Applications of GPT in Political Science Research

Kyuwon Lee[*], Simone Paci[†], Jeongmin Park[‡], Hye Young You[§], Sylvan Zheng[¶]

Abstract

This paper explores the transformative role of GPT in political science research, demonstrating its potential to streamline data collection and analysis processes. By automating the extraction of information from diverse data sources—such as historical documents, meeting minutes, news articles, and unstructured digital content—GPT significantly reduces the time and financial resources traditionally required for data management. We explore how GPT's capabilities complement the work of human research assistants, combining automated efficiency with human oversight to enhance both the reliability and depth of research outputs. The integration of GPT not only makes comprehensive data collection and analysis accessible to researchers with limited resources, it also enhances the overall efficiency and scope of research in political science. This article underscores the increasing importance of artificial intelligence tools in advancing empirical research within the field.

[*]Department of Politics, Princeton University, kyuwon.lee@princeton.edu
[†]Department of Politics, Princeton University, sp0954@princeton.edu
[‡]Department of Politics, Princeton University, jp49@princeton.edu
[§]Department of Politics and Public Affairs, Princeton University, hyou@princeton.edu
[¶]Department of Politics, New York University, saz310@nyu.edu

Introduction

In the expanding landscape of political science research, the integration of advanced artificial intelligence 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 received considerable attention as a potential tool for increasing research productivity and expanding the scope of empirical research (Ziems et al. 2024). This article explores the diverse applications of GPT in political science research, focusing on its ability to improve the efficiency and accuracy of data collection processes that traditionally require extensive manual handling.

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. 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. Moving on to more complex applications, 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 final example, we show GPT's ability to perform an advanced task of synthesizing data from multiple internet sources.

Each of these applications demonstrates how GPT can not only perform labor-intensive tasks 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 handle large volumes of data, a capability that is particularly useful in political science where researchers are often faced with extensive but not fully structured datasets. The examples we present in this article highlight GPT's strengths in natural language processing while mitigating its weaknesses in complex reasoning and hallucination (i.e., false information).

GPT's potential to reduce the gap in unequal research resources is another significant benefit of integrating it into the political science toolbox. Traditionally, large-scale research projects often have been the purview of well-funded researchers who 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. Using GPT for various data collection tasks improves efficiency but it may limit student positions as RAs, which are critical for their financial support and professional growth. We recommend retaining student roles that require critical thinking, such as validation and complex data management, which complement the GPT-assisted task of data collection.

Applications

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. While OCR technology has advanced, the quality of OCR output remains dependent on the scanned image quality and the choice of OCR tool, which often results in errors, such as misspellings and odd spacing. High-quality OCR tools, like Google Cloud Vision, produce cleaner text but are often impractical due to issues such as document accessibility or resource constraints. To address these challenges, we use GPT to clean texts produced with the open-source OCR tool, Tesseract.

We use records from the Security Classified Reports and Memorandums Concerning Race Relations in the United States and Overseas, August 1944–January 1946 series, retrieved from the National Archives and Records Administration. This series, consisting of five boxes, is a collection of race-related incidents involving military personnel, media, communist organizations, politicians, labor unions and the NAACP in the United States against the backdrop of the Second World War. These reports (sample image in Figure A1) include important details about these incidents, such as the date and place, the people involved, and the measures taken by key actors. Unfortunately, the available OCR tools demonstrated mixed levels of accuracy (Table A1).

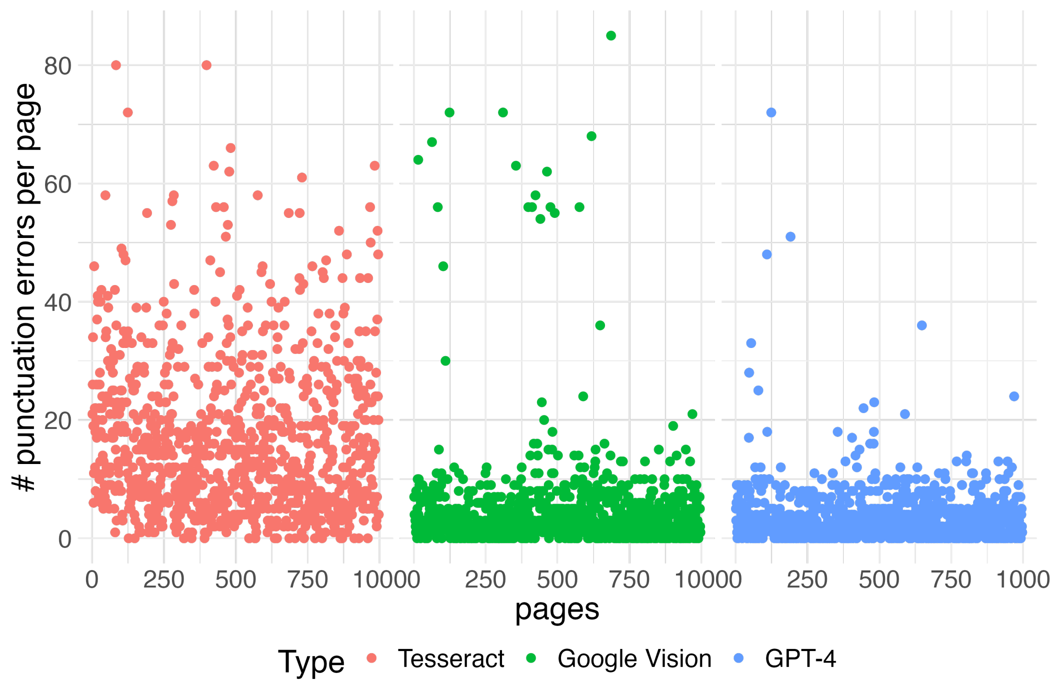
Here, we propose a time-saving approach that combines open-source tools (Tesseract) with GPT. We take the noisy text generated by Tesseract and use the GPT API to clean the noise, a process illustrated in Table A2. The result is almost exactly identical to the original text. We then visualize the effectiveness of this process. When comparing the performance of OCR tools, we consider two categories of noise: punctuation errors[1] and typos.[2] We then repeat the process for the entire document (997 pages) or one full box out of the five boxes in the series. and generate a count of noise per page. The results, presented in Figure 1, illustrate the effectiveness of GPT in removing noise; it removes most of the misplaced punctuation (Figure 1a) and typos (Figure 1b), yielding a cleaned text that is on par with Google Vision's quality. Summary statistics are provided in Figure A2.

We also demonstrate the use of GPT in extracting and summarizing information from this data. To demonstrate the efficiency of GPT, we first aggregate all strings from the 997 images to create individual race-related incidents between 1944 and 1945—a total of 7,638 cases. We then use GPT to extract critical details from each incident, including the location, main actors involved, and the targets. Finally, we extract a 10% sample from the cases, and manually check the accuracy of the information extracted by GPT, which is presented in Figure A3. Taken together, the evidence indicates the effectiveness of GPT in cleaning and analyzing historical documents.

