Policy Extraction for Wildfire Resilience: A PIRS-Based Approach

Bernette Chan, Jun Noh, Nidhi Shinde, and Zarina Yunis

Abstract

Wildfires continue to grow in frequency and intensity across California, requiring legislators to write policies that enhance wildfire resilience. Counties have several plans with conflicting and wordy policies, making it difficult for city planners to evaluate the efficacy of these initiatives. This report explores natural language processing (NLP) techniques used to identify policies related to wildfires. By applying methods like rule-based text extraction and Large Language Model (LLM) prompting, we extract policy statements from four Atascadero County plans. These policies are classified into categories using topic modeling. The ultimate goal of this project is to partially automate a system that scores each policy based on its potential to improve wildfire resilience. Using geographic information systems (GIS), we can map scored policies spatially so that counties can prioritize areas with a high risk of fire-related damage. Moving forward, we aim to extract policies across more counties and refine our categorization strategy to begin the scoring and evaluation process.

I. Introduction

WILDFIRES have become an increasingly severe and frequent threat across the United States, especially in the Western states. Over the past 30 years, the number, size, and intensity of wildfires have risen dramatically due to both natural and human made causes. Climate change has led to hotter temperatures, prolonged droughts, and increased fuel availability, while urban expansion into fire-prone areas and human activities such as power line failures and accidental ignitions have further exacerbated the problem. According to the National Interagency Coordination Center, the number of acres burned annually has steadily increased, causing significant economic, environmental, and human losses.

In response to this growing threat, policymakers and urban planners must integrate wildfire resilience strategies into land use planning and hazard mitigation policies. However, a major challenge in this process is the inconsistent and fragmented documentation of these policies. Many local governments struggle with poorly formatted planning documents, conflicting hazard mitigation strategies, and a lack of standardized methods for evaluating wildfire policies. This lack of cohesion makes it difficult for city planners to systematically assess and improve wildfire resilience efforts

To address this issue, we apply data science techniques to assist city planners in identifying fire-related policies within their planning documents. By analyzing policy language and categorizing relevant content, our approach allows planners to efficiently mark policies within the Planning Integration for Resilience Scorecard (PIRS). This tool provides a structured framework for evaluating how well wildfire resilience is integrated into existing plans, ultimately helping cities strengthen their wildfire preparation and response strategies.

II. Dataset Description

The data was provided to us in the form of several large city planning documents. Throughout this project, we worked with several California counties - Atascadero, Napa, Temecula, and Santa Barbara. While each document varied slightly in structure and content, they all outlined a plan for the city, along with actionable items in various realms of hazard mitigation, including wildfires.

Due to the varying nature of each city planning document, we initially had trouble creating a pipeline to pinpoint and extract policies in a way that was consistent across all four documents. Some of the documents had their policies clearly listed with a bold heading containing the measure number and title followed by the measure's text. This structure allowed us to iterate through each page of the document, using regular expressions to extract relevant policies and reformat them into a dataframe. While this technique was successful for structured documents, we quickly realized that it failed to extract implicitly defined policies. This made it difficult to use the same Python libraries and direct text extraction techniques. To resolve this issue, we tested and implemented several different techniques.

All authors are with the Departments of Mathematics and Statistics, California Polytechnic State University, San Luis Obispo, CA, USA.
Emails: bchan42@calpoly.edu, jnoh02@calpoly.edu, nashinde@calpoly.edu, zyunis@calpoly.edu.

III. Our Approach

Two of the methods we used to find policies were table extraction and Large Language Model (LLM) approaches. After examining the documents, particularly the City of Atascadero Final Climate Action Plan, we discovered that most of the relevant policies were organized into tables with categories for the measure label, actions, and more. We used this structure to split the relevant text into its respective tables. Then we removed extra sections from the text unrelated to policies, focusing on pages with policy-related text. This made our extraction method time-efficient because the text data we were examining was on a much smaller scale. Using the Pandas library in Python, we stored cleaned policies and their labels (number and letter combinations) in a dataframe. After finding relevant policy tables from each page, we wrote the extracted tables to CSV files containing a policy number and policy description. While some of the documents' policies were easy to extract, other documents had policies with varying formats, which made it difficult to implement just one technique for identifying policies, goals, and measures. For example, the Wildfire Protection Plan had policies in tabular format and also integrated into paragraphs of text.

