# The GELATO Dataset for Legislative NER

**Matthew Flynn[1], Timothy Obiso[1], Sam Newman[1,*]**
[1]Brandeis University
{matthewflynn, timothyobiso}@brandeis.edu, snewman.aa@gmail.com

**Abstract**

This paper introduces GELATO (Government, Executive, Legislative, and Treaty Ontology), a dataset of U.S. House and Senate bills from the 118th Congress annotated using a novel two-level named entity recognition ontology designed for U.S. legislative texts. We fine-tune transformer-based models (BERT, RoBERTa) of different architectures and sizes on this dataset for first-level prediction. We then use LLMs with optimized prompts to complete the second level prediction. The strong performance of RoBERTa and relatively weak performance of BERT models, as well as the application of LLMs as second-level predictors, support future research in legislative NER or downstream tasks using these model combinations as extraction tools.

**Keywords:** NER, Dataset, Ontology

## 1. Introduction

The automatic extraction of named entities in the legislative domain is important for supporting various types of research and more advanced NLP methods and applications. Most work done in NER has focused on the more general newswire domain; however, texts from other domains provide an interesting challenge due to differences in mention or text density, specialized terminology, and necessary real-world knowledge.

To this end, we introduce the Government, Executive, Legislative, and Treaty Ontology (GELATO), a two-level NER ontology designed for U.S. congressional legislation. We release an annotated dataset of 131 U.S. House and Senate bills using the GELATO tag set. Our work provides the baseline for an automatic entity extraction pipeline as new bills are released in the U.S. Congress.

We fine-tune and release several state-of-the-art transformer-based models on the GELATO dataset for first-level predictions. These predictions are then forwarded with their context to LLMs for second-level prediction, and we release optimized modules for this step. We compute standard NER performance metrics (F1) for both level-one and level-two predictions on our test data.

Our contributions are as follows:

- We introduce GELATO, a two-level NER ontology
- We release an annotated dataset using GELATO
- We fine-tune transformer models on GELATO data for first-level prediction and to provide baselines for future improvement
- We propose methods using LLMs to predict second-level NER labels
- We release optimized modules for second-level prediction with LLMs

All code and data is publicly available on our Github[1].

## 2. Related Work

Previous research in natural language processing on documents in the legal domain includes domain-specific datasets, benchmarks, and models, many of which are summarized by Küçük and Can (2025). Of particular note is Legal Knowledge Interchange Format (LKIF) Hoekstra et al. (2007, 2009), an extension of the Knowledge Interchange Format (KIF), a computer language for interacting with knowledge bases designed for the legal domain. This ontology can be used with relational logic systems to represent, compare, and contrast legal systems. LKIF provides a rich framework for representing legal knowledge broadly, though it was not designed with NER annotation in mind. GELATO draws on LKIF's conceptual structure but adapts it into a two-level annotation scheme suitable for sequence labeling.

While NER in the legislative domain is underresearched, much work has been done in the adjacent legal domain. Cardellino et al. (2017a,b) create a high-level ontology as a superclass to LKIF, which is a superclass to YAGO (Suchanek et al., 2007, 2024). There are also NER datasets for the legal domain in American English (Au et al., 2022), Brazilian Portuguese (Luz de Araujo et al., 2018), German (Leitner et al., 2020), Indian English (Kalamkar et al., 2022), and Chinese (Dai et al., 2025). These datasets primarily address

*Independent researcher. Work completed while at Brandeis University.

[1]https://github.com/Wollaston/gelato



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 7163–7177
11–16 May 2026. 

court judgments and case law, and their ontologies reflect this focus, including Location and Time/Date classes relevant to judicial proceedings. By contrast, GELATO targets congressional legislation and introduces categories for legislative constructs such as Acts, Funds, Programs, and Protected Classes.