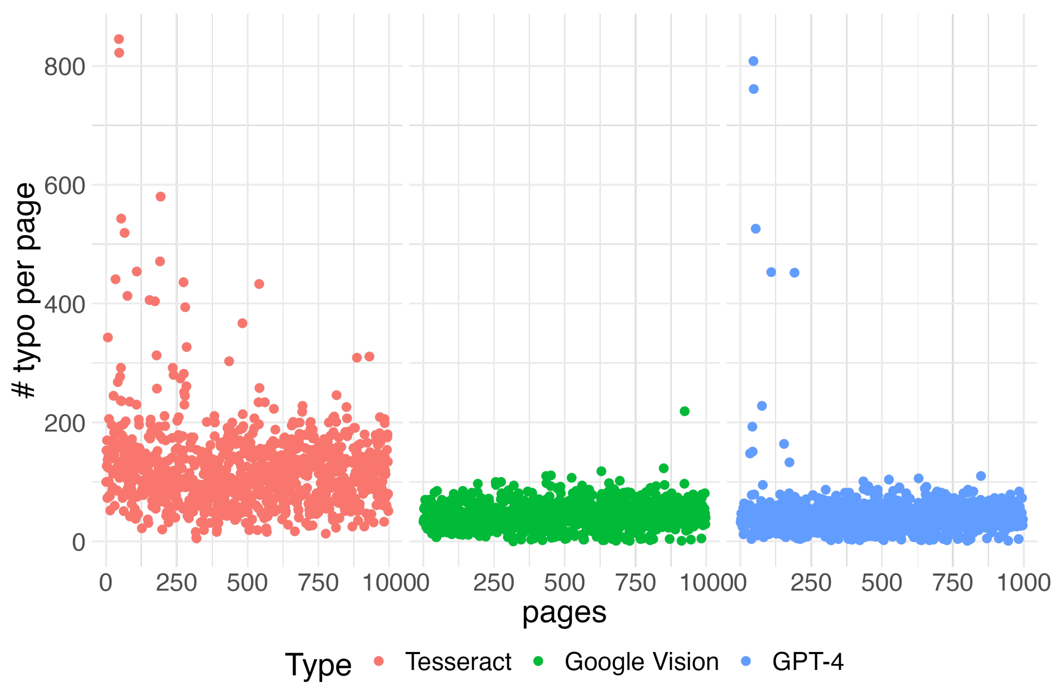
[1]We define punctuation error as (1) any occurrence of one or more spaces followed immediately by any common punctuation mark, or (2) any common punctuation mark that is not followed by a space or the end of a string.

[2]To identify typos, we use the `hunspell` package in `R`, a spell-checking library.

Figure 1: Comparison of Tesseract-processed Original Texts and GPT-cleaned Texts



(a) Number of punctuation errors per page



(b) Number of typos per page

Example 2: Extracting Unstructured Administrative Data

In this section, we show how GPT can be used to collect and clean administrative data provided in a semi-structured format (often in PDFs). We focus on meeting minutes from federal advisory committees (FACs) within federal agencies. A significant number of FACs serve as independent advisors that make policy recommendations to federal agencies. These committees hold over 5,000 public meetings annually, bringing together committee members, federal agency officials, and interest groups to discuss agency policy. Figure 2 shows two examples of committee meeting minutes, one from the Environmental Protection Agency (EPA) and one from the Centers for Disease Control and Prevention (CDC). Each contains the names, positions (e.g., chair, members, agency staff, or public attendees), and affiliations (e.g., Karmanos Cancer Institute) of meeting participants.

Figure 2: Examples of Advisory Committee Meeting Minutes

(a) EPA Meeting Minute

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.

(b) CDC Meeting Minute

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

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

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

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: chisml@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

We use GPT API and R to extract the name, affiliation, and position of each meeting participant from the FAC meeting minutes documents. Table B1 shows the API prompt and R command we use.

Extracting participants' position labels from meeting minutes is particularly challenging because the labels are so diverse and wide ranging. To address this, researchers can include the extensive set of position labels that appear in meeting minutes in the prompt. However, we also found that simply adding "etc" at the end of a list of example positions also addresses the issue, which is reflected in Table B1. After GPT created datasets from the meeting minutes, undergraduate RAs validated each dataset to ensure that all meeting attendee information was included.

Our example shows that the data collection and cleaning process for FAC meetings still requires human validation. However, having research assistants go through the GPT-generated data is much less resource-intensive and time-consuming than hiring research assistants to build data based on meeting minutes. If the minutes from one meeting includes 50,000 characters (5-6 pages), it costs 30 cents to run the GPT code on the transcript.

Example 3: Extracting Primary Sources from News Articles

In this section, we detail our approach of using GPT to extract semi-structured data from the extensive, unstructured texts of news articles, focusing on identifying the varied sources journalists employ. Newspaper articles typically reference a diverse array of sources—ranging from politicians and bureaucrats to private citizens and activists—which significantly influences the information conveyed to the public(Shapiro 2016). Although we focus on newspapers our approach could be used for similar tasks, such as extracting witness information from court records or guest appearances in news transcripts.

Identifying sources is particularly challenging because of the length of the input documents and the nuanced integration of source information within the article text, including variations in name and context. Despite the sophistication of GPT-4, this task requires too many reasoning steps to be solvable in a single prompt Wei et al. 2023. We divide the task into subtasks and solve each subtask separately with its own GPT prompt, feeding the output of one subtask prompt directly into the next. This way we avoid context window limitations and make the logic of each subtask explicit, which also makes debugging easier.

The details of the method are shown in Figure 3. First, we identify all quotes and information attributed to third parties in the news article. Second, we aggregate the quotes and information to the speaker or organization level, leveraging GPT's summarization capabilities. Finally, we transform the data into structured JSON, which can be processed with any data tool of choice. The full set of prompts and sample output can be found in the Appendix.

To validate our approach, we used the described method to extract 214 sources for 50 articles and hired crowd workers to identify errors in the extracted sources. We identify three types of errors: minor details (i.e., incorrect title, name, or organization); false sources (type I) where the extracted source was not cited in the article; and missing sources (type II) where a source present in the article was not extracted. We manually review each error identified by the crowd workers and estimate overall error rates. Our results show that the GPT-based system is highly accurate in extracting source details and rarely makes type I or type II errors (all error rates are less than 5%). Figure 4 shows error rates with 95% confidence intervals. Further details on crowd worker sourcing and screening are available in the Appendix.

We use this set of prompts to extract 31,431 sources from 5,795 New York Times articles about climate change over the period of 2012-2022 using the 'GPT-4 Turbo' model. Appendix Figure C1 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

In this section, we leverage GPT to extract specific information from an unstructured corpus of sources obtained through systematic Google searches. This exercise reflects a broad cate-

Figure 3: Source extraction process outline

News Article Text

Constance Millar ascended the trail to Telescope Peak — the highest point in Death Valley National Park — and discovered hundreds of dead and dying bristlecones. **"It's like coming across a murder scene," said Millar, an emerita research ecologist with the U.S. Forest Service who has studied bristlecone pines for the better part of 40 years.**

...

A new study, published Wednesday in the journal Nature, found that climate change has pushed almost a quarter of Earth's best-protected forests to a "critical threshold" for lost resilience — the point at which even a minor drought or heat wave could tip them into catastrophic decline. **Giovanni Forzieri, a professor of sustainable development and climate change at the University of Florence and lead author of the Nature study warned it could be a strong signal. ...**

