Challenges in End-to-End Policy Extraction from Climate Action Plans

Nupoor Gandhi[1], **Tom Corringham**[2], **Emma Strubell**[1]
Carnegie Mellon University[1], University California San Diego[2]
{nmgandhi,estrubel}@cs.cmu.edu, tcorringham@ucsd.edu

Abstract

Gray policy literature such as climate action plans (CAPs) provide an information-rich resource with potential to inform analysis and decision-making. However, these corpora are currently underutilized due to the substantial manual effort and expertise required to sift through long and detailed documents. Automatically structuring relevant information using information extraction (IE) would be useful for assisting policy scientists in synthesizing vast gray policy corpora to identify relevant entities, concepts and themes. LLMs have demonstrated strong performance on IE tasks in the few-shot setting, but it is unclear whether these gains transfer to gray policy literature which differs significantly to traditional benchmark datasets in several aspects, such as format of information content, length of documents, and inconsistency of document structure. We perform a case study on end-to-end IE with California CAPs, inspecting the performance of state-of-the-art tools for: (1) extracting content from CAPs into structured markup segments; (2) few-shot IE with LLMs; and (3) the utility of extracted entities for downstream analyses. We identify challenges at several points of the end-to-end IE pipeline for CAPs, and we provide recommendations for open problems centered around representing rich non-textual elements, document structure, flexible annotation schemes, and global information. Tackling these challenges would make it possible to realize the potential of LLMs for IE with gray policy literature.

1 Introduction

Gray policy literature — non-commercial and non-academic documents which can include white papers, technical reports, and working papers — is an information-rich resource that is generally difficult to navigate due to the volume and diversity of format (Pandita and Singh, 2011; Lawrence et al., 2015; Turner et al., 2005). Paid for by public funds, these documents are usually freely available and often the most timely resource on policy issues (Rothstein and Hopewell, 2009). Lawrence et al. (2015) found that half of surveyed policymakers would be more likely to use gray policy literature if information were easier to find and access.

Information extraction (IE) tasks in the NLP space are designed to make it possible to efficiently sift through such information, but gray policy literature poses several challenges for traditional IE. They are distributed as long PDFs with inconsistent document structure, such that relevant sections cannot be easily automatically extracted. They are designed to be visually appealing with crucial information organized in rich non-textual elements such as tables and graphics (Turner et al., 2005). In contrast, the standard datasets that IE is designed to perform well on take the form of short, plain text documents from domains such as webtext or news articles (Riedel et al., 2010; Roth and Yih, 2004).

In general, IE models that are trained on these standard datasets can be adapted to new domains by finetuning with annotated examples, but this may not be feasible for gray policy documents (Gururangan et al., 2020). Collecting a large number of manually annotated examples for a static set of entities can be prohibitively expensive with gray policy literature due to fast-paced and diverse developments in the field. Over standard datasets, large language models (LLMs) have demonstrated strong performance in entity and relation extraction in the few-shot setting (Yuan et al., 2022; Wan et al., 2023; Wadhwa et al., 2023).

But, it is unclear to what extent LLMs can be used to extract information from gray policy literature in the few-shot setting. In this work, we present a case study of few-shot IE with LLMs over climate policy text. We specifically consider climate action plans (CAPs) from the state of California. CAPs are distributed as long PDFs, sharing many of the challenging properties of gray policy

Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024), pages 156–167
August 16, 2024 ©2024 Association for Computational Linguistics

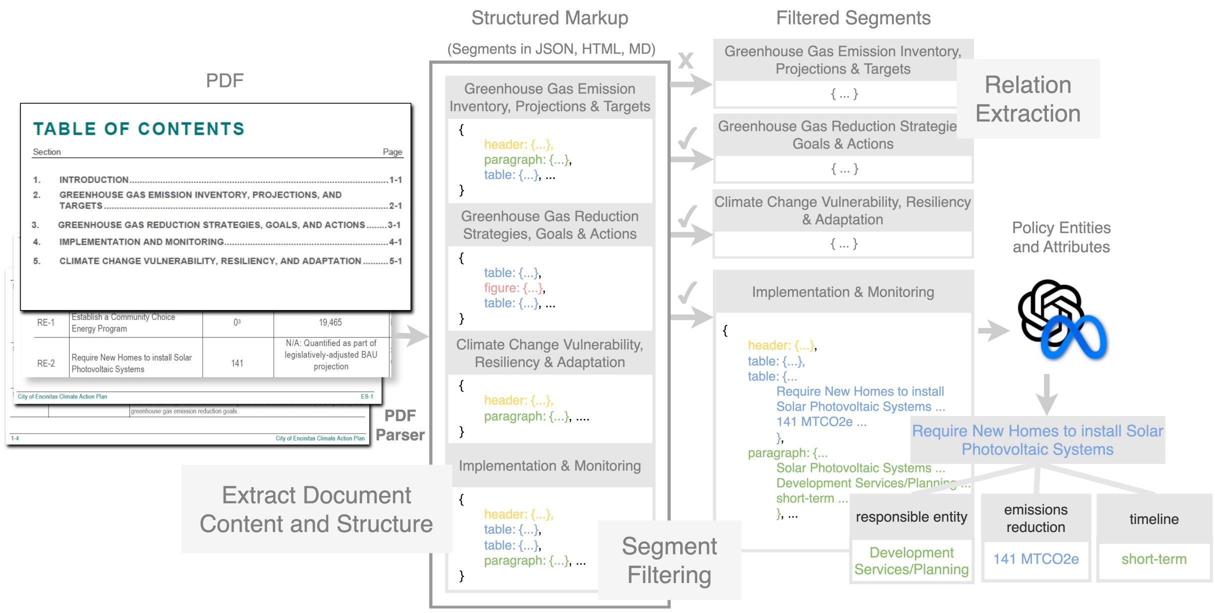
PDF
TABLE OF CONTENTS
Section
Page
1. INTRODUCTION 1-1
2. GREENHOUSE GAS EMISSION INVENTORY, PROJECTIONS, AND TARGETS 2-1
3. GREENHOUSE GAS REDUCTION STRATEGIES, GOALS, AND ACTIONS 3-1
4. IMPLEMENTATION AND MONITORING 4-1
5. CLIMATE CHANGE VULNERABILITY, RESILIENCY, AND ADAPTATION 5-1
RE-1
Establish a Community Choice Energy Program
19,465
RE-2
Require New Homes to install Solar Photovoltaic Systems
141
N/A: Quantified as part of legislatively-adjusted BAU projection
City of Encinitas Climate Action Plan
PDF Parser
Extract Document Content and Structure
Structured Markup
(Segments in JSON, HTML, MD)
Greenhouse Gas Emission Inventory, Projections & Targets
header: {...},
paragraph: {...},
table: {...}, ...
Greenhouse Gas Reduction Strategies, Goals & Actions
table: {...},
figure: {...},
table: {...}, ...
Climate Change Vulnerability, Resiliency & Adaptation
header: {...},
paragraph: {...}, ...
Implementation & Monitoring
header: {...},
table: {...},
table: {...},
paragraph: {...}, ...
Segment Filtering
Filtered Segments
Greenhouse Gas Emission Inventory, Projections & Targets
{ ... }
Greenhouse Gas Reduction Strategie Goals & Actions
{ ... }
Climate Change Vulnerability, Resiliency & Adaptation
{ ... }
Implementation & Monitoring
header: {...},
table: {...},
table: {...
Require New Homes to install Solar Photovoltaic Systems ...
141 MTCO2e ...
},
paragraph: {...
Solar Photovoltaic Systems ...
Development Services/Planning ...
short-term ...
}, ...
Relation Extraction
Policy Entities and Attributes
Require New Homes to install Solar Photovoltaic Systems
responsible entity
Development Services/Planning
emissions reduction
141 MTCO2e
timeline
short-term

