rate of human coders and of GPT-4. We considered as ground truth all cases in which human coders and GPT-4 agreed. For all disagreements, we adjudicated between the two sets with a third round of human coding, assisted by new RAs. For cases in

Figure 4

Source-Extraction Process Outline



which all three rounds disagreed—only seven of the total sample—the authors manually coded the ground truth.

Figure 6 illustrates the error rates against the ground truth by the original group of human coders and GPT-4. Across the three main pieces of relevant information, GPT-4 outperformed human coders. Figure D1 in online appendix D sorts the overall error rate into categories of mistakes: Type 1 (false positives), Type 2 (false negatives), and Type M (magnitude).⁶ Compared to human coders, GPT-4 made fewer Type 1 errors and more Type 2 errors. On the one hand, this pattern is reassuring because GPT-4’s output may not require extensive validation given its lower rate of false positives. On the other hand, it also suggests that GPT-4

may omit some information, probably whenever it is encoded in an ambiguous or complex way.

We also tested GPT-4’s ability to self-assess and found mixed results. The prompt asked GPT-4 to produce confidence ratings, on a scale of 0 to 100, about the accuracy of its output. The results are shown in Figure D2 in online appendix D. Whenever GPT-4 expressed a confidence rating less than 100, the error rate increased significantly, from 2.8% to 27.3%. However, GPT-4 often expressed overconfidence, giving a rating of 100 to half of the errors found in this exercise. As such, confidence ratings can be considered only as a noisy indicator of potential error.