↓

Extract information from sources

↓

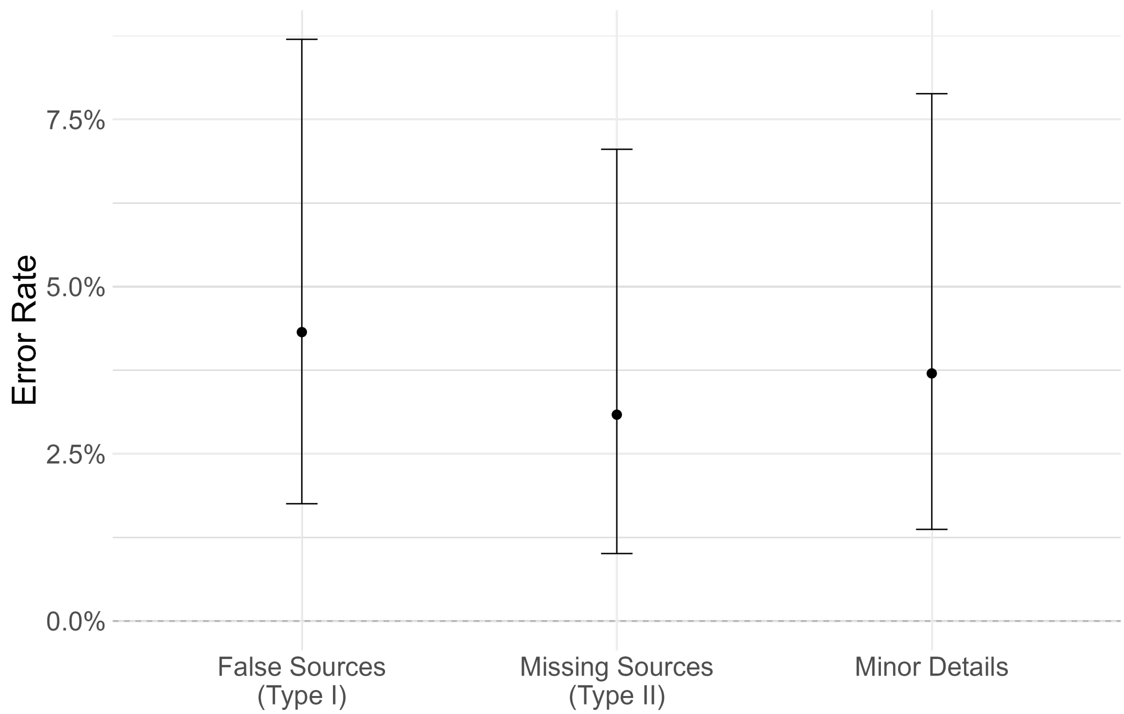
Aggregate information about sources

↓

Output JSON

```
{
  "person_name": "Constance Millar",
  "person_title": "Emerita Research Ecologist",
  "organization": "U.S. Forest Service",
  "document": "N/A",
  "summary": "Provided insights on the condition of bristlecone
     pines in Death Valley National Park, expressing shock at the
     scene of dead and dying trees."
},
{
  "person_name": "Giovanni Forzieri",
  "person_title": "Professor",
  "organization": "University of Florence",
  "document": "Nature study",
  "summary": "Discussed the implications of a study on the
     vulnerability of forests to climate change, highlighting the
     importance of the findings for societal awareness
     and climate mitigation strategies."
}
...
```

Figure 4: Performance of GPT-based source extraction



gory of data collection tasks where researchers cannot rely on a specific set of source material or a corpus of structured text. In these cases, data collection involves both searching for sources and extracting the relevant information. As a result, data collection draws from a variety of sources, such as websites, news articles, and academic and expert texts.

We replicate a large human-coded data collection effort by Montano, Paci, and Superti (2024), which examines whether having a daughter influences the pro-women policies of Italian mayors. The original study reflects a growing interest in political science in the role of elite biographical characteristics (Krcmaric, Nelson, and Roberts 2020). However, this approach faces a significant data availability problem, as systematic biographical data are rarely 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 research assistants from July 2023 to February 2024. For each mayor, the RAs reviewed up to the first 20 available search results for a total of over 7,300 Italian web pages.[3] 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). This task tests GPT-4's ability to parse through ambiguous and heterogeneous data. Most sources (about 90%) do not contain relevant information. The relevant information is encoded in myriad ways and the nuance of textual clues can be misleading. Table 1 shows illustrative examples of GPT-4 output. In three cases, GPT-4 correctly recovered the source information. The fourth case is an example of an error where the information is encoded in a complex way. The text mentions the mayor's "only son" and his two daughters. GPT-4 understood this as the mayor having three children while in truth the two are the mayor's son's daughters and thus the mayor's granddaughters, not to be counted as his direct offspring.

Table 1: Illustrative 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:** **Number of Kids:** **Number of Daughters:** **Confidence:**	1 1 1 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:** **Number of Kids:** **Number of Daughters:** **Confidence:**	1 1 NA 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:** **Number of Kids:** **Number of Daughters:** **Confidence:**	NA NA NA 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:** **Number of Kids:** **Number of Daughters:** **Confidence:**	1 3 2 90

Given the same set of search result links, we estimate the error rate of human coders and of GPT-4. We consider as ground truth all cases where human coders and GPT-4 agree. For

[3]In subsection D.1 in the Appendix, we provide additional information on this process and how to automatize it.

all disagreements, we adjudicated between the two sets by a third round of human coding, with the help of new RAs. For cases where all three rounds disagree, which were only 7 out of the total sample, the authors personally hand-coded the ground truth.

Figure 5: Human Coders and GPT-4 Coding Error Rates

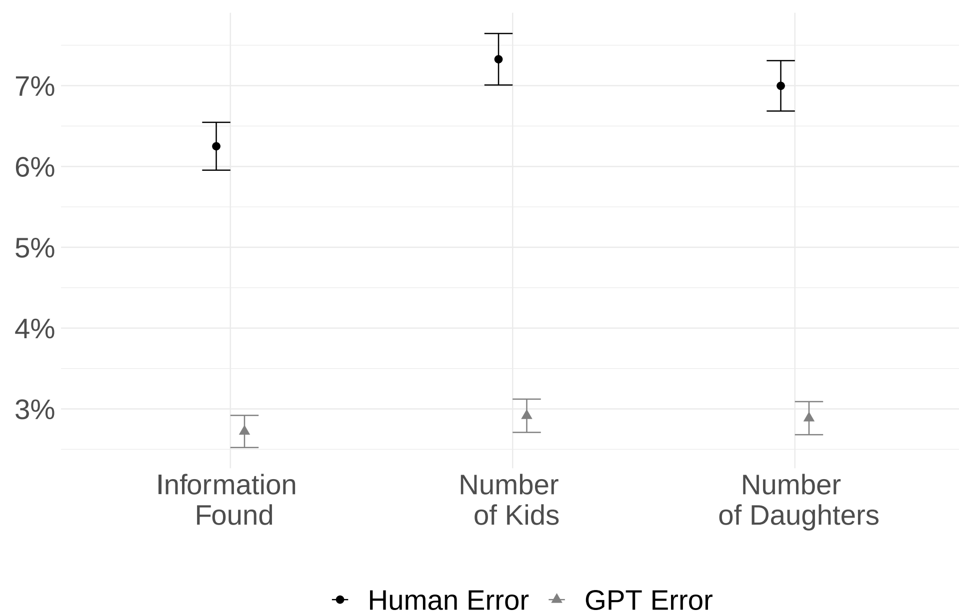


Figure 5 illustrates the error rates against the ground truth by the original group of human coders and GPT-4. Across the three main pieces of relevant information, GPT-4 outperformed human coders. Figure D1 in the Appendix sorts the overall error rate into different categories of mistakes: Type 1 (false positives), Type 2 (false negatives), and Type M (magnitude).[4] Compared to human coders, GPT-4 makes fewer Type 1 errors and more Type 2 errors. On the one hand, this pattern is reassuring, since 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 leave some information on the table, probably whenever it is

[4]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.

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 the Appendix. Whenever GPT-4 expressed a confidence rating below 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 only be taken as a noisy indicator of potential error.

Limitations and Best Practices

Our four applications focus on data collection, cleaning, and extraction tasks that are arduous, yet widespread in quantitative political science research. These types of tasks permit a straightforward application of LLM-based methods while minimizing their main weaknesses: reasoning and hallucinations. GPT can have difficulty completing reasoning tasks with multiple steps, such as solving arithmetic word problems or synthesizing sophisticated clues about a speaker's policy attitudes. Meanwhile, hallucinations are increasingly prevalent when asking GPT for information embedded in its training data as opposed to a specific input document.[5] Application of LLMs to more sophisticated tasks including sentiment analysis, topic labeling, stance detection, text annotation, information retrieval, and ideological scaling is an ongoing field of development, with researchers leveraging advanced prompting strategies, fine-tuning, and sophisticated post-processing to overcome these limitations (Wei et al. 2022; Wu et al. 2023a; Wu et al. 2023b; Argyle et al. 2023). Unfortunately, a full review of these cutting-edge techniques is out of scope of this article. We advise practition-

[5]For example, asking GPT-3.5-Turbo for relevant research papers on the intersection of artificial intelligence and political behavior yields a list of citations that sound plausible but are completely fictional, such as "The Politics of Artificial Intelligence" by David Runciman, *Political Quarterly* (2018).

ers interested in developing such applications to exercise optimistic caution and to always validate against a source of ground truth data.

Within the scope of data cleaning and collection, we additionally highlight important limitations and best practices. These recommendations integrate our experience and evidence from our validation exercises in this study along with advice concerning emerging best practices for LLM use and prompt engineering (Ekin 2023; Marvin et al. 2024).