Figure 1: Given a Climate Action Plan in PDF format, the end-to-end IE pipeline includes first extracting the content from PDFs into structured markup segments using extracted headers or the table of contents (§6.2). Then, the segments are filtered for policy relevance (§6.3). For each structured markup segment, we can perform few-shot RE with an LLM to extract policies (e.g. "Require New Homes to install Solar Photovoltaic Systems") and referring policy attributes (e.g. responsible entity, emissions reduction) (§6.4).

literature including information-dense non-textual elements and inconsistent structure. If possible, extracting rich, structured representations of policies with minimal annotation effort would be useful for many applications such as auditing emissions reductions, searching for relevant adaptation strategies for a specific climate hazards, or aggregating local government actions to the state or federal level.

We evaluate few-shot IE performance of LLMs for an *end-to-end* setup, where the input is a CAP in PDF form, and the output is a set of policy entities and relations (Figure 1). We inspect to what extent existing PDF parsers are able to preserve crucial policy information (*raw text recognition*), recognize textual and non-textual elements (*element recognition*), and extract a document structure that would be useful to a domain expert to segment the document (*structured segmentation*). Then, using the parsed segments of the CAP, we analyze how well entities can be extracted from information-dense non-textual elements (*intra-element extraction*), and more generally the relation extraction performance of LLMs in a few-shot setting (*segment-level extraction*). Finally, we experiment with modes of useful representations of the extracted entities and relations (*extraction utility*).

Based on our analysis over the CAPs, we provide recommendations for future directions in NLP that would improve IE with gray policy literature. We propose: (1) more flexible annotation schemes to account for inconsistencies in how entities are expressed across documents; (2) better representations of non-textual elements in the context; (3) methods that extract vague, imprecise, and subjective entity types pervasive in gray policy literature; and (4) the use of rich document structures.

2 Climate Policy Extraction

CAPs generally contain sections describing proposed policies and an inventory of emissions produced by the jurisdiction. Climate policies can be classified as adaptation to climate changes or emissions reduction measures. An example CAP for the city of Encinitas, California can be found here.

California municipal and county CAPs exhibit significant variability in structure but typically contain a similar set of sections. These include front matter, a discussion of the regulatory framework, a description of local and regional climate projections, and, critically for our purposes, chapters on specific and detailed mitigation policies, associated emissions reductions, and sometimes also chapters on implementation and on adaptation policies. Climate policies can be broadly categorized as either mitigation policies which reduce greenhouse gas emissions or adaptation policies which build community resilience to the impacts of climate change. The average length of a CAP is approximately 62k tokens.[1] CAPs are distributed by jurisdictions as PDFs.

There are at least three user groups could derive significant value from extracting policies from CAPs. Local sustainability officers seek to understand the policies of similar jurisdictions and how their own policies compare. State agencies wish to monitor the progress of CAP implementation and to aggregate policy commitments to the state and federal levels. Academic researchers seek to understand CAP characteristics, credibility, evolution over time, and effectiveness in shaping policy. To date, efforts to systematically extract policies have been expensive and have progressed sporadically (Berrang-Ford et al., 2021; Goonesekera and Olazabal, 2022)

3 Task Definition

Given a CAP containing policy information, we want to perform relation extraction, where the entities are policy names $p_1, \ldots, p_j$ and a closed set of policy attributes $a_1, \ldots, a_i$. We consider a single relation type of *reference* between a policy p_k and an attribute, indicating that the attribute describes the policy p_k. Each attribute can refer to at most one policy. For each policy p_k, we expect that the referential attributes $a_1, \ldots, a_i$ construct a sufficiently informative representation to perform some downstream analysis.

4 Method

Given a set of CAPs in the form of PDF documents, we first parse the PDFs into a structured markup format (e.g. HTML, Markdown, JSON) using a PDF parser. The resulting file would contain various *elements* such as headers, paragraphs, tables, and lists. As a part of the raw file content, the parser produces some document structure to mark sections and sub-sections or in some cases the table of contents. We can formalize the document structure as a graph of section headers and segments containing the section content, where headers are linked to corresponding segments, and headers can be subsumed by other headers to reflect the hierarchy induced by the PDF parser.