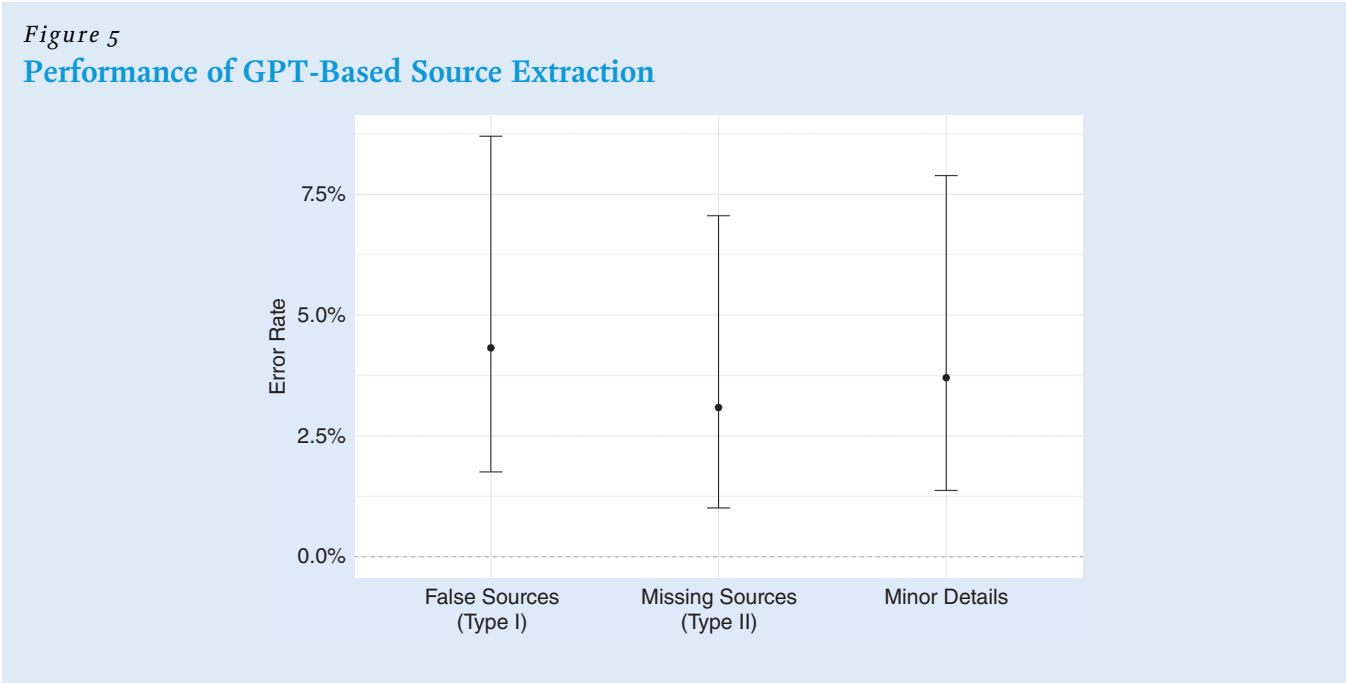


Table 3

Examples of GPT-4 Information from Google Search Results

SOURCE TEXT RELEVANT MENTIONS (TRANSLATED FROM ITALIAN)	EXTRACTED DATA
Success Case: Direct Information Encoding He lives in Gualdo Tadino with his partner Consuelo and their daughter Asia.	Information Found: 1 Number of Kids: 1 Number of Daughters: 1 Confidence: 90
Success Case: Indirect Information Encoding For me, these last few months have been full of surprises. The first, the most beautiful is the growth of my family which will soon expand.	Information Found: 1 Number of Kids: 1 Number of Daughters: NA Confidence: 80
Success Case: Complex Information Encoding As institutions and educational communities, we have a strong responsibility: to offer alternative and healthy models of sociality that allow our children to enjoy and rejoice in their age without exposing themselves to unnecessary risks.	Information Found: NA Number of Kids: NA Number of Daughters: NA Confidence: 100
Failure Case: Complex Information Encoding Alessandro Zanonato, 35 years old, is the mayor's only son and lives with Chiara, a lawyer like him, and two daughters.	Information Found: 1 Number of Kids: 3 Number of Daughters: 2 Confidence: 90

LIMITATIONS AND BEST PRACTICES

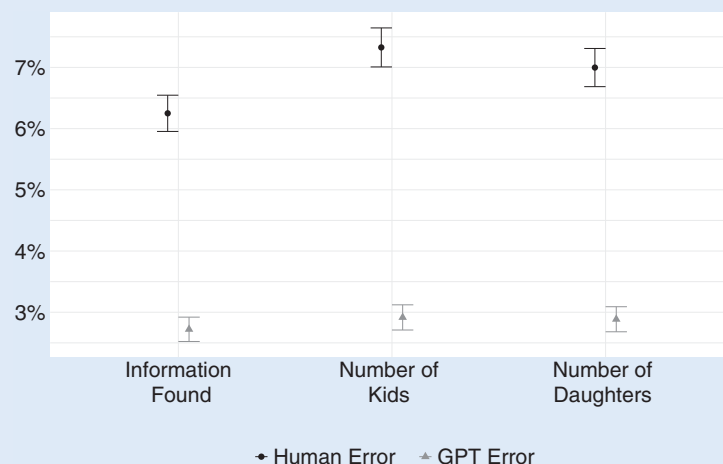
These four applications focus on data collection, cleaning, and extraction tasks that are tedious but common in quantitative political science research. These types of tasks allow for a

straightforward application of LLMs while minimizing the potential for reasoning errors and hallucinations. However, despite their straightforward nature, our applications also have limitations. In the context of data cleaning and collection, we also highlight important limitations and best practices. These recommendations integrated our experience and findings from our validation exercises in this study along with advice on emerging best practices for LLM use and prompt engineering (Ekin 2023).

First, LLM performance is extremely sensitive to the specific prompt used. The term “prompt engineering” has emerged to describe the process of tailoring the LLM prompt to the task at hand. This task is iterative and potentially idiosyncratic to the specific application. However, general guidelines can improve the process. In our experience, the best-performing prompts include several common components. First, the prompt should describe the task context, including the main objective and the type of input data. In addition, researchers should specify the output format, providing detailed descriptions of each data field. Prompts also may include examples of common information-encoding patterns or even be constructed computationally to incorporate document-specific context. For complex tasks, we encourage researchers to explore multistep prompts, as demonstrated in Example 3, or to ask the model to explain its reasoning before providing data, as recommended by Wei et al. (2022).

Second, the context window of LLMs limits the length of both input and output text generated by the model. LLM performance also degrades as the text length increases, even for documents that fit comfortably within the context window. Figure D3 in online appendix D shows that GPT made more errors in identifying the mayor’s children as the length of the input text increased. A practical guideline is to limit texts to well under half of the advertised context window by selecting portions of the text that contain relevant keywords or by breaking tasks into smaller segments. (See figure D3 in online appendix D for the relationship between source text length and coding errors.)

Figure 6
Human Coders and GPT-4 Coding Error Rates



Third, GPT occasionally does not follow the task instructions. This behavior can manifest as incomplete responses, incorrect column names, or incorrect data output. Whereas prompt engineering can mitigate these issues, we found that, in most cases, simply rerunning the same prompt multiple times until the output was well formed was sufficient. Similarly, researchers can leverage

standard of performance. However, alternative models, such as Google Gemini and Anthropic’s Claude, also may be worth considering.⁷ In particular, open-source models, such as Llama and Mistral, offer significant cost and reproducibility advantages (Spirling 2023). However, they require more-technical setup procedures and potentially lower generalized performance.