First, LLM performance is highly sensitive to the specific prompt used. The term "prompt engineering" has emerged to describe the process of tailoring the LLM prompt for the task at hand. We encourage practitioners to use a hand-labeled dataset to systematically evaluate prompt performance. In our experience, the best performing prompts included the task context, the main objective, and a detailed specification of the output format and description of each data field. Prompts may also include examples of common information encoding patterns or even be computationally constructed to incorporate document-specific context. For complex tasks, we encourage researchers to explore multi-step prompting as demonstrated in Example 3 or asking the model to explain its reasoning before providing data as recommended by Wei et al. 2022). A full account of prompt engineering tactics can be found in Ekin 2023 and Marvin et al. 2024.

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 (Kamradt 2024). Appendix Figure D3, shows that GPT makes more errors in identifying the mayor's children as the length of the input text increases. 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 breaking tasks into smaller pieces.

Third, GPT occasionally exhibits "laziness" and does not follow the task instructions. This can manifest as incomplete responses or incorrect column or data label names. While prompt engineering can help mitigate these issues, we have found that simply re-running the same prompt multiple times until output is well formed is sufficient for most cases. Similarly, researchers can choose to double code information and re-prompt inconsistent responses. For instance, in example 4, we check that the number of children (of both genders) is 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 variance model uses when sampling each token. Lower values reduce the likelihood of hallucinations but increase sensitivity to prompt wording and reduce reasoning ability. We employed temperature values between zero and one in all of our examples. A full exploration of the relationship between temperature and performance is beyond the scope of this article.

Finally, 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 ease output data management, we recommend directing the model to limit output to JSON or CSV/TSV format ("provide only the table and nothing else").

Ethical Considerations

Relying on LLMs outsources tasks traditionally performed by student research assistants. While this improves the efficiency and cost-effectiveness of data collection, it undermines student employment opportunities. These opportunities not only provide students with financial support, they also provide valuable research experience, insight into academic work, and potentially influence some to pursue graduate studies. Research assistantships strengthen student résumés and also provide an important pedagogical opportunity for experiential learning.

We encourage researchers to continue the practice of hiring promising students as research assistants. The use of LLMs does not completely eliminate the need for RAs, as validation still requires thorough human coding. Outsourcing menial 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 or more complex data management.

References

Argyle, Lisa P, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. "Out of one, many: Using language models to simulate human samples." *Political Analysis* 31 (3): 337–351.

Ekin, Sabit. 2023. *Prompt Engineering For ChatGPT: A Quick Guide To Techniques, Tips, And Best Practices,* May. Accessed April 24, 2024. https://doi.org/10.36227/techrxiv.22683919.v2.

Gelman, Andrew, and John Carlin. 2014. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9, no. 6 (November): 641–651. ISSN: 1745-6916, accessed January 9, 2020. https://doi.org/10.1177/1745691614551642.

Kamradt, Greg. 2024. *Pressure Testing GPT-4-128K With Long Context Recall.* [Online; accessed 17. Apr. 2024], April. https://twitter.com/GregKamradt/status/1722386725635580292/photo/1.

Krcmaric, Daniel, Stephen C. Nelson, and Andrew Roberts. 2020. "Studying Leaders and Elites: The Personal Biography Approach." *Annual Review of Political Science* 23, no. Volume 23, 2020 (May): 133–151. ISSN: 1094-2939, 1545-1577, accessed April 17, 2024. https://doi.org/10.1146/annurev-polisci-050718-032801.

Marvin, Ggaliwango, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. 2024. "Prompt Engineering in Large Language Models." In *Data Intelligence and Cognitive Informatics,* edited by I. Jeena Jacob, Selwyn Piramuthu, and Przemyslaw Falkowski-Gilski, 387–402. Singapore: Springer Nature. ISBN: 978-981-9979-62-2. https://doi.org/10.1007/978-981-99-7962-2_30.

Montano, Beatrice, Simone Paci, and Chiara Superti. 2024. "The Mayor's Daughter: How Patriarchal Norms Moderate The Daughter's Effect on Mayoral Policy-Making."

Reed, Genna. 2021. "Administrator Regan Resets EPA Science Advice." *The Equation* (March 31, 2021). Accessed October 27, 2023. https://blog.ucsusa.org/genna-reed/administrator-regan-resets-epa-science-advice/.

Shapiro, Jesse M. 2016. "Special interests and the media: Theory and an application to climate change." *Journal of public economics* 144:91–108.

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* 35:24824–24837.

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,* arXiv:2302.10205, February. Accessed April 17, 2024. https://doi.org/10.48550/arXiv.2302.10205. arXiv: 2302.10205 [cs].

Wu, Patrick Y, Jonathan Nagler, Joshua A Tucker, and Solomon Messing. 2023a. "Concept-Guided Chain-of-Thought Prompting for Pairwise Comparison Scaling of Texts with Large Language Models." *arXiv preprint arXiv:2310.12049.*

Wu, Patrick Y, Joshua A Tucker, Jonathan Nagler, and Solomon Messing. 2023b. "Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting." *arXiv preprint arXiv:2303.12057.*

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. ISSN: 0891-2017, accessed April 17, 2024. https://doi.org/10.1162/coli_a_00502.

Online Supplementary Appendix

Table of Contents

A Example 1: Cleaning and Analyzing Historical Data

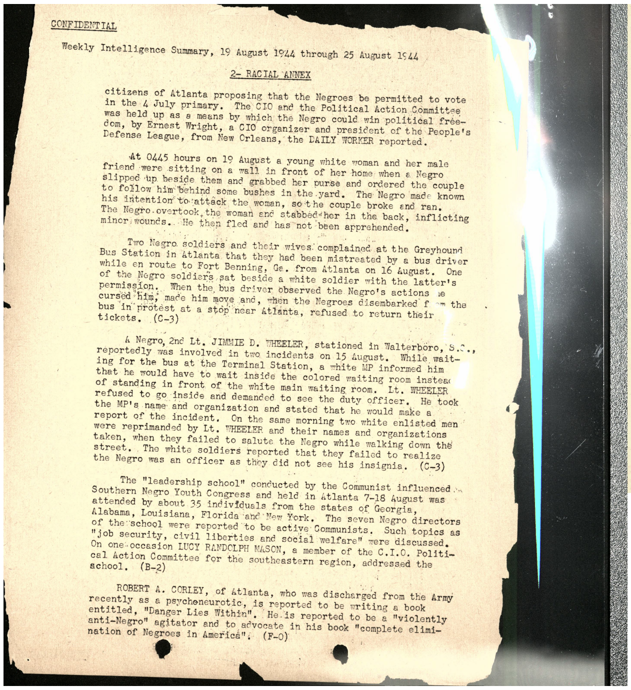
A.1 Original Image and Text

Original Text (human-typed)

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).[1]

Figure A1: Examples of Military Intelligence Reports from Box 262

(a) Example Image (Whole)



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 free-dom, 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.

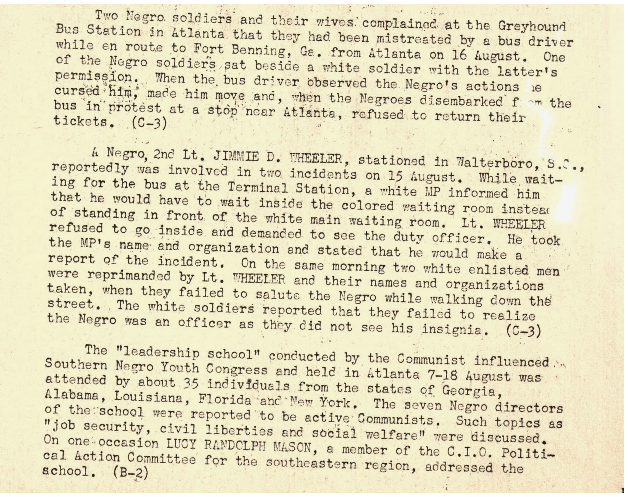
Two Negro soldiers and their wives complained at the Greyhound Bus Station in Atlanta that they had been mistreated by a bus driver while en route to Fort Benning, Ga. from Atlanta on 16 August. One of the Negro soldiers sat beside a white soldier with the latter's permission. When the bus driver observed 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 wait-ing 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. Politi-cal 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 elimi-nation of Negroes in America". (F-O)

(b) Example Image (Cropped)



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 wait-ing 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. Politi-cal Action Committee for the southeastern region, addressed the school. (B-2)

[1]Weekly Intelligence Summary, 19 August 1944 to 25 August 1944, box 262, folder 2, Security Classified Reports and Memorandums Concerning Race Relations in the United States and Overseas, August 1944–January 1946, Record Group 107; National Archives Building, Washington, DC.