IV. Policy Extraction Techniques

A. Ruled-Based Extraction

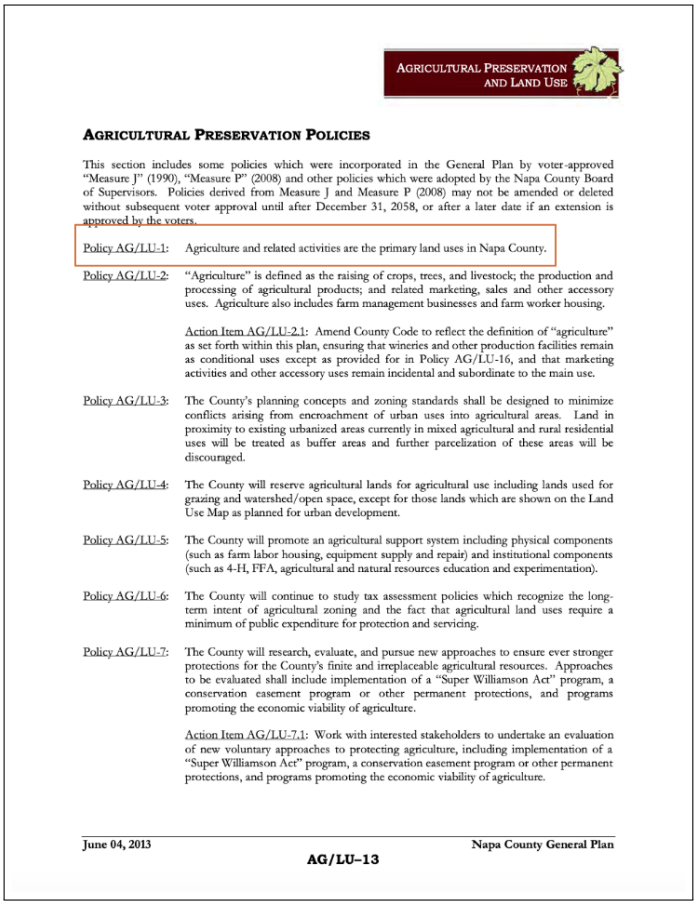
Rule-based text extraction involves chunking each document line by line to identify which text is a policy given regular expressions and other structure-based logic. In structured planning documents, visual or typographic emphasis (such as bold, italic, or enlarged fonts) often signals the beginning of key items like policies, programs, or actions. Because formatting conventions vary across documents, extraction logic must be adapted to each document's specific structure.

For example, one document may organize content hierarchically by goals, followed by nested policies and programs. Goals may follow naming conventions like all-caps titles with unique identifiers (e.g., "Goal LOC 1"), while policies and programs follow consistent numbering schemes (e.g., "Policy 1.1", "Program 1.1.1"). Regular expressions are used to identify these patterns and capture their associated text. Section boundaries are typically inferred from layout cues such as line breaks or spacing.

Extracted items are stored in a hierarchical data structure to preserve relationships (e.g., which policy belongs to which goal). This structured output can then be exported to a spreadsheet or database for further analysis or client review.

A similar strategy is used for documents with less consistent formatting, such as tables of measures and actions. In such cases, well-numbered entries and tabular layout make it easier to isolate and extract action items systematically.

For instance, the Napa General Plan is organized by thematic chapters such as Agriculture and Land Use, Housing, and Safety, each containing clearly labeled goals and corresponding policies. A typical structure includes a series of goals (e.g., *Goal AG/LU-1*, *Goal AG/LU-2*), followed by related policies (e.g., *Policy AG/LU-1*, *Policy AG/LU-2*). These elements were extracted using regular expressions tailored to the document's formatting conventions and compiled into a spreadsheet to support structured policy analysis (see Figure 1 below).



AGRICULTURAL PRESERVATION AND LAND USE

AGRICULTURAL PRESERVATION POLICIES

This section includes some policies which were incorporated in the General Plan by voter-approved "Measure J" (1990), "Measure P" (2008) and other policies which were adopted by the Napa County Board of Supervisors. Policies derived from Measure J and Measure P (2008) may not be amended or deleted without subsequent voter approval until after December 31, 2058, or after a later date if an extension is approved by the voters.

Policy AG/LU-1: Agriculture and related activities are the primary land uses in Napa County.

Policy AG/LU-2: "Agriculture" is defined as the raising of crops, trees, and livestock; the production and processing of agricultural products; and related marketing, sales and other accessory uses. Agriculture also includes farm management businesses and farm worker housing.

Action Item AG/LU-2.1: Amend County Code to reflect the definition of "agriculture" as set forth within this plan, ensuring that wineries and other production facilities remain as conditional uses except as provided for in Policy AG/LU-16, and that marketing activities and other accessory uses remain incidental and subordinate to the main use.

Policy AG/LU-3: The County's planning concepts and zoning standards shall be designed to minimize conflicts arising from encroachment of urban uses into agricultural areas. Land in proximity to existing urbanized areas currently in mixed agricultural and rural residential uses will be treated as buffer areas and further parcelization of these areas will be discouraged.