Outside of the legal domain, the 2002 and 2003 CoNLL shared-tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) are standard datasets for NER evaluation in Spanish, Dutch, and English. CoNLL's flat ontology of Person, Organization, Location, and Miscellaneous has proven effective for newswire text but was not designed for specialized domains where entities like statutory references, government programs, and classes of people are central. GELATO extends this paradigm with a two-level ontology made of six top-level classes and 30 subclasses tailored to U.S. legislative text.

| Annotators 1 | 2 | 3 | House Bills | Senate Bills |
|---|---|---|---|---|
| ✓ | ✓ | | 141-170 | 169, 174, 176, 183, 187, 193, 196, 198, 199, 200 |
| ✓ | | ✓ | 171-200 | 102, 104, 107, 109, 112, 117, 121, 126, 128, 143 |
| | ✓ | ✓ | 100-140 | 144, 145, 148, 158, 160, 163, 164, 165, 167, 168 |

Table 1: Bills annotated by each annotator

## 3. Data and Annotation

We compiled and annotated 131 U.S. House and Senate bills using the GELATO tag set. The training ($n = 80$) and dev ($n = 21$) sets are made up of House bills; and the test set contains only Senate bills ($n = 30$). All bills were pulled from the 118th Congress with bill numbers 100-200.

All House bills in the 100–200 range were annotated, along with 30 Senate bills from the same range selected for their optimal length (3,000–9,000 characters). As detailed in Table 1, each annotator processed approximately 80 bills: 60 from the House and 20 from the Senate. Pairwise F1 scores (Table 2) demonstrate high inter-annotator agreement (IAA). While some disagreement is expected given the ontology's complexity, these scores confirm the clarity of our guidelines and the stability of the underlying framework.

| Annotators 1 | 2 | 3 | IO F1 | Level 1 F1 |
|---|---|---|---|---|
| ✓ | ✓ | | 66.35 | 62.53 |
| ✓ | | ✓ | 77.27 | 73.50 |
| | ✓ | ✓ | 80.09 | 79.68 |

Table 2: Inter-Annotator Agreement (F1)

### 3.1. Dataset

The dataset consists of legislative bill text from the House of Representatives and the Senate of the 118th Congress. All data was queried from the Congress.gov API[2] and tokenized. As legislative text is punctuation-dense due to the inclusion of many citations, we split each punctuation mark into its own token.

The training set consists of House bills ($n = 80$) with a total of 77,372 tokens and a mean of 967 tokens per bill. The longest bill contains a total of 10,653 tokens, while the shortest contains 240.

The dev set also consists of House bills ($n = 21$) with a total of 23,819 tokens and a mean of 1,134 tokens per bill. The longest bill contains a total of 3,859 tokens, while the shortest contains 205.

The test set contains Senate bills ($n = 30$) of more consistent length with a total of 31,740 and a mean of about 1,058 tokens per bill. The longest Senate bill contains 2,005, while the shortest contains 641.

Table 3 summarizes the GELATO dataset. Table 13 provides mention counts for both levels. These counts were obtained via SeqScore (Palen-Michel et al., 2021; Lignos et al., 2023), an evaluation toolkit for sequence labeling tasks. Additionally, all tags and label transitions were validated using SeqScore.

### 3.2. Ontology

The GELATO entity type ontology has two levels. The top level has six tags: PERSON, ORGANIZATION, DOCUMENT, ACT, ABSTRACTION, and CLASS. PERSON and ORGANIZATION are borrowed from the prototypical CoNLL ontology, though their subclasses are significantly informed by the legislative domain. The other four classes were largely derived from analyzing sample bills through the lens of the LKIF Ontology. All subclasses are informed by the clustering evaluation of these top-level annotations.