A.2 Comparison of OCR Results

In this section, we compare the performance of GPT with other OCR tools, including the freely available Tesseract, which is also the most prominent and widely used solution; the subscription-based Adobe Acrobat; and the cloud-based, paid OCR service, Google Vision. We first provide an example of the mixed quality of these tools using a single case from Record Group 107 (Appendix A.1), typed by a human. The results (Table A1) highlight the different levels of accuracy between these tools: text recognized by Tesseract and Adobe Acrobat is noisy, with multiple misspellings and unusual characters. While Google Vision typically provides near-perfect text recognition, using it to process thousands of images and waiting for results can be a time-consuming and sometimes unnecessary step, especially if the text has already been processed, albeit imperfectly.

Table A1: Comparison of OCR Results by OCR Tools

OCR Tools	Result[*]
Tesseract	Two Negro**.** soldiers and their wives**?** complained**.** at the Greyhound **ay ae .** 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 **4ugust.** One **ee: Be }** of the Negro soldiers, sat beside a white soldier with the latter's [| **fo a cS** permission, When the bus driver observed the **. Negro'ts** actions **16 as Ls.** curséd him**}** made him move **.and, whén** 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)
Adobe Acrobat	Two Negro. **-oldiets:|** and their **wiv-os.' complainc:. at the Greyhounrl Bus Ste.tion ;in b.tlanta** that they had been mistreated **,**by a bus driver while en **rout.e: to Fort Benning, Ge . •** from Atlanta on **16 /1.ugust.** One of the **l'1e::gro soldiat; • ,pat** beside a white soldier with the latter's **permi;;ilon • . When the. bu · s dr:i.vGr bbserved** the Negro's actions **1e curs, :;(.ma · e him - }nd, • ,1 • }1.n the Negroe__s** disembarked **· m the bus 1n .** protest at a stop **· · · V.?near h.tla:nta,** refused to return their **!** .tickets **• .C-.3)**
Google Vision	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 **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)

[*] The errors are highlighted in bold text.

A.3 OCR Results with GPT

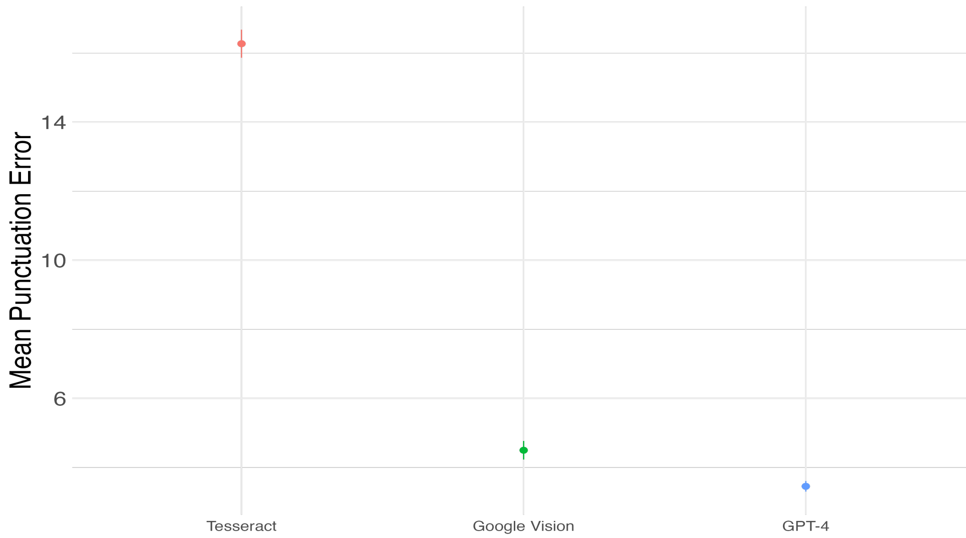
Table A2: OCR Results Using Tesseract and GPT

<table>
<tr><th>Description</th><th>Result</th></tr>
<tr><td>R Code</td><td><code>paste0(c("Hi GPT, can you please clean this text? 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[i])</code></td></tr>
<tr><td>OCR
(Tesseract)</td><td>Two Negro. soldiers and their wives? complained. at the Greyhound ay ae . 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 4ugust. One ee: Be } of the Negro soldiers, sat beside a white soldier with the latter’s [| fo a cS permission, When the bus driver observed the. Negro’ts actions 16 as Ls. curséd him} made him move .and, whén 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)</td></tr>
<tr><td>Cleaned with
GPT+</td><td>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)**.</td></tr>
</table>

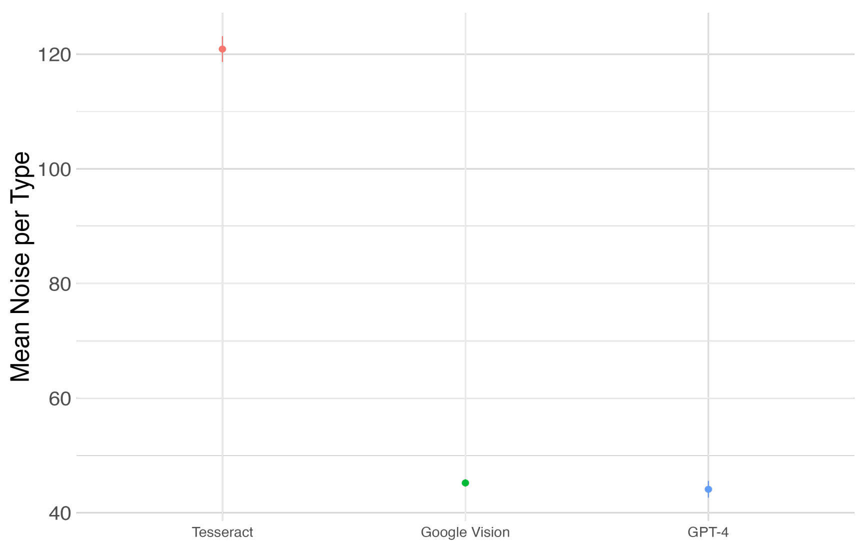
+ The corrections are highlighted in bold text.

A.4 Summary of OCR Results with GPT

Figure A2: Comparison of Tesseract-processed Original Texts and GPT-cleaned Texts



(a) Performance of GPT-based Text Cleaning (punctuation)



(b) Performance of GPT-based Text Cleaning (typo)

Table A3: GPT Prompt in R

Description	R Code
R Code	`paste0(c("This is a case of race-related incidents in WW2 US. Provide a csv table consisting of 3 columns. The column names should be: location, main actor, target. The columns are: (1) main actor: the person/organization who is doing the key action in the case (2) target: the person/organization to whom the action is targeted (3) location: the locality in which the case is happening:", OCR[i])`