Since the average CAP in our dataset is 170 pages, it is necessary to divide CAPs into coherent, topically focused segments. Given a fixed level in the document structure hierarchy, the CAP can be segmented according to the structure. For example, a CAP can be broken down into a series of segments, each corresponding to the content of a sub-section. This results in a corpus of structured markup segments to perform relation extraction over.

We perform zero and few-shot relation extraction by prompting autoregressive LLMs. We draw in-context examples from the set of annotated segments. We use a two-step entity-extraction procedure to extract relations, where the second prompt is dependent on the model output of the first. For each segment, we first prompt the LLM to produce the set of policy names that appear in the segment. Then, for each policy name p_k we prompt the LLM to produce the attributes $a_1, \ldots a_l$ referring to the policy name.

The relation extraction task is linearized as is conventional with IE using LLMs (Paolini et al., 2021). Extracted entities are expressed as a JSON mapping between entity types and mention spans. Spans that do not appear in the segment are discarded to reduce the effect of model hallucination.

5 Related Work

In contrast to the understudied climate policy domain, there have been extensive studies about the performance of PDF parsing tools for documents with complex layouts over scientific domains (Ramakrishnan et al., 2012; Bast and Korzen, 2017; Meuschke et al., 2023). PDF parsers generally perform well on scientific text (Bast and Korzen, 2017), but isolating performance on non-textual elements Meuschke et al. (2023) show that PDF parsers struggle to extract tables more than all other content elements. (Deng et al., 2024) has found that processing tables as images using multi-modal LLMs can be more effective than parsing the tables into text.

For few-shot IE using LLMs, there have been mixed results over biomedical and clinical text domains (Hu et al., 2023; Li and Zhang, 2023; Jimenez Gutierrez et al., 2022; Li and Groth, 2023). There has been some limited work to perform few-shot IE in the climate policy domain: Buster et al. (2024) extract features of wind energy systems from PDF ordinances using LLMs. In our work, we systematically evaluate the end-to-end IE pipeline at both the extraction and PDF parsing stage for a

[1] As tokenized by the GPT-3.5-turbo-0125 model.

more general climate policy taxonomy.

A key limitation in using LLMs for IE over long, complex documents is limited context length. While there has been work to build models that can process long contexts (Beltagy et al., 2020), memory and attention constraints result in parts of the context being ignored. With scientific documents, Dagdelen et al. (2024) found that failures occurred when the number of tokens exceeded the model context window. Accordingly, we focus on making use of CAP structure to produce segments that fit within the context window of LLMs.

Historically, NLP tools have been used for climate policy text to identify salient topics using clustering or topic modeling (Brinkley and Stahmer, 2021). There has been some work to classify policy type or targets (e.g. pledge net-zero vs. emissions reduction) (Sachdeva et al., 2022; Biesbroek et al., 2020; Juhasz et al., 2024).

More recently, there have been multiple large-scale initiatives to extract structured representations of policies. Sewerin et al. (2023) spent over 600 hours to annotate 412 documents with 42 policy instrument and design types. Similarly, Berrang-Ford et al. (2021) have annotated climate hazards and adaptation efforts in 1,682 articles with the assistance of 126 researchers. Accordingly, in our work, we study the capacity of LLMs to assist and reduce the effort required to collect and maintain information resources about the state of climate policy.

6 Experiments and Analysis

To perform an evaluation of state-of-the-art end-to-end IE, we annotate the documents at multiple levels of granularity: raw PDFs, structured markup CAPs, filtered CAP segments, and elements of each segment.

6.1 Dataset

We collect a dataset of 227 publicly available CAPs scraped from California city and county government websites published between 2006 and 2022 (Boswell and Greve, 2023). This dataset is used for each of the annotation tasks.

To verify that core policy information can be retained in PDF parsing, we annotated descriptions of climate policies at up to five levels of granularity for 17 raw PDF CAPs from San Diego County (16 municipal and one county CAP). On average, the most concise descriptions of a policy were on 7.5 words (e.g. "Promote Installation of Commercial and Industrial Photovoltaic Systems"), and the most granular descriptions were 48.9 words (e.g. "Implement and enforce Title 18, Chapter 18.30, Section 18.30.130 of the Carlsbad Municipal Code, mandating solar photovoltaic energy generation systems on existing non-residential buildings undergoing major renovations."). We collected 1,183 policy entities.

To evaluate relation extraction performance, we annotate richer representations of policies with policy mentions and corresponding attribute mentions over a sample of parsed, structured markup segmented CAPs. Over 65 segments, we marked 102 climate policies and 838 attributes, with an average segment length of 401 words. Based on existing climate policy taxonomies (Boswell et al., 2019), we developed a minimal closed set of 11 policy attributes. Frequent subjective and ambiguous cases resulted in inter-annotator agreement Fleiss' Kappa 0.39 and Krippendorf's Alpha 0.41.

Segments were annotated by six in-house undergraduate annotators with backgrounds in public policy and computer science. All annotators used INCEpTION (Klie et al., 2018).

In addition to segments, we inspect the utility of document structure produced by the PDF parsers. For each parser, we extract a hierarchical list of sections and sub-sections. This is either explicitly generated by the parser, or induced by the header tags produced in the parsed output. Given only the ordered document structure induced by a PDF parser, we prompted an annotator to mark a subset of section headers that suggest the section likely contains policy information. In this annotation task, we determine the extent to which induced structure can be used to narrow the space of candidate segments. For the 17 San Diego CAPs, we annotated the structure produced by four PDF parsers: Nougat, Marker, GROBID, and Adobe Extract.

6.2 Extracting Document Content and Structure

We experiment with common parsers to extract and structure text from PDF documents:

Nougat (Blecher et al., 2023) does not rely on an external OCR engine. Instead, it uses a visual encoder, an mBART decoder, and a tokenizer specialized in scientific text. The parsed output is in a markup language that supports headers, which are used for segmentation, and LaTeX tables.

Marker is a widely-used pipeline of deep-learning models including a Tesseract OCR engine to extract text, detect page layout, and convert to markdown.[2] Marker supports hierarchical headers in the parsed output, which we use for segmentation.