[2] api.congress.gov

| | Train | Dev | Test | Total |
|---|---|---|---|---|
| Bills | 80 | 21 | 30 | 131 |
| Tokens | 77,372 | 23,819 | 31,740 | 132,931 |
| Unique Mentions | 1,108 | 448 | 556 | 2,112 |
| Total Mentions | 3,388 | 1,029 | 1,393 | 5,810 |

Table 3: GELATO Summary

#### 3.2.1. Person

The top-level PERSON class has three subclasses based on the bill domain: MEMBER **MBC**, TITLE **TTL**, and INDIVIDUAL **IND**. A MEMBER is a member of Congress and is almost always listed as the one introducing a bill or sponsoring a bill. They are present in every bill, appear in a distinct context, and serve a downstream interest. A TITLE is a mention of a person in a specific position, such as *Speaker* or *President*. An INDIVIDUAL is a PERSON mention that falls into neither of the above categories.

(1) January 9, 2023 Mr. Biggs MBC introduced the following bill

(2) for a period to be subsequently determined by the Speaker TTL

#### 3.2.2. Organization

The top level ORGANIZATION class has eight subclasses based on the bill domain: NATION **NAT**, STATE **STA**, LOCALITY **LOC**, COMMITTEE **COM**, AGENCY **AGC**, LEGISLATIVE BODY **LEG**, INTERNATIONAL INSTITUTION **INT**, and ASSOCIATION **ASC**. NATION, STATE, and LOCALITY are the three levels of geopolitical entities covered in the ORG class. The NATION class is named as such rather than *country* to avoid issues of official UN recognition and to include American Indian nations, etc. (3) is an example of a state.

The classes LEGISLATIVE BODY and COMMITTEE cover a fairly static set of entities, where **LEG** mostly encompasses {*Senate, House of Representatives, Congress*} and **COM** covers their committees, as demonstrated in (4) and (5). AGENCY and INTERNATIONAL INSTITUTION encompass mentions of U.S. government agencies, (*departments, bureaus, administrations,* etc.) and mentions of international organizations like the *United Nations*, respectively. ASSOCIATION is the **ORG** class's catch-all subclass for capturing mentions like *Planned Parenthood*, etc.

(3) Mr. Biggs MBC , Mr. Nehls MBC , Mrs. Miller MBC of Illinois STA

(4) Be it enacted by the Senate LEG and House of Representatives LEG of the United States of America NAT

(5) which was referred to the Committee on Ways and Means COM

#### 3.2.3. Document

The DOCUMENT class takes more inspiration from the LKIF ontology and its mereological and dependency relationships. It encompasses those physical entities on which ACTS depend. The top level DOCUMENT class has six subclasses based on the bill domain: CODE **CODE**, BILL **BILL**, REFERENCE **REF**, PARENTHETICAL **PAR**, TREATY **TRE**, and REPORT **REP**. CODE covers the foundational documents of United States law. This includes documents such as the *United States Code*, which is frequently referenced and amended in legislation. The BILL class exists almost exclusively for capturing the extremely consistent mentions of bill numbers, as demonstrated in (6).

The REFERENCE and PARENTHETICAL classes capture the much more varied mentions of specific *titles*, *sections*, *paragraphs*, etc. that are directly referenced and amended in most bills, as demonstrated in (7). The REPORT class is the somewhat catch-all subclass of DOCUMENT, and includes mentions such as *President's Budget* and similar documents. The TREATY class is for treaty mentions, which did not show up in our dataset. We include this subclass in the GELATO ontology because treaties don't fit the concept of REPORT, which isn't a true catch-all, and are important to capture, though rare.

(6) [H.R. 189 BILL Introduced in House (IH)]

(7) statement pursuant to section 102 of the National Environmental Policy Act of 1969 REF (42 U.S.C. 4332) PAR

#### 3.2.4. Act

The ACT class depends on the DOCUMENT class, and encompasses two well-defined subclasses:

| Ontology | Locale | Person | Organization | Location | Legal Text | Time/Date | Other |
|---|---|---|---|---|---|---|---|
| GELATO | American English | ✓ | ✓ | | ✓ | | Abstraction, Act, Class |
| Cardellino *et al.* | American English | ✓ | ✓ | | ✓ | | Abstraction, Act |
| Leitner *et al.* | German | ✓ | ✓ | ✓ | ✓ | | Case-by-case Regulation |
| Kalamkar *et al.* | Indian English | ✓ | ✓ | ✓ | ✓ | ✓ | Court, Case Number |
| LeNER-Br | Brazilian Portuguese | ✓ | ✓ | ✓ | ✓ | ✓ | |
| E-NER | American English | ✓ | ✓ | ✓ | ✓ | | Court, Miscellaneous |