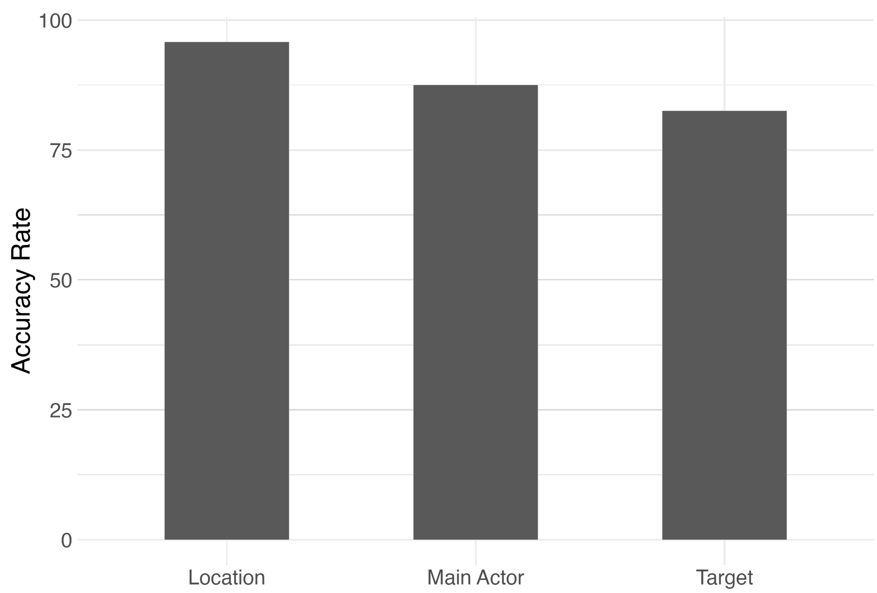


Figure A3: Accuracy Rate of Information Summarized Using GPT-4, on 10% Sample

B Example 2: Extracting Unstructured Administrative Data

B.1 R Code and ChatGPT Command

Table B1: GPT Prompt and API Command in R

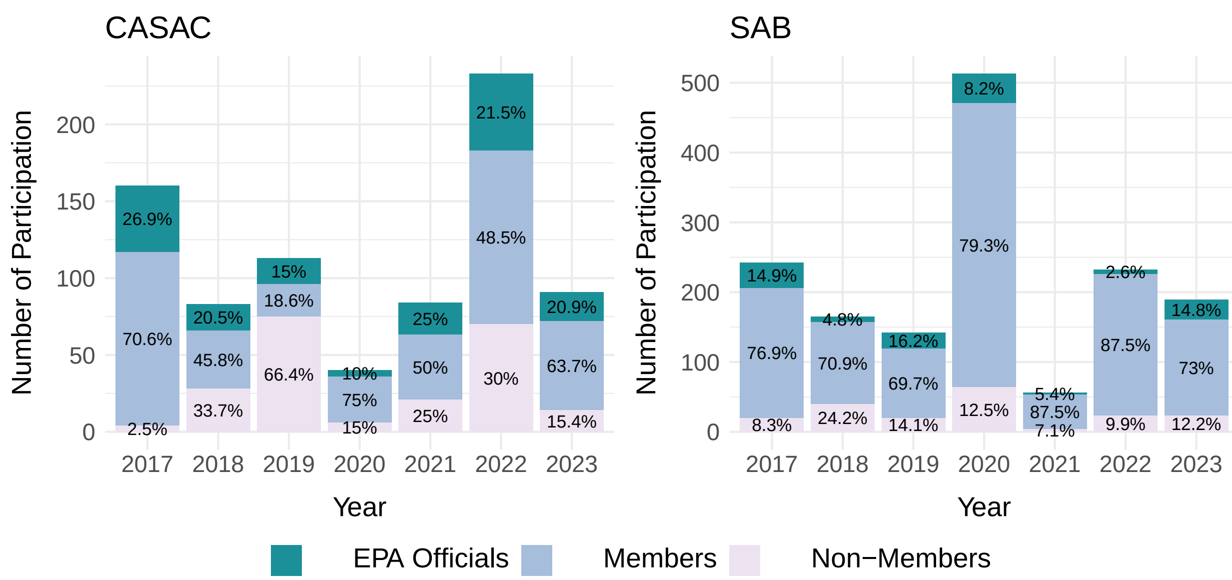
Description	Command in R
Prompt	`prompt=` 'Return me a csv delimiter table of three columns, "name," affiliation," and "position." Do not return anything else except for the table. The first column "name" has the names of meeting participants and people, if any, who made public comments. When writing 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$page[I])))])), collapse=" ")))))`
Extract API response	`capture.output(cat(content(response)$choices[[1]]$message$content))`

B.2 Analyses of FAC Meetings Based on Extracted Data

As an example, we focus on two FACs in the Environmental Protection Agency (EPA), which are the Clean Air Scientific Advisory Committee (CASAC) and Science Advisory Boards (SAB). CASAC provides independent advice to the EPA administrator on the technical basis for the EPA's National Ambient Air Quality Standards (NAAQS). To do so, the CASAC committee creates recommendation reports based on their reading of the EPA's drafts of Integrated Science Assessment (ISA) on NAAQS, EPA officials' presentations, and inputs from interest groups. SAB also reviews the quality and relevance of the scientific and technical information used by the EPA or proposed as the basis for agency regulations. The George W. Bush and Trump administrations attempted to fill many EPA advisory committees with their ideological allies, and the Biden administration reset EPA advisory committees in which the Trump administration intervened (Reed 2021).

Figure B1 shows the number and percentage of participation in FAC meetings. There are largely three types of participants: EPA agency officials who are career bureaucrats, members of FACs who are appointed by the EPA administrator, and interest groups that can voluntarily participate in FAC meetings and provide comments. If the same interest group participated in two FAC meetings in a given year, their participation counts as two in the statistics. The figure shows that the proportion of interest group participation in CASAC and SAB meetings is largest when there is ideological divergence between EPA political appointees and career officials. The reason there was low interest group participation in 2017 is that the Trump administration began disbanding and resetting some FACs in 2018. The Biden Administration disbanded these FACs in March 2021 to reset the committee membership. Thus, the data on FAC meetings provide an invaluable opportunity for researchers to examine how ideological disagreement between political leaders and bureaucrats affects interest groups' participation in bureaucratic policymaking.

Figure B1: Participation in CASAC and SAB Meetings, 2017-2023



C Example 3: Extracting Primary Sources from News Articles

C.1 Full Prompts

Table C1: Extracting Sources from News Articles with GPT: Identifying Quotes and Information (Step 1)

1: Quotes and Information
System: You are a research assistant whose task is to extract quotes and other external information used by journalists in news articles. The user will provide the text of a news article. Generate a numbered list of all quotes and external information attributed to specific people, organizations, or documents such as studies, reports, or press releases. Format your response as follows: 1. Quote or information - Name of the source - Background of the source - Context of the quote or information Some articles do not mention any information drawn from external sources. In these cases, simply say "No sources mentioned." **User:** [Original News Article Text] The trees had stood for more than 1,000 years. The rings of their trunks told the story of everything they'd witnessed. Weather patterns shifted; empires rose and fell. But here, in one of the harshest environments on the planet, the bristlecone pines survived. Until the day in 2018 when Constance Millar ascended the trail to Telescope Peak — the highest point in Death Valley National Park — and discovered hundreds of dead and dying bristlecones. "It's like coming across a murder scene," said Millar, an emerita research ecologist with the U.S. Forest Service who has studied bristlecone pines for the better part of 40 years. In a study published this spring, she and fellow researchers showed that the West's worst drought in at least 1,200 years had critically weakened the trees. Voracious bark beetles — a threat to which bristlecones were previously thought immune — delivered the death blow. Standing amid the devastation atop Telescope Peak, Millar realized, "This could be a harbinger of what could be happening in the future." Millar immediately contacted Barbara Bentz, a research entomologist for the Forest Service based at the Rocky Mountain Research Station in Logan, Utah. Just a few years earlier, Bentz had surveyed bristlecone populations across the West and found that few were succumbing to the bark beetle outbreaks that decimated so many other species. She felt confident that the trees would remain resilient even as rising temperatures caused beetle populations to explode. Yet, when Bentz peeled a piece of bark off one of the Death Valley bristlecones, she found scores of tiny tunnels created by beetle larvae as the insects chewed through the tree's living tissue. "Oh, no," Bentz thought to herself. "It really is beetles. Oh, no." She and her colleagues uncovered further evidence of insect damage in Utah's Wah Wah Mountains, where young bristlecones were attacked by a small brown beetle called a pinyon ips. A new study, published Wednesday in the journal Nature, found that climate change has pushed almost a quarter of Earth's best-protected forests to a "critical threshold" for lost resilience — the point at which even a minor drought or heat wave could tip them into catastrophic decline. "It's a strong warning, I think, for society," said Giovanni Forzieri, a professor of sustainable development and climate change at the University of Florence and lead author of the Nature study. He pointed out that most of the world's climate plans count on forests to pull planet-warming gases out of the atmosphere. If these ecosystems collapse, humanity will find it difficult, if not impossible, to stave off catastrophic warming. **Response:** 1. "It's like coming across a murder scene." - Constance Millar - Emerita research ecologist with the U.S. Forest Service, with over 40 years of experience studying bristlecone pines. - Context: Describing her reaction upon discovering the dead and dying bristlecone pines in Death Valley National Park. 2. "This could be a harbinger of what could be happening in the future." - Constance Millar - Emerita research ecologist with the U.S. Forest Service. - Context: Reflecting on the implications of the dying bristlecone pines for future forest health. 3. "Oh, no. It really is beetles. Oh, no." - Barbara Bentz - Research entomologist for the Forest Service based at the Rocky Mountain Research Station in Logan, Utah. - Context: Expressing her dismay upon finding evidence of beetle larvae damaging bristlecone pines in Death Valley. 4. "It's a strong warning, I think, for society." - Giovanni Forzieri - Professor of sustainable development and climate change at the University of Florence and lead author of the Nature study. - Context: Commenting on the findings of a study published in Nature regarding the vulnerability of the world's forests to climate change and the implications for climate plans that rely on forests to sequester carbon.