Policy AG/LU-4: The County will reserve agricultural lands for agricultural use including lands used for grazing and watershed/open space, except for those lands which are shown on the Land Use Map as planned for urban development.

Policy AG/LU-5: The County will promote an agricultural support system including physical components (such as farm labor housing, equipment supply and repair) and institutional components (such as 4-H, FFA, agricultural and natural resources education and experimentation).

Policy AG/LU-6: The County will continue to study tax assessment policies which recognize the long-term intent of agricultural zoning and the fact that agricultural land uses require a minimum of public expenditure for protection and servicing.

Policy AG/LU-7: The County will research, evaluate, and pursue new approaches to ensure ever stronger protections for the County's finite and irreplaceable agricultural resources. Approaches to be evaluated shall include implementation of a "Super Williamson Act" program, a conservation easement program or other permanent protections, and programs promoting the economic viability of agriculture.

Action Item AG/LU-7.1: Work with interested stakeholders to undertake an evaluation of new voluntary approaches to protecting agriculture, including implementation of a "Super Williamson Act" program, a conservation easement program or other permanent protections, and programs promoting the economic viability of agriculture.

(a) Example of the Napa General Plan document structure.

	B
1	**Policy**
2	Policy AG/LU-1: Agriculture and related activities are the primary land uses in Napa County.
3	Policy AG/LU-2: "Agriculture" is defined as the raising of crops, trees, and livestock; the production and processing of agricultural products; and related marketing, sales and other accessory uses. Agriculture also includes farm management businesses and farm worker housing. Action Item AG/LU-2.1: Amend County Code to reflect the definition of "agriculture" as set forth within this plan, ensuring that wineries and other production facilities remain as conditional uses except as provided for in Policy AG/LU-16, and that marketing activities and other accessory uses remain incidental and subordinate to the main use.
4	Policy AG/LU-3: The County's planning concepts and zoning standards shall be designed to minimize conflicts arising from encroachment of urban uses into agricultural areas. Land in proximity to existing urbanized areas currently in mixed agricultural and rural residential uses will be treated as buffer areas and further parcelization of these areas will be discouraged.
5	Policy AG/LU-4: The County will reserve agricultural lands for agricultural use including lands used for grazing and watershed/open space, except for those lands which are shown on the Land Use Map as planned for urban development.
6	Policy AG/LU-5: The County will promote an agricultural support system including physical components (such as farm labor housing, equipment supply and repair) and institutional components (such as 4-H, FFA, agricultural and natural resources education and experimentation).
7	Policy AG/LU-6: The County will continue to study tax assessment policies which recognize the long- term intent of agricultural zoning and the fact that agricultural land uses require a minimum of public expenditure for protection and servicing.
8	Policy AG/LU-7: The County will research, evaluate, and pursue new approaches to ensure ever stronger protections for the County's finite and irreplaceable agricultural resources. Approaches to be evaluated shall include implementation of a "Super Williamson Act" program, a conservation easement program or other permanent protections, and programs promoting the economic viability of agriculture. Action Item AG/LU-7.1: Work with interested stakeholders to undertake an evaluation of new voluntary approaches to protecting agriculture, including implementation of a "Super Williamson Act" program, a conservation easement program or other permanent protections, and programs promoting the economic viability of agriculture.

(b) Extracted General Plan policies in spreadsheet format.

Fig. 1: Document layout and extracted policy spreadsheet from the Napa General Plan with Policy 1 highlighted.

We also tested this approach for table extraction.

Upon examining the documents, particularly the City of Atascadero Final Climate Action Plan, we discovered that most of the relevant policies were organized into tables with categories for the measure label, actions, and more. We used this structure to split the relevant text into its respective tables. After coming to this realization, we also removed extra sections from the text that were unrelated to policy, and solely focused on the pages with policy-related text. This made our extraction method much more time-efficient because the text data we were examining was on a much smaller scale. Using the Pandas library in Python, we stored cleaned policies and their labels (number and letter combinations) in a dataframe. After finding relevant policy tables from each page in a document, we wrote the extracted tables to new CSV files, each containing a policy number and a corresponding policy description.

B. Querying LLMs

Unstructured documents pose a unique challenge for extraction because policies are contained in random paragraphs, lists, and other chunks of text that do not contain headings.

To address this issue, we leveraged tools that work like a human reader, by replicating the process of reading through text and identifying relevant information. Large language models (LLMs) like Google's Gemini and OpenAI's ChatGPT accomplish this task by reading a prompt and interpreting it to extract important information.

However, these models are black-box technologies, so we cannot directly analyze the model parameters to understand the reasoning behind the mistakes they could make. For instance, we could ask Gemini to extract all actionable policies from a single page; this prompt could work perfectly on one page and miss key policies on another. To maximize accuracy, we tested different queries and found some that captured as many potential policies as possible.