Table 4: Comparison of legal-domain NER ontologies across different papers and languages

PUBLIC ACT **PA** and AMENDMENT **AMD**. PUBLIC ACT mentions are those that refer to an act by name, and AMENDMENT mentions are those that refer to amendments to the constitution.

(8) prior to the enactment of the American Rescue Plan Act PA

(9) as guaranteed by the Second Amendment to the Constitution AMD

### 3.2.5. Abstraction

The ABSTRACTION class has 9 subclasses: 8 based on the legislative domain and a MISC subclass. The general goal of ABSTRACTION is to capture entities that are the result of legislative action. The 8 non-miscellaneous subclasses are: PROGRAM **PRO**, SESSION **SES**, SYSTEM **SYS**, INFRASTRUCTURE **INF**, FUND **FUND**, DOCTRINE **DTR**, SPECIFICATION **SPC**, and CASE **CASE**.

The PROGRAM subclass encompasses programs established and managed by the government, such as *Medicare* and *Social Security*. The SYSTEM and INFRASTRUCTURE subclasses cover the frameworks established and maintained by the government, often in support of a PROGRAM. The distinction lies in whether the entity is tangible, such as *International Outfall Interceptor*, for the latter, or not, such as the *National Forest System* in (10), or a database or website for the former. Mentions of the FUND subclass include those referring to the *General Fund of the Treasury* or other specific funds such as those in (12). The DOCTRINE subclass covers somewhat abstract policies, principles, and rights, such as *Free Speech* or the *Fairness Doctrine* in (13). The CASE subclass, like the TREATY subclass, is unpopulated in our dataset, but encompasses mentions of court cases such as *McCulloch v. Maryland*, which are referenced in bills when they establish relevant legal precedents or support the need for legislative changes. The SPECIFICATION subclass covers mentions of specific, named legal specifications and classifications like *Schedule I* and *340b* in (14).

(10) forest management activity conducted on National Forest System SYS land

(11) maintained in the National Firearms Registration and Transfer Record SYS

(12) held by the Old-Age and Survivors Insurance Trust Fund FUND

(13) present opposing viewpoints on controversial issues of public importance, commonly referred to as the 'Fairness Doctrine DTR'

(14) subject to an agreement under section 340B of the Public Health Service Act REF shall include the 340B SPC modifier established by the Secretary TTL

### 3.2.6. Class

The CLASS class refers to a class of people. The initial motivator for including the class was capturing mentions of protected classes qualified for special protection by law, i.e. {*Age, Ancestry, Color, Disability, Ethnicity, Gender, HIV/AIDS status, Military Status, National Origin, Pregnancy, Protected Veteran Status, Race, Religion, Sex, Sexual Orientation*}. For these mentions we created the subclass PROTECTED CLASS **PC**. When annotating the pilot, we found instances where other classes of people were a subject of a bill and consistently referenced

throughout (16), and we decided to add the subclass NON-PROTECTED CLASS **NPC** to capture their mentions.

We also started noticing repeated spans like *State* and *Committee* referenced throughout bills, leading us to recognize that GELATO lacked a type for classes of organization, which would complete the logical parallel (i.e. **PER** : **CLS** :: **ORG** : ____). However, we ultimately decided that the mentions that would fall within the class were not worth capturing, as their presence in a bill did not meaningfully differentiate that bill from another.

(15) processing of claims for temporary disability ratings for veterans **PC** described

(16) submit to Congress **LEG** a recommendation on the number of refugees **NPC** who may be admitted

Table 4 shows a comparison between GELATO and other ontologies for adjacent domains.