Table C2: Extracting Sources from News Articles with GPT: Source Aggregation (Step 2)

2: Source Aggregation
System: You are a research assistant whose task is to understand the sources of quotes and other external information used by journalists in writing news. The user will provide a list of quotes and external information referenced in a news article. Aggregate together the quotes and information drawn from the same sources into a new numbered list. Provide a description of the source and a summary of the source's contribution. Format your response as follows: 1. Name of the source - Description of the source - Summary of the source's contribution - Documents authored by the source, if relevant **User:** [List of quotes and information from output of Step 1] **Response:** 1. Constance Millar - Description of the source: Emerita research ecologist with the U.S. Forest Service, with over 40 years of experience studying bristlecone pines. - Summary of the source's contribution: Constance Millar provided insights on the condition of bristlecone pines in Death Valley National Park, expressing shock at the scene of dead and dying trees and suggesting that this situation may foreshadow future forest health challenges. 2. Barbara Bentz - Description of the source: Research entomologist for the Forest Service based at the Rocky Mountain Research Station in Logan, Utah. - Summary of the source's contribution: Barbara Bentz shared her reaction to discovering beetle larvae in bristlecone pines, indicating concern for the trees' health and the potential impact of pest infestations. 3. Giovanni Forzieri - Description of the source: Professor of sustainable development and climate change at the University of Florence and lead author of the Nature study. - Summary of the source's contribution: Giovanni Forzieri discussed the implications of a study on the vulnerability of forests to climate change, highlighting the importance of the findings for societal awareness and climate mitigation strategies. - Documents authored by the source: Lead author of the study published in Nature on the vulnerability of the world's forests to climate change.

Table C3: Extracting Sources from News Articles with GPT: Organization and JSON (Step 3)

3: Structured Data

System:
For each entry in the given list of sources, identify the person's name (if the source is a person), their title (such as President, Professor, Spokesman, etc.), the organization name, the document name (for sources that are written documents such as academic studies, press releases, books, etc.), and the category (already provided). If any of these are not relevant provide a value of N/A. For unnamed or anonymous sources provide a value of "Anonymous" for the person's name. Do not include titles in the person name field (eg, Barack Obama instead of President Obama). Finally, include the summary of the primary source's role in the article.
Format your response as JSON array with keys "person_name", "person_title", "organization", "document", "category" and "summary"
User: [List of aggregated sources from output of Step 2]
Response:


```
[
  {
    "person_name": "Constance Millar",
    "person_title": "Emerita Research Ecologist",
    "organization": "U.S. Forest Service",
    "document": "N/A",
    "summary": "Provided insights on the condition of bristlecone pines in
    Death Valley National Park, expressing shock at the scene of dead and dying trees."
  },
  {
    "person_name": "Barbara Bentz",
    "person_title": "Research Entomologist",
    "organization": "Forest Service Rocky Mountain Research Station",
    "document": "N/A",
    "summary": "Shared her reaction to discovering beetle larvae in bristlecone pines,
    indicating concern for the trees' health."
  },
  {
    "person_name": "Giovanni Forzieri",
    "person_title": "Professor",
    "organization": "University of Florence",
    "document": "Nature study",
    "summary": "Discussed the implications of a study on the vulnerability of forests
    to climate change, highlighting the importance of the findings for societal awareness
    and climate mitigation strategies."
  }
]
```

C.2 Validation Details

We collected corrections for 462 sources for 87 articles from 25 crowd workers hired on the Prolific platform with an undergraduate or graduate degree and English fluency. In the main text, we report data based on results that remove data from inattentive crowd workers, resulting in a smaller validation set of 241 sources. In order to determine crowd worker attentiveness, we asked each crowd worker to correct errors for a "test" article that contained manufactured errors known to us in advance. This test article was embedded in the task, making it impossible for the crowd worker to determine whether there was a test article or which article it was. "Inattentive" workers were those who did not catch these manufactured errors. Thus, we analyze corrections made by eight attentive crowd workers, who each read two articles and made corrections on a total of 214 sources extracted by the GPT-based system. Error rates are even lower if we include data from inattentive workers (ranging from 1.1% and 1.5% for the three error types).

C.3 Additional Figure

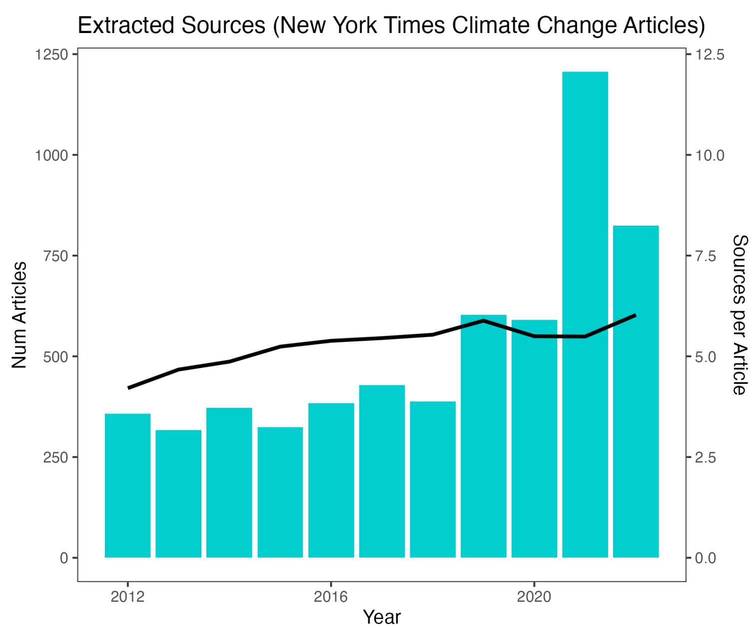


Figure C1: We extract 31,431 sources from 5,795 *New York Times* articles about climate change over the course of 2012-2022.

D Example 4: Extracting Information from Unstructured Sources

D.1 Google Search Algorithm

The first part of the data collection in Montano, Paci, and Superti (2024) relies on systematic Google searches. This procedure offers several key advantages. Relying on search engine results is highly flexible and can be applied to various data collection projects that rely on publicly available information. It can collect a wide range of sources, thus triangulating information retrieval for topics discussed across areas of the internet. Alternatively, search strategy can be refined to rely on specific search engines, such as Google News or Google Scholar, or to target only specific types of sources, such as institutional websites or blogs.

At the same time, this strategy has clear drawbacks. Source selection depends on the internal algorithm of the chosen search engine and opaque dynamics such as SEO (Search Engine Optimization) indexing. This may skew search results toward sources with greater resources and web visibility. In addition, the underlying information base is constantly changing as websites are shut down and search results change. As a result, the data collection process is difficult to replicate. However, researchers can ensure source traceability by maintaining a database of the underlying sources, such as text scraped from search result links.

This first step can also be automatized. Relying on the JSON Google Search API, we created an algorithm that iterates over predetermined combinations of search terms, collecting the top ten search result links for each search. We then scrape the resulting set of web pages for their text content. The full R code is available upon request.