Our process involves four steps:

1) Splitting the document into chunks
2) Writing a prompt

3) Querying the Gemini API
4) Saving the responses to a text file

1) Chunking: We began by splitting each document into chunks (pages or paragraphs) to stay within the usage limits of Gemini's API.

- The Gemini API free tier has the following restrictions:
  1) Daily limits on token (character) processing
  2) Limits on how many queries can be sent per minute

According to Google AI for Developers, a token is equivalent to about 4 characters in the English language. We used the Gemini 2.0 Flash model, which has a limit of 8,192 tokens per query. To ensure that none of our queries exceeded the limit, we counted the tokens in each chunk and printed a warning if any chunk went over. If any lines exceeded the token limit, the document was further split so that fewer tokens were processed per request.

After counting tokens, we also restricted the number of queries sent per minute. The Gemini 2.0 Flash model has a limit of 15 queries per minute, or 1,500 queries per day. Therefore, we can only send 1 query to Gemini every 4 seconds. Our code introduces a 6 second delay between each request to stay within this limit.

2) Prompt Engineering: The next step involves prompt design / engineering, the process of writing prompts that invoke accurate and interpretable responses from an LLM.

We did this by giving clear and specific instructions, constraints on how to format output, and defining what a policy is.

Consider the prompt we engineered for Atascadero's Community Wildfire Protection Plan (CWPP):

```
Extract both explicit and implicit policies from this text. A policy can be a rule,
guideline, or a recommended action. Provide the exact wording.
```

This prompt includes clear instructions to extract implicit and explicit policies, maximizing the number of policies pulled. Then it provides a definition for a policy so that the LLM can optimize its search. We also included a request to preserve the original wording so that the extracted policies were not paraphrased.

3) Querying: Now we can query Gemini by prompting the LLM to extract policies for each chunk in the document. As mentioned earlier, we kept a 6 second delay between requests to stay within the rate limits for Gemini's free tier.

4) Saving the Results: Finally, we saved the responses to a CSV file to be manually checked for accuracy. Currently, the process of cleaning the responses is not automated; this can be done in Excel or by manipulating a dataframe in Python.

To clarify how this process works, consider an unstructured document like Temecula's Quality of Life Master Plan.

This plan has policies listed out in lines or bullet points, with inconsistent headings for each section. We split the text into chunks of 300 words or less, wrote a prompt to extract policies, queried Gemini, and saved the responses to a CSV file. The prompt used for this plan is identical to the one outlined earlier in Step 3.

V. Policy Categorization: Latent Dirichlet Allocation

A. Motivation

After extracting policies from the planning documents, the next challenge was categorizing them into meaningful topics. Given the large volume of extracted text, manually sorting policies into relevant themes would have been inefficient and subjective. To address this, we implemented LDA, a topic modeling technique that helps uncover hidden themes in a collection of documents. LDA assumes that each document consists of a mixture of topics, where a distribution of words represents each topic. Once we obtained the distribution of words for each topic, we manually assigned labels to the topics based on their most relevant keywords, ensuring that they aligned with the themes of wildfire resilience. This additional step helped refine the automated categorization and improve interpretability. By leveraging this probabilistic approach, we enabled a more systematic classification of policies, making it easier for city planners to analyze and interpret wildfire-related policies. We evaluated topic quality using specificity, computed against the PIRS scorecard as a ground truth for policy relevance. For each topic, specificity reflects its ability to exclude irrelevant policies or policies not identified as wildfire-resilient by our client. High-specificity topics are more precisely aligned with wildfire-related policies, whereas low-specificity topics tend to be overly general. This metric enabled us to identify which LDA topics best captured the core policy themes most relevant to wildfire.

While our dataset included eight documents—the Atascadero General Plan, the SLO County Multi-Jurisdictional Hazard Mitigation Plan, the Community Wildfire Protection Plan (CWPP), and the Atascadero Final Climate Action Plan (CAP) and the same four documents for Napa Valley—we primarily focused our LDA analysis on the Atascadero General Plan. This

Fig. 2: Screenshot of Temecula's Quality of Life Master Plan showing the unstructured policy text.

document contained well-structured policies with dense content, making it an ideal starting point for topic modeling before expanding to the other documents.