### 3.3. Annotation

The two levels of labels for each document and mention in GELATO were completed in two steps with three total annotators.

The first level of GELATO annotation for each document was completed by two annotators using Label Studio (Tkachenko et al., 2020-2025). All annotators shared the same configuration and Label Studio instance. After annotation, all three annotators participated in the adjudication process. All tokens with differing tags were discussed by the group of annotators and resolved through group consensus. After adjudication, the annotators proofread these adjudicated tags to obtain the final gold-standard first-level data, ensuring quality and consistency.

To obtain the second-level GELATO labels, the three annotators reviewed the gold labels together. For each gold label mention, the annotators decided which second-level label of the respective first-level label was most appropriate, and tagged it accordingly. Together, this two step process resulted in the complete and fully adjudicated two-level, gold-standard GELATO dataset.

## 4. Multi-level NER with LLMs

We propose a no-training approach to level-two NER label prediction that evokes the generalized knowledge LLMs have obtained during pretraining through optimized prompts that include specific instructions. There are many challenges when performing NER with LLMs, and there have been many ways proposed of doing so (Labrak et al., 2024; Subedi et al., 2024; Wang et al., 2025).

In this paper, we use DSPy (Khattab et al., 2022, 2024) to find optimized prompts. The base prompt for the LLM is generated with its level-one label, a context window of 50 tokens on each side (totalling 100), and the possible level-two labels. This includes input and output field descriptions, as well as a template structure to frame the response. In the user message, this template structure is filled in with the necessary context before the model is prompted to generate a structured response from which the level-two label can be extracted. We chose 50 tokens for the context window as it is reasonably sized for model performance while limiting the total number of tokens per prompt. With this optimized configuration, we prompt the LLM to classify the provided mention.

This approach allows us to take advantage of the strengths of LLMs and smaller, fine-tuned transformer language models. We use the smaller model to process every token and classify them according to a simpler ontology. Once there is a level-one label for a mention, we prompt an LLM with few-shot examples to perform level-two classification.

## 5. Experiments

### 5.1. Level One

We fine-tune a series of transformer language models on the GELATO dataset. These results are summarized in Table 5. Our results show the strengths and weaknesses of models of different sizes and pretraining strategies. We report test results after fine-tuning base and large models of BERT and RoBERTa (Devlin et al., 2019; Liu et al., 2019) on GELATO.

To obtain the optimal hyperparameters for each model, we conducted a sweep of 50 runs on Weights and Biases (Biewald, 2020) using Bayesian hyperparameter optimization (Snoek et al., 2012) over `learning_rate`, `batch_size`, `epochs`, `weight_decay`, and `warmup_ratio`. Learning rate was selected from a log uniform distribution from 1E-5 to 6E-4; batch size was varied between 1, 2, 4, 8, 16, 32, 64, and 128. Epochs between 1 and 50 were tested. Weight decay and warmup ratio were both varied between 0.0 and 0.5. The best set of hyperparameters for each model is included in Table 5.

### 5.2. Level Two

We obtain level-two labels using the labels predicted by the transformer models in level one. To find optimized prompts for each level-one label, we use the MiPROv2 optimizer (Opsahl-Ong et al., 2024) from DSPy with Qwen3/Qwen-32B (Team,

| Model | LR | BS | E | WD | WR | Lvl 1 F1 | Lvl 2 F1 |
|---|---|---|---|---|---|---|---|
| bert-base-cased | 3E-5 | 1 | 44 | 0.4 | 0.5 | 68.09 | 53.26 |
| bert-large-cased | 1E-5 | 1 | 34 | 0.0 | 0.2 | 67.16 | 53.14 |
| roberta-base | 1E-4 | 2 | 47 | 0.1 | 0.2 | 84.87 | 64.79 |
| roberta-large | 5E-5 | 1 | 48 | 0.5 | 0.1 | **88.72** | **67.44** |
| Gold Level One Labels | - | - | - | - | - | - | 76.60 |