GROBID (Lopez, 2009) structures PDFs into an XML/TEI encoded document using maximum chain Markov models and linear-chain CRF. GROBID also extracts the table of contents, which we use for segmentation.

Adobe Extract API uses Adobe Sensei ML to extract paragraphs, lists, headings, tables.[3] We convert the output to HTML format. We are able to extract the table of contents using Adobe Extract, and segment the documents according to varying levels of depth of the table of contents.

To evaluate raw text and element recognition, we use Levenshtein distance (Levenshtein et al., 1966), or the number of character insertions, deletions substitutions necessary to transform a contiguous span from an extracted segment into a reference piece of text. We measure the recall of a set of parsed segments using a threshold of 10 edits. Among matching extracted segments, we normalize distance over the length of the reference text.

PDF Parser	Policy Description Recall	
	Struct. Markup	Filtered Struct. Markup
Nougat	0.47	0.41
Marker	0.86	0.68
GROBID	0.21	0.18
Adobe	0.81	0.81

Table 1: We report policy description recall over the structured markup document and the subset of segments that are annotated as policy-relevant. We can observe that annotating with the structure given by the Adobe PDF parser suffers no policy information loss while significantly reducing the content to perform inference over.

Raw Text Recognition: In Table 1, we inspect to what extent the PDF parsers preserve the policy descriptions and uncover an underlying document structure that would make it feasible to extract the policy segments under annotation resource constraints.

First, we find that for most PDF parsers, core policy information is retained after parsing the PDF into structured markup formats. We estimate how much of the core policy information is dropped or heavily distorted in the parsing process by comparing the annotated policy descriptions and the CAP segments using the `fuzzysearch` library[4].

PDF Parser	Element Recall			
	Tables	Paragraphs	Lists	Headers
Nougat	0.24	0.72	0.62	0.86
Marker	0.78	1.00	0.82	1.00
GROBID	0.68	1.00	0.74	1.00
Adobe	1.00	1.00	0.79	1.00

Table 2: Element-wise recall of PDF parsers over a sample of 10 segments, where element is considered recognized using a fuzzy string match between the textual content of the PDF reference and the parsed element. Tables and lists are generally most challenging for parsers to recognize.

Element-wise Recognition: We also evaluate the PDF parsers for *element recognition*. Critical elements in CAPs include tables, paragraphs, lists, and headers. For a random sample of 10 CAPs, we identify a policy-rich segment and annotate critical elements in the segment from the raw PDF. We can measure recall of these elements in the structured markup form of the CAP.

In Table 2, we observe that table and list elements are typically more challenging to exract in CAPs than purely textual elements like headers and paragraphs. In Table 1, we observe that most policy text segments can be matched in the PDF parser output to with the exception of GROBID. Almost all PDF parsers struggle to recognize tables in CAPs. In contrast with scientific articles, Blecher et al. (2023) report table recall 50 points higher on open-access ArXiv articles than CAPs, Poor parsing performance on tables is an important bottleneck for policy extraction over CAPs, since tables are often the most information-dense elements of the document.

6.3 Segment Filtering

The PDF parsers convert the documents to a structured markup format, which includes ordered lists of section headers and segments containing the

[2] https://github.com/VikParuchuri/marker

[3] Adobe PDF Extract API Technical Brief

[4] https://github.com/taleinat/fuzzysearch

	Llama2				GPT-3.5			
	k=0	k=1	k=2	k=3	k=0	k=1	k=2	k=3
Entity Extraction	0.00	1.89	13.04	**15.48**	5.83	10.21	10.16	9.46
+ GOLD policies	34.28	53.95	55.28	**64.95**	47.10	49.53	52.00	51.10
Relation Extraction	0.00	0.00	3.77	4.87	0.00	1.64	6.65	**6.79**
+ GOLD policies	2.38	44.06	54.28	**57.84**	25.55	40.42	42.48	41.77

Table 3: Entity and Relation Extraction F1 in zero and few-shot settings. Given GOLD policy names, extraction performance is much stronger. Including multiple in-context examples also seems to improve performance. This may be a result of ambiguous levels of abstraction for policy mentions that are challenging to specify using instructions of in-context examples.

section content. In practice, inference can be expensive over potentially hundreds of sections in the document. Additionally, inference over less relevant sections such as front matter, policy landscape, and climate history and projection sections may yield false positives in the policy extraction task as these sections often contain boilerplate discussions of policies not specific to the relevant jurisdiction (Scott et al., 2022). Given only the high-level document structure produced by the PDF parser (i.e., section headers, table of contents), we filter the subset of policy-rich segments. After annotating 25 document structures, we pass BERT-based representations of the section headers through a 2-layer feed-forward neural network with a binary classification head indicating policy informativeness. We perform this *structured segmentation* using the parsed document structure to identify a subset of segments that contain policy information.

Structured Segmentation: We measure the utility of the structure that the parser extracts. In Table 1, we find that after annotating a sufficient substructure, Adobe suffers no loss in policy information, while reducing the amount of segments to process by 58%. The structure produced by Nougat and Marker is generally longer, often containing extraneous elements, since the structure elements include any header in the documents. This may result in annotator mistakes and consequently a small loss in policy information.

6.4 Relation Extraction

Given a corpus of segments from the CAPs, we perform inference in zero and few-shot settings. To select in-context examples we use the k nearest neighbors from the target example based on cosine distances of Sentence-BERT representations (Reimers and Gurevych, 2019) to the target context. Using 10-fold cross-validation over the set of segments annotated for both policies and attributes, all results are computed over 6 random seeds.

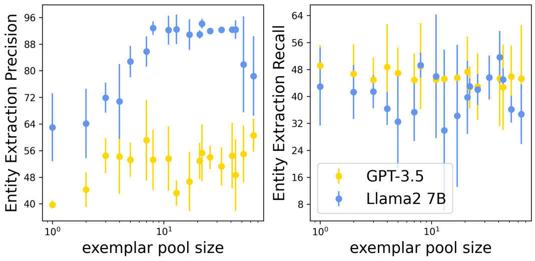

Figure 2: Precision and Recall for 3-shot entity extraction where we vary the exemplar pool size to select ICL examples from and GOLD policy names are given. For higher-quality ICL examples, precision shows clear improvements for both models, but this is not the case for recall.