B. Model Developments and Refinements

Our first LDA model treated each policy as a separate document, allowing us to analyze how individual policies aligned with different topics. We initially set the number of topics to 15, as this provided the best alignment with the PIRS scorecard. This approach offered valuable insights but also revealed certain limitations. Due to the density of the policies, which contained multiple programs and action items, the model sometimes failed to assign a dominant topic to a given policy. Instead, it distributed low probability values across multiple topics, with some policies receiving probabilities as low as 0.067 across all 15 topics, making classification less meaningful. To improve topic assignment, we adopted a more granular approach by treating each program within a policy as a separate document. This method allowed LDA to work with more fine-grained text units, capturing important distinctions between different action items and ensuring a more precise topic classification. Specifically, we segmented policies into individual programs and increased our number of topics to 16 to account for the additional documents created by this segmentation. After refining our approach, the updated LDA model was able to map relevant topics to each policy/program more effectively. Each policy/program had a clear, dominant topic with a high probability.

C. Choosing the Number of Topics

Selecting the optimal number of topics is one of the most crucial steps in the LDA process, as it directly influences the interpretability and usefulness of the resulting topics. For this specific document, we set the number of topics to 16 after extensive human evaluation. First, we attempted to use perplexity score, where perplexity measures how well a probability model predicts, and the elbow curve, similar to k-means, to determine the optimal number of topics. However, this approach proved to be unhelpful as the rate of change decreased, or "elbow" would only occur at large values, such as 50 topics. When we ran the LDA model with the said 50 topics, we obtained many repeated keyword sets and topics; therefore, we decided against the perplexity score as a tool to determine the number of topics.

Instead, we used a starting point of 15 topics, and we would decrease the number of topics when keyword sets started to repeat or overlap across topics. This signified that the model was oversplitting a specific topic, producing repetitive themes, and reducing the clarity of the topics. On the other hand, we would increase the number of topics when keywords were vague, and policies grouped under a certain topic point to many different themes, such as transportation and housing, which are completely unrelated in this context. One of the obvious indicators for increasing the number of topics occurred when keywords such as shall, county, and plan were in the same topic. These are general keywords that every policy document contains, and therefore

indicate the need to increase the number of topics. This process was highly effective for the various documents we analyzed throughout this project; however, it was a very time-consuming task as it required constant human intervention.

D. Results & Analysis

Following the implementation of LDA with both 15 and 16 topics, we evaluated the extracted topics and their probability distributions to measure the model's effectiveness. Below is a list of the topics identified for each model, with the most relevant topics to the PIRS scorecard highlighted in gray:

TABLE I: Topics Identified in the 15-Topic LDA Model

Topic	Topic Category
1	Community Development & Risk Management
2	Environmental Hazards & Anti-Discrimination Policies
3	Urban & Residential Zoning
4	Downtown Planning & Design Standards
5	Historic Preservation & Zoning
6	Natural Resource Conservation & Emergency Management
7	Housing Development & Smart Growth
8	Parks, Trails, & Public Spaces
9	Tourism & Real Estate Policy
10	Agricultural & Geologic Land Use
11	Park & Water Resource Planning
12	Infrastructure & Public Facilities
13	Transit & Disaster Preparedness
14	Waste Management & Public Services
15	Mixed-Use & Commercial Development

TABLE II: Topics Identified in the 16-Topic LDA Model

Topic	Topic Category
1	Building Accessibility & Design Regulations
2	Housing Policy & Neighborhood Compatibility
3	Residential & Mixed-Use Development
4	Land Use & Public Facilities Planning
5	Environmental & Archaeological Considerations in Development
6	Transportation & Traffic Management
7	Zoning & Aesthetic Standards
8	Noise Regulations & Land Use
9	Master Planning & Infrastructure Development
10	Economic Development & Housing Goals
11	Emergency Preparedness & Land Use
12	Open Space & Recreation Planning
13	Historic Preservation & Community Identity
14	Education & Sustainable Growth
15	Transportation & Parking Standards
16	Environmental Conservation & Emergency Response

We also analyzed the probability values assigned to each policy/program to determine how well the model captured dominant themes. In the 15-topic model, some policies were spread too thin across multiple topics, leading to lower probability values (e.g., 0.067) across all topics, making classification less distinct. However, with the transition to 16 topics, each policy/program was more strongly associated with a single dominant topic, with probability values ranging from 0.4 to 0.8, leading to clearer topic assignments.

To illustrate the impact of our refinements, we include a comparison showing how the LDA model initially struggled to identify relevant policies but improved after segmentation and increasing the number of topics. Table VI in Appendix presents a before-and-after visualization of how the model mapped policies/programs to topics in the PIRS scorecard. The document column only contains policies/programs in the PIRS scorecard given to us by the client.

The 15-topic model struggled to confidently assign relevant topics to several policies, often labeling them as "None" with low probability values (e.g., 0.067 across all topics). This indicates that the model was unable to differentiate distinct themes within the dense policy text.