Table 5: Best F1 scores and hyperparameter configuration for each model; LR = learning rate, BS = batch size, E = epochs, WD = weight decay, WR = warmup ratio

2025) as the inference model.[3]

For the prediction step, we also used Qwen/Qwen3-32B self-hosted with vLLM (Kwon et al., 2023). We loaded each optimized DSPy module and routed the prompting for the level-two label based on the level-one prediction. Because of this, if the level-one label was incorrect, the level-two label would be incorrect as well. The predicted level-two label was mapped back to its respective level-one label for alignment and scoring. The level-two results are summarized alongside the level-one results in Table 5, and a detailed breakdown of F1 by label for the best-performing `roberta-large` is presented in Table 6.

The F1 scores for the level-two predictions show a wide variance in performance by level-one and level-two label types. Models perform better with more concrete types that exist in specific contexts in the training data, such as BILL and TITLE, while struggling with more abstract or referential types, such as PARENTHETICAL. There could be some performance loss as well due to the relative contextual or semantic similarity of certain types, such as MEMBER and NAME.

## 6. Results

As expected, the larger RoBERTa models perform better at this task; the best-performing model is `roberta-large` with an F1 of 88.72. The BERT base-sized model performed the worst with an F1 of 68.09. Because of the timing of the 118th Congress (2023-2025), it is impossible that the exact text of any of these bills was included in the training data. U.S. Congress bills are publicly available so it cannot be concluded that there were no legislative bills or texts in any of the pretraining data; despite this, our results showcase the ability of transformer models to perform domain-aligned named entity extraction in a novel domain.

Additionally, legislative bills have many unique features such as their structure, format, references, style, etc. Documents of this format may not have been included in significant numbers in the pretraining of the language models we experiment with. As a result, the token sequences we are providing to the models are very unique and likely novel. In addition, we also train with fewer tokens than large-scale general NER datasets such as CoNLL. As such, the evaluation metrics from models trained for our task are lower.

For the level-two prediction task, in which the Qwen model makes fine-grained predictions based off of level-one labels, the F1 score is approximately 20 points lower for RoBERTa models, and approximately 14 to 15 points lower for BERT models, when compared to their level-one F1 counterpart. Due to the cascading nature of level one errors on level two predictions, we also report the performance of the Qwen model conditioned on the gold level one labels.

It is unclear if Qwen/Qwen3-32B's pretraining data included these specific bills or content from the 118th Congress, but it can be reasonably concluded that it has seen similar content. This is generally expected with LLMs and their vast pretraining corpora, and it is this knowledge and generalizability that makes using LLMs an interesting choice for second-level prediction in two-level ontologies.

## 7. Error Analysis

BERT models performed much worse than RoBERTa models. To investigate why, we first produced confusion matrices to gain better insight into what tags the models struggled with in general, and relative to one another. These are shown in Figure 1.

Of particular note is that all models performed poorly when classifying CLASS mentions. We believe this is due to the nuance required in order to correctly annotate and tag CLASS mentions, especially relative to a PERSON mention. This is likely due to the fact that many CLASS mentions are plural versions of words that would otherwise be tagged as PERSON if in their singular forms.

And while GELATO distinguishes between PERSON and CLASS mentions, which differ primarily in number, GELATO notably does not make the dis-

[3] https://huggingface.co/Qwen/Qwen3-32B

| Type | | F1 | Type | | F1 |
|---|---|---|---|---|---|
| Abstraction | | 67.80 | | Code | 5.26 |
| | Case | – | | Parenthetical | 13.33 |
| | Doctrine | 0 | | Reference | 61.54 |
| | Fund | 66.67 | | Report | 0 |
| | Infrastructure | 61.54 | | Treaty | – |
| | Misc | 0 | Organization | | 90.64 |
| | Program | 51.06 | | Agency | 88.37 |
| | Session | 100.00 | | Association | 31.11 |
| | Specification | – | | Committee | 91.25 |
| | System | 28.57 | | International Institution | 85.71 |
| Act | | 88.24 | | Legislative Body | 95.27 |
| | Amendment | 0 | | Locality | 50.00 |
| | Public Act | 89.31 | | Nation | 83.46 |
| Class | | 27.52 | | State | 69.39 |
| | Non-Protected Class | 9.76 | Person | | 96.65 |
| | Protected Class | 14.81 | | Member | 46.84 |
| Document | | 94.99 | | Name | 0 |
| | Bill | 95.24 | | Title | 93.22 |