In the context of data cleaning and collection, we also highlight important limitations and best practices. These recommendations integrated our experience and findings from our validation exercises in this study along with advice on emerging best practices for LLM use.

logical dependencies across data fields to check for response coherence. For instance, in Example 4, we checked that the number of children (of both genders) was greater than or equal to the number of daughters. A related concern is the production of hallucinations, or false information. In our experience with data-collection and cleaning tasks, outright hallucinations have not occurred. Researchers can experiment with the temperature parameter, which affects how much the LLM relies on current input data relative to its training data. Lower values reduce the likelihood of hallucinations but increase sensitivity to prompt wording and reduce reasoning ability. Temperature values range from zero to two; we kept it less than one in all of our examples.

Fourth, we make a few recommendations to improve the ergonomics of interacting with the GPT API. We recommend allowing the model to record the portions of the texts from which it extracts information. This addition can facilitate validation and shed light on the inner workings of the LLM information processing to aid in debugging. To simplify output data management, we recommend instructing the model to limit output to JSON or CSV/Tab Separated Values format (e.g., “provide only the table and nothing else”).

Fifth, we note the existence of competing LLMs in addition to OpenAI’s GPT. We focused on the GPT family of models in this study because of their ease of use, widespread adoption, and high

ETHICAL CONSIDERATIONS

The use of LLMs raises ethical concerns related to professional, privacy, and environmental issues. Researchers should consider whether the potential costs of these novel tools outweigh the added efficiencies. Similarly, we encourage practitioners to consider strategies that limit or offset any negative downstream consequences of integrating LLMs into the research process.

First, the applications presented in this article outsource tasks traditionally performed by student RAs. Although this improves the cost effectiveness of data collection, it undermines student-employment opportunities. These opportunities provide students with not only financial support but also valuable research experience and insight into academic work, potentially influencing some to pursue graduate studies. The RA experience strengthens students’ résumés and also provides an important pedagogical opportunity for experiential learning. We encourage researchers to continue the practice of hiring promising students as RAs. The use of LLMs does not completely eliminate the need for RAs because validation requires thorough human coding. Outsourcing repetitive data-entry tasks to LLMs can free up time and resources to offer students more rewarding and intellectually stimulating tasks, such as exploratory literature reviews and more complex data management.

Second, LLMs raise potential privacy concerns. Given the rapid development of these models, no clear consensus has emerged on the confidentiality risks associated with input data (Wu, Duan, and Ni 2024; Yao et al. 2024). Therefore, we recommend that researchers exercise caution and avoid using the proposed techniques for any sensitive data.

Third, both research and anecdotal evidence shows that LLMs may exhibit social biases embedded in their training data (Hida, Kaneko, and Okazaki 2024). As a result, information-extraction tasks may produce output data that are consistent with the model's underlying biases, such as relying on stereotypes to decide ambiguous cases. Researchers should evaluate whether their applications may be susceptible to this problem and focus validation efforts on detecting social biases in LLM output.

Fourth, the development and operation of LLMs requires significant energy consumption, which raises environmental concerns (Strubell, Ganesh, and McCallum 2020). Researchers should consider limiting their use of LLMs to cases in which efficiency gains are clear and justify an increased environmental footprint. Similarly, for larger projects, researchers should evaluate the benefits of carbon-offsetting strategies.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit <http://doi.org/10.1017/S1049096525000046>.

DATA AVAILABILITY STATEMENT

Research documentation and data that support the findings of this study are openly available at the Harvard Dataverse at <https://doi.org/10.7910/DVN/7KJLH7>.

CONFLICTS OF INTERESTS

The authors declare that there are no ethical issues or conflicts of interest in this research. ■

NOTES

1. The terms “LLMs” and “GPT” are related but refer to different concepts in the field of natural-language processing and AI. LLMs are a category of natural-language models trained on massive datasets containing diverse language data, and they are designed to produce coherent human-like text by predicting the next word in a sequence. The GPT family of LLMs refers to a specific set of models developed by the company OpenAI. These models were among the first to combine unprecedentedly large training sets (e.g., more than 570gb of text for GPT-3) with the transformer architecture—a neural network design that allows the model to give varying amounts of attention to different parts of the text. This allows it to more effectively understand context and relationships within the text. In summary, GPT is a specific implementation of the broader category of LLMs.
2. Box 262, Security Classified Reports and Memorandums Concerning Race Relations in the United States and Overseas, August 1944–January 1946. Records of the Office of the Secretary of War, Record Group 107. Washington, DC: National Archives Building.
3. Most errors regarding the main actor and target identification are due to contextual factors. GPT has difficulty identifying the target when it is not explicitly mentioned, such as in nonviolent cases, or when the main actor does not have an explicit target (e.g., National Association for the Advancement of Colored People meetings). It also has difficulty identifying the main actor in indirect descriptions (e.g., news articles).
4. Note that this success rate is based on rerunning the same prompt if GPT did not retrieve the complete information from the meeting minutes in the first trial. In R, this process can be automated by checking for missing information in the CSV columns.
5. Subsection D.1 in online appendix D provides additional information on this process and how to automatize it.