At inference time, generated outputs are parsed as a JSON object. We assess GPT-3.5-turbo-0125 using the function-calling feature of the API to constrain the output to a JSON format. We use a function-calling finetuned 7B Llama2 model as well.[5] This results in a set of policies, where each policy name is linked to a set of attributes. We evaluate these widely-used models to measure both *segment-level extraction* and *intra-element extraction* performance. In post-processing, predicted spans that do not appear in the context are dropped. We compute standard metrics (P, R, F1) for linearized, typed relation and entity tuples. We use a relaxed string matching setup between reference and generated spans similar to previous work with generative models that do not produce standardized outputs (Wadhwa et al., 2023).

Segment-level Extraction: We can observe in Table 3 that when gold policy names are given, the model performance is strong for both Llama2

[5] https://huggingface.co/Trelis/Llama-2-7b-chat-hf-function-calling-v2

and GPT-3.5-turbo. Without gold policy names, the performance is extremely poor, especially for RE. Including multiple in-context examples in the prompt is necessary for reasonable performance. Upon manual inspection of model output, we observe that this is partially a result of the ambiguous nature of annotation. For example, a policy may be broadly previewed in an introduction section with abstract terms and concretely enumerated as a list of measures in the appendix of the CAP. With global context and knowledge about how CAPs are typically structured, an annotator can correctly ignore abstract mentions of policies. The model, however, is limited to only a single segment.

The model performance does not seem to improve with higher quality examples. We experiment with varying the size of the exemplar pool from which we select in-context examples in Figure 2. We find that for both models, annotating more than 10 exemplars does not seem to improve overall F1 performance significantly. In general, higher quality exemplars improves precision, but has little effect on model recall. One explanation for this is that the ICL examples have erratic levels of policy abstraction, so that similar contexts are not necessarily more useful for policy recall.

Intra-element Extraction: We inspect performance of models over policy-rich document elements such as tables and lists. We annotate an additional 20 segments that contain table and list elements and report zero-shot performance in Table 4 for policy name extraction. For all table and text formats, policy name extraction performance is poorer over segments containing non-textual elements than segments that contain only paragraph elements. Upon manual analysis of the model output, we observe that models can easily identify policies from well-formed tables (i.e. there are no breaks in columns, cells are merged consistently).

We can observe instances of hallucination in Table 5. In the first example, the model hallucinates a policy called "Severe Storm Preparedness Measures" intended to target the climate hazard described in the content. While this can be avoided altogether by enforcing that extracted spans occur in the context, models would be most prone to policy hallucination if the segment maps to a section that does not contain policies. In the second example, we can observe that well-behaved tables with contiguous rows and columns can also be interpreted incorrectly. For example, the model incorrectly

Text Format	Table Format	P	R	F1
HTML	HTML	26.47	4.31	7.41
Plain	MD	29.69	9.00	13.82
Plain	CSV	37.93	10.89	16.92
Plain	TSV	17.11	7.47	10.40
Plain	JSON	48.28	13.33	**20.90**

Table 4: We evaluate policy extraction performance over a sample of 20 challenging segments containing complex tables and lists in a zero-shot setting using a GPT-3.5-turbo model. We can observe that regardless of prompt format, models struggle to extract policies from non-textual elements.

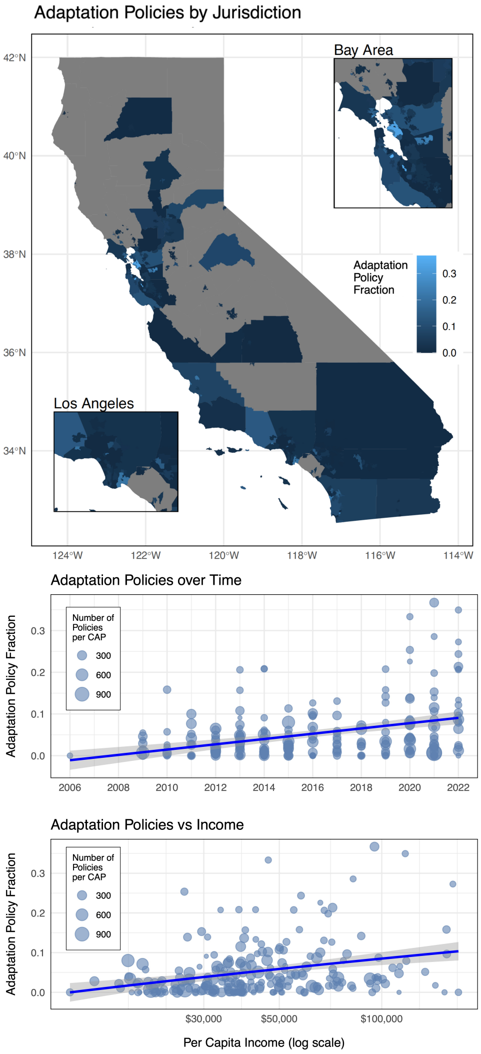

Figure 3: Top: Based on adaptation policy share across jurisdictions, we observe coastal jurisdictions are more focused on adaptation ($r = 0.03$, $p < 1 \times 10^{-10}$). Middle: Number of adaptation policies has increased over time ($R^2 = 0.13, p < 2 \times 10^{-8}$). Bottom: Number of adaptation policies increases with the logarithm of community per capita income ($R^2 = 0.091$, $p < 5 \times 10^{-6}$).