Table 6: F1 scores for level-two predictions using roberta-large and Qwen/Qwen3-32B; – indicates the label was not present

tinction between ORGANIZATION and some group of organizations. Perhaps the models would have performed better if this distinction also exists to help better generalize over number.

The models in our experiments also differ in their vocabulary sizes and tokenizers. BERT has a vocabulary size of 30,522 and uses the WordPiece subword tokenization strategy (Schuster and Nakajima, 2012; Song et al., 2020). In contrast, RoBERTa has a vocabulary size of 50,265 and uses the byte version of Byte-Pair Encoding for tokenization (Sennrich et al., 2015). We believe that the differences in vocabulary size, as well as the different tokenization strategies, also contribute to the relative performance differences of these models.

As for LLMs, they have vast amounts of general knowledge from their pretraining, and this can impact performance when these general expectations do not align with those of domain-specific interests, like those of GELATO. For example, LLMs frequently confused the NAME subclass of the first-level PERSON class with the other MEMBER subclass. This can be expected confusion with LLMs as MEMBER is a specific type of name in GELATO. This suggests that when designing NER ontologies that LLMs will use, it is important to have clear semantic distinction to avoid LLM confusion.

Interestingly, in rare cases, the LLM will refuse to choose one of the provided level-two labels, noting that none of the options are satisfactory in its generated reasoning and leading to an automatic error. This occurred between one and seven times with level-two predictions for each of the four generated level-one prediction sets, and more frequently with level-one labels from the less-performant BERT models.

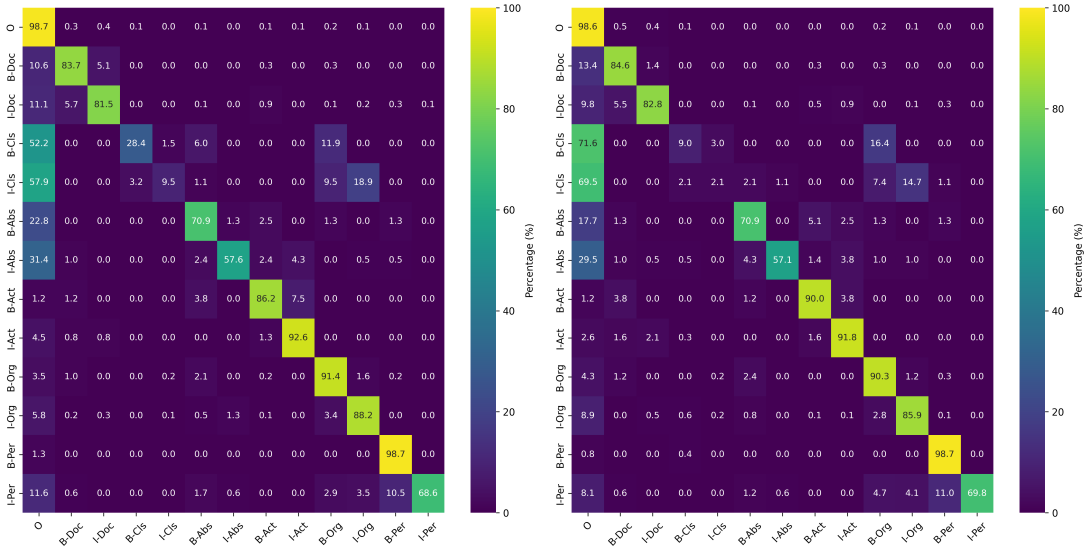
As well, with the two-level ontology and predictions, any error made in level one would cascade to level two. Therefore, it is expected that level-two F1 scores would lag behind their level-one counterparts, and that level-two predictions generated from RoBERTa predictions would strongly outperform those from BERT models. Any increase in performance for level-one predictions would therefore lead to a likely performance boost for level two as well. This is particularly notable for any misalignment errors in level one, as level two does not perform any BIO-encoding and instead relies on the encoding generated in level one.