6. Type M errors are adapted from Gelman and Carlin (2014) and refer to differences in magnitude between the ground truth and the collected information. For example, the number of children may be coded as 4 when the real number is 2.
7. For more details on the differences between the models, see Korinek (2023).

REFERENCES

- Argyle, Lisa, Christopher Bail, Ethan Busby, and David Wingate. 2023a. “Leveraging AI for Democratic Discourse: Chat Interventions Can Improve Online Political Conversations at Scale.” *Proceedings of the National Academy of Sciences* 120 (41): e2311627120.
- Argyle, Lisa, Ethan Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023b. “Out of One, Many: Using Language Models to Simulate Human Samples.” *Political Analysis* 31 (3): 337–51.
- Bisbee, James, Joshua Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer Larson. 2024. “Synthetic Replacements for Human Survey Data? The Perils of Large Language Models.” *Political Analysis* 32 (4): 401–16. DOI:10.1017/pan.2024.5.
- Chiu, Ke-Li, Annie Collins, and Rohan Alexander. 2022. *Detecting Hate Speech with GPT-3*. <https://arxiv.org/abs/2103.12407>.
- Ekin, Sabit. 2023. *Prompt Engineering for ChatGPT: A Quick Guide to Techniques, Tips, and Best Practices*. <https://doi.org/10.36227/techrxiv.22683919.v2>.
- Feinstein, Brian D., and Daniel J. Hemel. 2020. “Outside Advisers and Inside Agencies.” *Georgetown Law Journal* 108 (5): 1139–211.
- Gelman, Andrew, and John Carlin. 2014. “Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors.” *Perspectives on Psychological Science* 9 (6): 641–51.
- Hida, Rem, Masahiro Kaneko, and Naoaki Okazaki. 2024. *Social Bias Evaluation for Large Language Models Requires Prompt Variations*. Accessed July 25, 2024. <https://doi.org/10.48550/arXiv.2407.03129>.
- Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. “Survey of Hallucination in Natural Language Generation.” *ACM Computing Surveys* 55 (12): 1–38.
- Korinek, Anton. 2023. “Generative AI for Economic Research: Use Cases and Implications for Economists.” *Journal of Economic Literature* 61 (4): 1281–317.
- Krcmaric, Daniel, Stephen C. Nelson, and Andrew Roberts. 2020. “Studying Leaders and Elites: The Personal Biography Approach.” *Annual Review of Political Science* 23 (May): 133–51.
- Montano, Beatrice, Simone Paci, and Chiara Superti. 2024. “The Mayor’s Daughter: How Patriarchal Norms Moderate the Daughter’s Effect on Mayoral Policy Making.” www.simonepaci.com/featured-project.
- Spirling, Arthur. 2023. “Why Open-Source Generative AI Models Are an Ethical Way Forward for Science.” *Nature* 616 (April): 413.
- Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2020. “Energy and Policy Considerations for Modern Deep Learning Research.” In *Proceedings of the AAAI Conference on Artificial Intelligence* 34:13693–96.
- Velez, Yamil Ricardo, and Patrick Liu. 2024. “Confronting Core Issues: A Critical Assessment of Attitude Polarization Using Tailored Experiments.” *American Political Science Review* 1–18. DOI:10.1017/S0003055424000819.
- Wang, Yu. 2023. “Topic Classification for Political Texts with Pretrained Language Models.” *Political Analysis* 31 (4): 662–68.
- Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, Denny Zhou, et al. 2022. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” *Advances in Neural Information Processing Systems* 35:24824–37.
- Wei, Xiang, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, et al. 2023. *Zero-Shot Information Extraction via Chatting with ChatGPT*. <https://doi.org/10.48550/arXiv.2302.10205>.
- Wu, Xiaodong, Ran Duan, and Jianbing Ni. 2024. “Unveiling Security, Privacy, and Ethical Concerns of ChatGPT.” *Journal of Information and Intelligence* 2 (2): 102–15.
- Yao, Yifan, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. “A Survey on Large Language Model (LLM) Security and Privacy: The Good, the Bad, and the Ugly.” *High-Confidence Computing* 4 (2): 100211.
- You, Hye Young, Kyuwon Lee, Simone Paci, Jeongmin Park, Sylvan Zheng. 2025. “Replication Data for “Applications of GPT in Political Science Research: Extracting Information from Unstructured Text.”” *PS: Political Science & Politics*. <https://doi.org/10.7910/DVN/7KJLH7>.
- Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. “Can Large Language Models Transform Computational Social Science?” *Computational Linguistics* (March): 1–55.