Context	Extracted Policy
Severe storms include **windstorms, hail, lightning, thunderstorms, and heavy rainfall**. Severe weather is usually caused by intense storm systems . . . The most vulnerable to severe weather are persons experiencing **homelessness; those who may live in less structurally resilient buildings**. . . The **energy delivery system** is especially vulnerable to windstorms, . . . cause **Public Safety Power Shutoff (PSPS) events**	*policy name:* "Severe Storm Preparedness Measures" *reference:* ["Public Safety Power Shutoff (PSPS) . . . "], *capacity:* ["energy delivery system"], *subpopulations:* ["homelessness", "less structurally resilient buildings . . . "], *climate-hazards:* ["severe storms", "windstorms", . . .]
Table 11: Measures to Reach AB 32 Target 0 | Sphere | GHG Sector | Focus Area Linkage | Measure AB 32/Kyoto | Annual Savings (Tonnes CO2e) | Job Creation Estimate | Annual Savings | Estimated Implementation Cost to City WORK – 2 | **Commercial Buildings** | **Peak Demand Reduction**: **Collaborate with SCE and encourage 100 businesses to enroll in Energy Efficiency and Demand** | A | 193 | 1.0 | $96,900 | **$2,000** . . .	*policy name:* "Peak Demand Reduction" *description:* ["Collaborate with SCE and encourage businesses to enroll in Energy Efficiency and . . . "], *responsible-entity:* ["SCE", "CVAG", "Los Angeles County Energy Enterprise Management . . . "], *cost:* ["$2000"], *sector:* ["Commercial Buildings"]

Table 5: Example annotations from GPT-3.5-turbo. Among the *entity types* and **extracted spans**, we observe hallucinated policy names and attributes when there are no ground-truth policies in the context (top) or when the model fails to localize row content from tables (bottom).

reports two management entities: "CVAG", "Los Angeles County Energy Enterprise Management Information System (EEMIS)" that appear in the following row of the table.

7 Downstream Extraction Utility

A question of interest to policy researchers is how CAP focus has shifted from mitigation to adaptation as faith in global mitigation efforts has declined (Hoesung Lee and José Romero (eds.), 2023). We classify a policy as an adaptation policy using a heuristic string match over the referring attributes with the regular expression "heat|precip|flood|fire|sea level". A keyword search with this regular expression over the entire document could extract mentions of hazards in introductory sections that are not associated with concrete policies.

Using a GPT-3.5-turbo model, we extracted 47,006 climate policies from 227 jurisdictions. 4.6 percent of the extracted policies mention the five hazards in their "climate-hazard" attribute. The low percentage is expected as CAPs have traditionally focused on greenhouse gas emission reductions rather than adaptation.

Linking the extracted set of policies to county and municipal characteristics reveals spatial variability in the fraction of adaptation policies (Figure 3) with a slight indication that coastal jurisdictions are more focused on adaptation. The share of adaptation policies has increased over the past 16 years. There is a significant positive linear relationship between the fraction of adaptation policies and the logarithm of community per capita income. This may indicate that wealthier communities show more interest in safeguarding their assets, though further investigation is required to rule out potential confounding factors, such as distance to the coast.

8 Recommendations

To improve end-to-end IE over gray policy literature, we propose several directions for future work based on our analysis of California CAPs.

There is a need for more **flexible annotation schemes** for inconsistently formatted documents. The same entities will frequently appear at different levels of abstraction or detail between documents and within a document. We need mechanisms to specify which mentions to extract. This is distinct from fine-grained entity typing annotation where there are no vertical coreferences between types.

IE systems need better **representation of non-textual elements** such as tables and rich graphics, and for semantic representations of the text to be built from information derived from all modalities. In a text-only modality, it may be useful to build representations that localize information content according to reading order.

Vague, imprecise, and subjective entity types are crucial for gray policy literature, and current methods to refine the set of extracted entities such as providing ICL examples or detailed instructions may be insufficient. In the case of CAPs, interpretation of policy mentions often require global document information or domain expertise.

IE systems should also be designed to **leverage**

rich document structure. In the case of CAPs, for example, we observed that that the table of contents in isolation is informative enough for annotators to infer which sections are relevant.

References

Hannah Bast and Claudius Korzen. 2017. A benchmark and evaluation for text extraction from pdf. In *2017 ACM/IEEE joint conference on digital libraries (JCDL)*, pages 1–10. IEEE.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Lea Berrang-Ford, AR Siders, Alexandra Lesnikowski, Alexandra Paige Fischer, Max W Callaghan, Neal R Haddaway, Katharine J Mach, Malcolm Araos, Mohammad Aminur Rahman Shah, Mia Wannewitz, et al. 2021. A systematic global stocktake of evidence on human adaptation to climate change. *Nature climate change*, 11(11):989–1000.

Robbert Biesbroek, Shashi Badloe, and Ioannis N Athanasiadis. 2020. Machine learning for research on climate change adaptation policy integration: an exploratory uk case study. *Regional Environmental Change*, 20(3):85.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*.

Michael R. Boswell and Adrienne I. Greve. 2023. California climate action plan database. Data set.

Michael R Boswell, Adrienne I Greve, and Tammy L Seale. 2019. *Climate action planning: a guide to creating low-carbon, resilient communities*. Island Press.

Catherine Brinkley and Carl Stahmer. 2021. What is in a plan? using natural language processing to read 461 california city general plans. *Journal of Planning Education and Research*, page 0739456X21995890.

Grant Buster, Pavlo Pinchuk, Jacob Barrons, Ryan McKeever, Aaron Levine, and Anthony Lopez. 2024. Supporting energy policy research with large language models. *arXiv preprint arXiv:2403.12924*.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

Naihao Deng, Zhenjie Sun, Ruiqi He, Aman Sikka, Yulong Chen, Lin Ma, Yue Zhang, and Rada Mihalcea. 2024. Tables as images? exploring the strengths and limitations of llms on multimodal representations of tabular data. *arXiv preprint arXiv:2402.12424*.

Sascha M Goonesekera and Marta Olazabal. 2022. Climate adaptation indicators and metrics: State of local policy practice. *Ecological Indicators*, 145:109657.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*.

Hoesung Lee and José Romero (eds.). 2023. *Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. IPCC, Geneva, Switzerland.

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2023. Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv:2303.16416*.