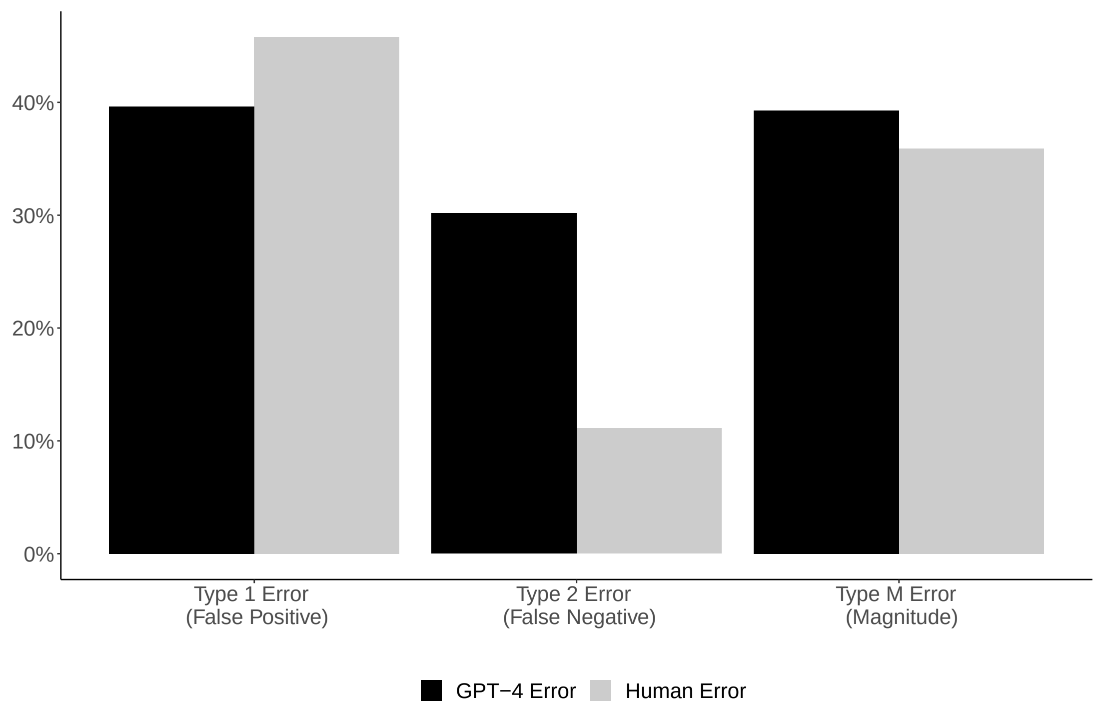
D.2 Full Prompts

Table D1: Full Prompts

Example 4 - Elite Biographical Information Extraction
In this task, you will read an Italian text scraped from Google results; it could be any webpage. Search the text for specific information, and return to me a csv delimiter table, using commas as separators, with the following columns: "info-found," "has-kids," "number-kids," "number-daughters," "source," "confidence." Do not return anything else except for the table. In the text I provide below, search information about whether [NAME], mayor of [TOWN], has kids, how many kids the mayor has, and how many of the kids are female. Be careful, the text may discuss the kids of other people and also mention the mayor. It is not enough to have the word "figli" and the mayor name. The text may also discuss other mayors of [TOWN]. Make sure that the text attributes the kids mentioned to mayor [NAME]. The text may also mention [NAME] as former mayor of [TOWN], in this case you can still consider the information as valid. The text may also connect the kids to the mayor indirectly, discussing the broader family of the mayor, but make sure that there is a clear link. Also, if the mayor is said to have a granddaughter or grandson, assume they must have at least one kid. In this case, unless it is mentioned which parent is the mayor's child, assume the child is male. The column "info-found" should have value 1 if you find any information on any of these queries. If you find no information about kids of the mayor, the column should have value 0. The column "has-kids" should have value 1 if you found information about the mayor having kids. Input 0 if there is information about the mayor not having kids. Otherwise, input NA. The column "number-kids" should have a numeric value for the number of kids, both male and female. If there is a specific number of kids, use that number. If the text specifies kid's names, use the number of names. Otherwise, input NA. The column "number-daughters" should have a numeric value for the number of daughters. In case you can find the names of the daughters, infer the gender from the names. In Italian, most names have a clear gender connotation. If there is a specific number of daughters, use that number. If the text specifies daughter names, use the number of names. Otherwise, input NA. The column "source" should have the sentences where you found information for any of the values imputed in the previous columns, pasted together with a semi-colon separating them. If you did not find info, input NA. The column "confidence" should contain a number from 0 to 100 indicating the degree of confidence in the information extraction for all previous columns. Give a rating both when you find and when you do not find information. Follow these guidelines to determine confidence: Clarity of Attribution: Increase confidence if the information about kids clearly refers to [NAME]. Decrease if other individuals could be the parents. Contextual Evidence: Increase confidence if multiple sources within the text confirm the information. Decrease if the information comes from ambiguous or unreliable parts of the text. Similar Names: Decrease confidence if there are multiple people with the same name or if the text discusses other mayors. Conflicting Information: If the text contains conflicting information regarding the kid, decrease confidence significantly. Use the whole scale, following these descriptions of different confidence ranges: 0-20: High ambiguity or multiple conflicting sources. 21-40: Moderate ambiguity, or some elements are unclear but likely correct. 41-60: Slightly unclear, minor contradictions or uncertainties present. 61-80: Mostly clear, with minor unresolved questions. 81-100: Clear and unambiguous information, strongly supported by the text. These ranges are just for reference, the confidence rating should have a number from 0 to 100, not a range. This is a really important task, and you will be rewarded if you do a great job. So put maximum effort into it, please! Use the following text to create the table: [Insert text scraped from web page]

D.3 Additional Results

Figure D1: Error Types



Error rates are calculated as the number of errors of a specific type over the number of sources checked.

Figure D2: GPT Confidence Rating and Coding Error

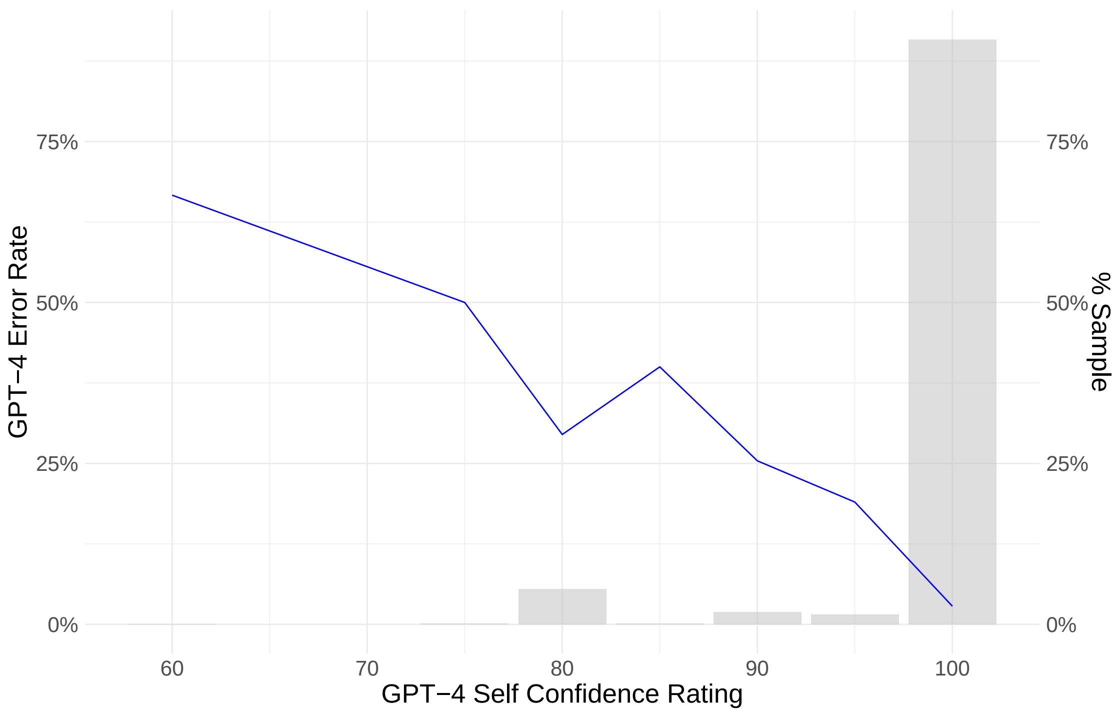


Figure D3: Source Text Length and Coding Error