However, the 16-topic model demonstrates a significant improvement in topic classification. For example:

- **Policy 1.1 (Page II-13** was previously unclassified in the 15-topic model but was correctly identified under "Noise Regulations & Land Use" with a **probability of 0.83** in the refined model.
- **Policy 5.1 (Page II-27**, which addresses multi-family density and slope restrictions, initially lacked a meaningful topic assignment. However, the updated model confidently categorized it under "Environmental Conservation & Emergency Response" with a **probability of 0.78**.
- **Policy 1.4 (Page III-28**, relating to city street design, saw an increase in topic clarity, being mapped to "Master Planning & Infrastructure Development" with **0.393 probability** in the improved model.

As mentioned earlier, to evaluate how effectively our LDA model aligned with wildfire-related policies, we validated topic assignments using the PIRS scorecard as a ground truth. Since our ground truth was very small (66 PIRS policies vs 558 overall policies), we decided to compute specificity values where each LDA topic was treated as a binary classifier to measure how well each topic excluded policies that were not relevant according to the PIRS scorecard. The specificity for the top six most populated topics is shown below in the table. A high specificity score indicates that the topic is focused and does not dominate policies unrelated to wildfire-related policies. Conversely, a low specificity score suggests that the topic frequently dominates irrelevant or noisy policies, making it thematically ambiguous. We see that our specificities for these six topics are rather high which means that the LDA model is effectively isolating wildfire-relevant policies.

TABLE III: Specificity Scores for Selected LDA Topics

Topic	Specificity
Environmental Conservation & Emergency Response	0.932
Open Space & Recreation Planning	0.923
Historic Preservation & Community Identity	0.923
Residential & Mixed-Use Development	0.921
Noise Regulations & Land Use	0.914
Zoning & Aesthetic Standards	0.901

We chose to analyze the six topics containing the most policies (around 49% with 271 policies out of the overall 558) because these are the topics where we would expect the majority of the PIRS policies to be located. Overall, 52% (34 out of 66 total policies) of our labeled policies or PIRS policies were found in these six topics, while the rest were sprinkled throughout the other ten topics.

These results highlight the impact of expanding the topic space and segmenting policies into individual programs. Moving forward, we would like to enhance the efficiency of our analysis because there is still a need for manual verification of the LDA outputs.

VI. Spatial Tagging: Named Entity Approach & LLM Prompting

Named Entity Recognition (NER) is a natural language processing (NLP) technique used to identify and classify key information (entities) in text into predefined categories such as names of people, organizations, locations, dates, and more. In our project, we apply NER for *spatial tagging*—extracting spatially relevant terms from policy documents to support geographic analysis and planning.

A. Motivation

We selected NER as our primary approach because many existing NER frameworks already provide built-in support for recognizing geographical entities such as cities, countries, and locations. From this base, we can extend the models by adding domain-specific entities relevant to planning and zoning, making NER a natural fit for our spatial text analysis needs.

B. Model Developments and Refinements

We primarily focused on the `spaCy` NLP library.

1) spaCy – Built-in Model: We first tested spaCy's built-in NER model to evaluate its baseline performance on policy documents. From spaCy's recognized entity types, we focused on those relevant to spatial tagging:

- **GPE (Geo-Political Entity):** Cities, countries, regions
- **LOC (Location):** Non-political locations such as mountains or rivers
- **ORG (Organization):** Named agencies or institutions
- **FAC (Facility):** Buildings and infrastructure

- **MISC (Miscellaneous):** Other relevant, uncategorized terms

TABLE IV: Example Built-in NER Outputs from Policy Text

Policy Text
Policy AG/LU-43: Lands along the west bank of the Napa River south of the City of Napa and specific urban areas within four miles of the high water mark of Lake Berryessa are appropriate areas for marine commercial zoning and development.
Policy AG/LU-44: For parcels fronting upon the west side of the Napa River south of the City of Napa which are designated Agriculture, Watershed, and Open Space or Agricultural Resource on the Land Use Map of this General Plan which have commercial zoning, additional commercial development will be allowed as follows: All existing commercial establishments that are currently located within a commercial zoning district shall be allowed to continue to operate and use the existing buildings and/or facilities. Additional commercial uses which are permitted by the existing commercial zoning of the parcel shall be permitted on that portion of the parcel zoned commercial.

Legend:
GPE (Geo-Political Entity) LOC (Location) ORG (Organization) FAC (Facility)

The NER model seems to do well, however certain spatial phrases like *"south of the City of Napa," "four miles of the high water mark of Lake Berryessa,"* and terms like *"commercial zone"* are not consistently identified.

The built-in spaCy model identified many relevant spatial entities but missed domain-specific spatial terms like zoning classifications, land use types, area measurements, and housing categories, which are critical for our analysis. We compared each policy's pulled entities from the NER models to the corresponding entities present in the PIRS scorecard. However, while there was some overlap, we noticed that the NER model severely lacked the spatial terms that we expected to see.