Pengcheng Jiang, Jiacheng Lin, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2024. Genres: Rethinking evaluation for generative relation extraction in the era of large language models. *arXiv preprint arXiv:2402.10744*.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 in-context learning for biomedical IE? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matyas Juhasz, Tina Marchand, Roshan Melwani, Kalyan Dutia, Sarah Goodenough, Harrison Pim, and Henry Franks. 2024. Identifying climate targets in national laws and policies using machine learning. *arXiv preprint arXiv:2404.02822*.

Uri Katz, Matan Vetzler, Amir Cohen, and Yoav Goldberg. 2023. NERetrieve: Dataset for next generation named entity recognition and retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3340–3354, Singapore. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Amanda Lawrence, Julian Thomas, John Houghton, and Paul Weldon. 2015. Collecting the evidence: Improving access to grey literature and data for public policy and practice. *Australian Academic & Research Libraries*, 46(4):229–249.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Mingchen Li and Rui Zhang. 2023. How far is language model from 100% few-shot named entity recognition in medical domain. *arXiv preprint arXiv:2307.00186*.

Xue Li and Paul Groth. 2023. How different is different? systematically identifying distribution shifts and their impacts in ner datasets.

Patrice Lopez. 2009. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries: 13th European Conference, ECDL 2009, Corfu, Greece, September 27-October 2, 2009. Proceedings 13*, pages 473–474. Springer.

Norman Meuschke, Apurva Jagdale, Timo Spinde, Jelena Mitrović, and Bela Gipp. 2023. A benchmark of pdf information extraction tools using a multitask and multi-domain evaluation framework for academic documents. In *International Conference on Information*, pages 383–405. Springer.

Ramesh Pandita and Shivendra Singh. 2011. Grey literature: A valuable untapped stockpile of information. *Journal of the Young Librarians Association*, 5.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *arXiv preprint arXiv:2101.05779*.

Cartic Ramakrishnan, Abhishek Patnia, Eduard Hovy, and Gully APC Burns. 2012. Layout-aware text extraction from full-text pdf of scientific articles. *Source code for biology and medicine*, 7:1–10.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III 21*, pages 148–163. Springer.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the eighth conference on computational natural language learning (CoNLL-2004) at HLT-NAACL 2004*, pages 1–8.

Hannah R Rothstein and Sally Hopewell. 2009. Grey literature. *The handbook of research synthesis and meta-analysis*, 2:103–125.

Siddharth Sachdeva, Angel Hsu, Ian French, and Elwin Lim. 2022. A computational approach to analyzing climate strategies of cities pledging net zero. *npj Urban Sustainability*, 2(1):21.

Tyler A. Scott, Nicholas Marantz, and Nicola Ulibarri. 2022. Use of boilerplate language in regulatory documents: Evidence from environmental impact statements. *Journal of Public Administration Research and Theory*, 32(3):576–590.

Sebastian Sewerin, Lynn H Kaack, Joel Küttel, Fride Sigurdsson, Onerva Martikainen, Alisha Esshaki, and Fabian Hafner. 2023. Towards understanding policy design through text-as-data approaches: The policy design annotations (polianna) dataset. *Scientific Data*, 10(1):896.

Anne M Turner, Elizabeth D Liddy, Jana Bradley, and Joyce A Wheatley. 2005. Modeling public health interventions for improved access to the gray literature. *Journal of the Medical Library Association*, 93(4):487.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context learning for relation extraction using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547, Singapore. Association for Computational Linguistics.

Siyu Yuan, Deqing Yang, Jiaqing Liang, Zhixu Li, Jinxi Liu, Jingyue Huang, and Yanghua Xiao. 2022. Generative entity typing with curriculum learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3061–3073, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Limitations

For downstream use of extracted policies, it is necessary to link policy mentions between segments. In this work, we do not evaluate the quality of cross-document coreference systems for linking policy mentions. Ideally, extracted document segments should be (1) long enough to such that there is a low probability that a single policy p_k is mentioned across many segments, and (2) short enough to fit into the context window of a LLM at inference time. We do not verify that policies are rarely mentioned across multiple segments. This results in some redundancy in the entities we extract.

There are also many cases where entities cannot be extracted without context that is trapped in figures or icons (e.g. icons are used to indicate climate hazards).

This work is also a case study of end-to-end IE over CAPs. While gray policy literature shares some of the challenging properties of CAPs, we have focused on a single type of gray policy document, and we have not measured how to what extent our findings generalize.

Policies are often mentioned in a CAP at varying levels of abstraction. For example, in the Encinitas table of contents, a policy about renewable energy is mentioned in multiple sections: the "Climate Action Plan Overview", "GHG Reduction Strategy Framework", "Table 3-2 Effect of Plan Actions on City of Encinitas Emissions and Target (MTCO2e) 3-2" in the Appendix. To extract all of the relevant attributes for the renewable energy policy, it would be necessary to link coreferent policy mentions and aggregate the set of attributes across mentions.

Policy extraction with CAPs can be subjective and require additional resources. For example, a policy scientist may want to measure projected emissions reductions for a region. Some jurisdictions may use soft language to describe a policy in a CAP (e.g. "Consider the implementation of renewables" vs. "Establishes a Renewable Portfolio Standard requiring …"). One criticism of CAPs is that there is no guarantee that a jurisdiction will implement a given policy, so it may be necessary to reference external documents (e.g. funding proposals and annual budgets).

Ethics Statement

Existing tools for end-to-end IE have significant performance limitations and are not necessarily robust enough to be used for decision-making. We highlight several areas of future work for extracting information from California CAPs, but it is unclear to what extent those areas would serve IE over CAPs from other parts of the world.

End-to-end IE for CAPs targets three user groups that would be affected: local sustainability officers, state agencies, and academic researchers. Poor performance of these models could result in additional work to correct model responses. Missing or incorrect extractions could also lead to an inaccurate understanding of progress in adaptation or emissions reduction. For that reason, it is crucial that user groups are aware that state-of-the-art tools have important limitations.

Acknowledgements

Thanks to Adrienne Greve and Mike Boswell for valuable feedback on this work. We additionally thank our annotators: McKenna Carlson, Nida Firdaws, Elvin Li, Wushuang Li, Matthew Le, and Calvin Nguyen.