2) spaCy – Custom NER Model: To better capture domain-specific spatial terms, we created a custom NER model by extending entity types and manually labeling a dataset that was tailored for us.

So, we decided to add these entities:

- **AGRICULTURE:** Agricultural areas, practices, or terms
- **RESIDENTIAL:** Residential designations, zones, or related terms
- **OPEN_SPACE:** Parks, undeveloped land, or designated open space
- **ZONING:** Zoning classifications or codes
- **LANDUSE:** General land use categories (e.g., agricultural, residential, industrial)
- **FACILITY_STATUS:** Operational or functional status of facilities
- **HOUSING:** Terms related to housing types, development, or density
- **MAP_SOURCE:** References to maps, planning diagrams, or cartographic sources
- **PLANNING_AREA:** Named districts, planning zones, or defined geographic units

Note: The current dataset and entity definitions primarily focus on the agricultural element. This framework is still in progress and can be expanded to cover other elements as the project evolves.

To prepare the dataset, we used the scorecard policies as a reference. Many policies already included underlined terms indicating potentially relevant spatial terms, which was helpful. However, due to inconsistent labeling of spatial phrases, we found it necessary to define our own set of entities.

Below is an example of our dataset, provided in the JSON format required for Named Entity Recognition (NER).

```
{
  "training_data": [
    {
      "text": "Policy AG/LU-1: Agriculture and related activities are the primary land uses in
          ↪ Napa County",
      "entities": [
        { "start": 16, "end": 50, "label": "AGRICULTURE" },
        { "start": 77, "end": 91, "label": "GEO" }
      ]
    },
    {
```

```
    "text": "Policy AG/LU-3: The Countys planning concepts and zoning standards shall be
        ↪ designed to minimize conflicts arising from encroachment of urban uses into
        ↪ agricultural areas. Land in proximity to existing urbanized areas currently in mixed
        ↪  agricultural and rural residential uses will be treated as buffer areas and further
        ↪  parcelization of these areas will be discouraged.",
    "entities": [
      { "start": 119, "end": 131, "label": "RESIDENTIAL" },
      { "start": 137, "end": 156, "label": "AGRICULTURE" },
      { "start": 173, "end": 204, "label": "RESIDENTIAL" },
      { "start": 218, "end": 242, "label": "AGRICULTURE" },
      { "start": 247, "end": 271, "label": "RESIDENTIAL" }
    ]
   },
   {
    "text": "Policy AG/LU-4: The County will reserve agricultural lands for agricultural use
        ↪ including lands used for grazing and watershed/open space, except for those lands
        ↪ which are shown on the Land Use Map as planned for urban development.",
    "entities": [
      { "start": 33, "end": 52, "label": "AGRICULTURE" },
      { "start": 57, "end": 77, "label": "AGRICULTURE" },
      { "start": 96, "end": 111, "label": "AGRICULTURE" },
      { "start": 116, "end": 138, "label": "OPEN_SPACE" },
      { "start": 199, "end": 217, "label": "RESIDENTIAL" }
    ]
   }
  ]
}
```

This snippet from the dataset shows how entities are labeled NER. In the first policy, the phrase “Agriculture and related activities” is tagged as AGRICULTURE, while “in Napa County” is tagged as a GEO (geographical location).

C. LLM Prompting with Gemini

To complement our NER pipeline, we explored prompt-based entity extraction using Gemini, a large language model. Unlike traditional NER models, Gemini can extract domain-specific terms with minimal setup by providing well-crafted instructions.

This approach is very similar to the Querying LLMs section stated above in this report. This replicates the prompt engineering process (with a different prompt, geared towards extracted spatial terms). Here is an example of a prompt.

```
prompt = f"""
   Extract the following types of spatial (mappable) terms from the policy text:
   1. Place names (either general, like "downtown", or specific, like "Angwin").
   2. Land use or zoning classifications.
   3. Geographical features (e.g., creeks, mountains, rivers).
   4. Structures, including facilities, buildings, and infrastructure.
   5. Mappable units of measure (e.g., distances, areas such as acres, hectares, square miles,
       ↪ parcels, buffers).
   6. Geospatial terms (e.g., raster, point, polygon, line, or file types such as GeoJSON, .shp
       ↪ , etc.).
   The response should be a comma-separated list of these terms from the following policy text
       ↪ :\n\n{policy_text}
"""
```

Using the same steps of querying the LLM, we have promisable results from the model:

TABLE V: Example Spatial Terms Highlighted from Policy Text

Policy Text
Policy AG/LU-43: Lands along the west bank of the Napa River south of the City of Napa and specific urban areas within four miles of the high water mark of Lake Berryessa are appropriate areas for marine commercial zoning and development.
Policy AG/LU-44: For parcels fronting upon the west side of the Napa River south of the City of Napa which are designated Agriculture , Watershed , and Open Space or Agricultural Resource on the Land Use Map of this General Plan which have commercial zoning , additional commercial development will be allowed as follows: All existing commercial establishments that are currently located within a commercial zoning district shall be allowed to continue to operate and use the existing buildings and/or facilities . Additional commercial uses which are permitted by the existing commercial zoning of the parcel shall be permitted on that portion of the parcel zoned commercial .