A Extended Related Work

In part, a core challenge in entity and relation extraction with generative models is inconsistent output format (Jiang et al., 2024). Previous work has shown that manual annotation of model output can reveal that unannotated spans are a major source of errors, and that soft matching of spans can make evaluation more precise (Wadhwa et al., 2023; Han et al., 2023). Katz et al. (2023) has shown that constraining the output to a JSON format can also improve the consistency of LLM outputs – a finding that we make use of in this work.

B Dataset and Additional Results

	Adobe	Marker	Nougat	GROBID
Lev. Distance	0.015	0.040	0.035	0.102
Tokens	0.34	0.27	0.28	0.43
Segments	0.42	0.55	0.46	0.84

Table 6: We report the Levenshtein distance to GOLD policy descriptions. In addition, we report policy description recall over all segments (Tokens) and the subset of segments annotated as policy-relevant (Segments). We can observe that annotating with the structure given by the Adobe PDF parser suffers no policy information loss while significantly reducing the content to perform inference over.

Policy Attribute	Instruction
description	Extract a description for the policy.
management	Extract mentions of the individual/entities responsible for implementation of the policy.
funding	Extract mentions of the funding source for the policy.
co-benefits	Extract mentions of co-benefits for the policy.
reference	Extract mentions of references/legislation (e.g. State Senate Bill, State Assembly Bill, County Ordinance, City General Plan reference, City Local Hazard Mitigation Plan reference) relevant to the policy.
capacity	Extract mentions of adaptive capacity relevant to policy.
subpopulations	Extract mentions of the affected subpopulations for the policy (e.g. people who work outdoors, hazardous materials facilities).
climate-hazard	List what climate hazards are relevant to the policy? (e.g. extended droughts, sea-level rise, extreme heat)
participation	Extract mentions of the level of participation necessary for the policy to be successful.
begin	Extract mentions of when the policy will begin.
complete	Extract mentions of when the policy will be complete.
evaluation	Extract mentions of how the policy will be evaluated.
cost	Extract mentions of how much the policy will cost.
feasibility	Extract mentions describing the feasibility of the policy (e.g. low-cost, existing policy).
jurisdiction	Extract mentions (if they exist) of whether the policy is a city, state, regional, or federal policy.
sector	Extract mentions (if they exist) of whether the impact sector for the policy is the built environment, economy, ecosystem, systems, or social justice.
target	Extract mentions (if they exist) of whether the policy targets the community-at-large or municipal assets.
assumptions	Extract mentions of assumptions about the policy .

Table 7: For each policy, we extract a set of attributes using a short description of the attribute.

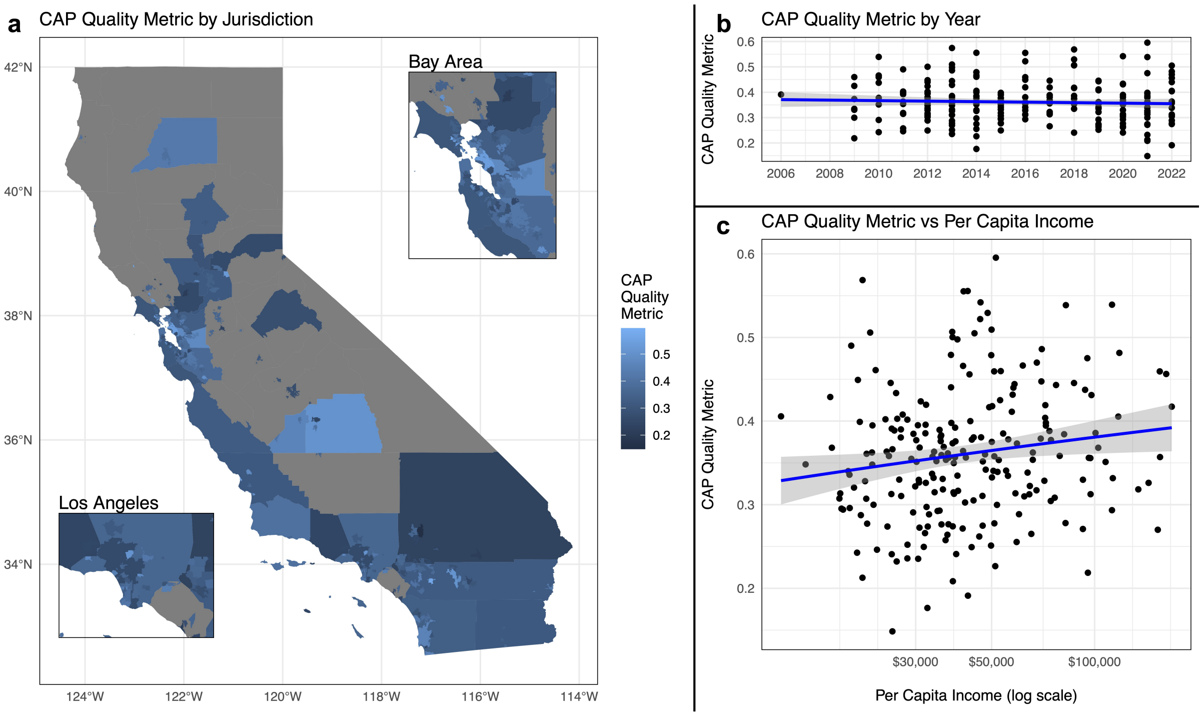

Figure 4: We define "CAP Quality" as the number of non-missing observations for each of the GPT-3.5 extraction fields (Table 7). Suppose a CAP has n policies of which a proportion p_j for characteristic or relation j (e.g., management, or funding) is non-missing. For $k = 17$ characteristics, we define an overall quality metric to be $\frac{1}{k}\sum_{j=1}^{k} p_j$. For this particular quality metric, there is wide spatial variability with no discernable patterns, a slight decline in quality over time (though not statistically significant), and a positive and statistically significant linear relationship with per capita income indicating that wealthier communities tend to produce higher quality CAPs according to this simple metric ($p = 0.0185$, $R^2 = 0.025$).