The LLM handles most missed terms that NER had missed. This suggests the LLM is better suited for capturing nuanced, context-dependent geographic and zoning-related expressions that traditional NER may overlook.

VII. Future Work & Conclusion

For our future work, we plan to improve the NER and Gemini models, expand the dataset, and continue with the text-to-mapping processing with GIS:

Model Ensembling: Combine the strengths of our custom NER and Gemini through ensembling techniques. This hybrid approach aims to boost accuracy and handle more complex or ambiguous spatial terms.

Expanded Entity Set & Dataset: Continue expanding the entity taxonomy to cover additional planning domains, such as fire, land use, safety element, etc. At the same time, grow and diversify the labeled policy dataset by including the rest of the policies from other plans.

Text-to-GIS Mapping: Develop a pipeline to connect extracted spatial entities directly to geographic features. This involves:

- Mapping NER outputs to GIS layers (e.g., zoning districts, land use boundaries)
- Implementing geocoding for named places and address-based entities
- Exploring spatial ontologies or standardized vocabularies to support reliable mapping.

Ultimately, the goal is to enable workflows that move from policy text $\rightarrow$ structured entities $\rightarrow$ mapped outputs, unlocking powerful new capabilities for planning and analysis.

Appendix

TABLE VI: Topic Assignments and Probability Distributions for Policies & Programs Between the 15-Topic and 16-Topic LDA Models

Page	Policy	Topic (15 Model)	Probability (15 Model)	Topic (16 Model)	Probability (16 Model)
II-13	Policy 1.1: Preserve the rural atmosphere of the community and assure "elbow room" in areas designated for lower density development by guiding new development into the Urban Core to conform to the historic Colony land use patterns of the City and to respect the natural environment, hillside areas, and existing neighborhoods.	None	0.067	Noise Regulations & Land Use	0.83
II-13	[1.1]2: Concentrate higher-density development downtown and within the Urban Core, and focus master-planned commercial uses at distinct nodes along arterial corridors.	None	0.067	Master Planning & Infrastructure Development	0.436
II-27	Policy 5.1: Reduce multi-family densities and increase single-family lot sizes as site slope increases.	None	0.067	Environmental Conservation & Emergency Response	0.78
II-30	[6.4]4: Utilize the Secretary of the Interior's Standards and Guidelines for Rehabilitating Historic Properties to assess proposed improvements to historic properties.	None	0.067	Housing Policy & Neighborhood Compatibility	0.774
III-28	Policy 1.4: Preserve the winding, tree-lined nature of the city street system in hillside areas. Programs: hillsides to preserve rural character and help limit vehicle speed.	None	0.067	Master Planning & Infrastructure Development	0.393

ACKNOWLEDGMENT

We sincerely thank Bill Siembieda, Margot McDonald, and Andrew Fricker for their guidance and support throughout this project. Their insights into wildfire resilience and policy analysis have been invaluable in shaping our approach.

REFERENCES

[1] City of Atascadero, *Atascadero General Plan 2025* [PDF]. Atascadero Planning Department, 2025. [Online]. Available: https://www.atascadero.org/sites/default/files/2023-06/Atascadero%20GP%202025.pdf
[2] San Luis Obispo County, *San Luis Obispo County Multi-Jurisdictional Hazard Mitigation Plan*, 2019. [Online]. Available: https://www.slocounty.ca.gov/Departments/Planning-Building/Forms-Documents/Plans-and-Elements/Elements/Local-Hazard-Mitigation-Plan/San-Luis-Obispo-County-Multi-Jurisdictional-Hazard.pdf
[3] City of Atascadero, *Community Wildfire Protection Plan*, 2019. Unpublished manuscript, available via personal archive.
[4] City of Atascadero, *Final Climate Action Plan*, 2014. Unpublished manuscript, available via personal archive.
[5] County of Napa, *Napa County General Plan*. [Online]. Available: https://www.countyofnapa.org/1760/General-Plan
[6] County of Santa Barbara, *Eastern Goleta Valley Community Plan*. Unpublished manuscript, available via personal archive.
[7] City of Temecula, *Quality of Life Master Plan*. Unpublished manuscript, available via personal archive.
[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Jan. 2003.
[9] D. Buscaldi and P. Rosso, "Named entity recognition for geographical information retrieval," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 189–198.