## 8. Discussion

Our experiments on GELATO reveal a number of insights into the challenges present in legislative NLP, particularly NER.

The superior performances of the RoBERTa model emphasizes how a larger vocabulary size and more diverse pretraining data can greatly improve model performance on the same task for a model of a similar architecture.

The complete absence of CLASS in most model predictions indicates an open challenge in legislative NER. According to the confusion matrices in Figure 1, CLASS was most commonly predicted to be 'O'. The similarity between items in the PERSON and CLASS classes are subtle; future work can explore why and how models may need more training




(a) bert-base-cased

(b) bert-large-cased

(c) roberta-base

(d) roberta-large

Figure 1: Level one confusion matrices

examples containing CLASS to learn to differentiate it from general lexical items and PERSON.

Based on the fine-tuning results, it seems that transfer learning is not applicable here and misalignment between NER domains leads to decreased performance over a single-task training paradigm. It also seems that there is room for BERT-era models to improve and reach the performance of RoBERTa.

Although not explored in this work, this approach could also be used to perform NER on traditional CoNLL-style, single-level datasets. First, a smaller language model would be trained to not classify, but rather identify named entities. In other words instead of 2n+1 labels for n classes, you would only need three labels (BIO). Our approach would take these identified mentions and prompt an LLM to do the level 1 classification. One shortcoming of this approach is that it is dependent on the performance of a level-one classifier.

Overall, our results show that pre-trained models offer competitive performance on the GELATO dataset and can be reliably deployed for downstream tasks, in addition to future research. GELATO is a useful benchmark for investigating the performance of models at NER in the legislative domain.

## Limitations

The GELATO dataset is limited to House and Senate Bills 100-200 from the 118th Congress. This limits our dataset temporally and stylistically. This may hurt the performance of our models on bills

from previous or future congressional sessions, from different legislative chambers at the state level, or from other English-speaking countries.

While GELATO aims to comprehensively cover the most common entities in U.S. legislative documents, Treaty and Case have no mentions across our train, dev, and test sets which limits our ability to use this dataset to extract those entity types.

One limitation of our two-stage prediction pipeline is that errors made in stage one propagate through the pipeline and are irrecoverable in stage two. When performing prompt optimization, we find one optimized prompt for each level-one label, giving us six optimized prompts in total. Other designs of this system may involve a single prompt or even a prompt for each level-two label.

## Ethical Considerations

All bills in the GELATO dataset are publicly available U.S. government documents obtained via the Congress.gov API and are therefore in the public domain and not subject to copyright restrictions. We release our code and models under a permissive open-source license to encourage more research with our models and data.

Three graduate student annotators (the authors) with training in linguistics and NLP collaboratively created GELATO through a two-stage process with full adjudication of any disagreements. This is a descriptive annotation; for example, this ontology includes Protected Class and Non-Protected Class subclasses that are consistent with U.S. anti-discrimination law definitions.

GELATO can support beneficial applications including legislative tracking, policy analysis, and government transparency initiatives. However, automated entity extraction could also enable potentially harmful uses such as targeted analysis of how specific groups are referenced in legislation or identification of individual legislators for inappropriate purposes.

## 9. Acknowledgments

We thank Constantine Lignos for his helpful comments and insightful feedback on this paper. We also thank anonymous reviewers whose feedback helped to improve this paper.

## 10. Bibliographical References

Ting Wai Terence Au, Vasileios Lampos, and Ingemar Cox. 2022. E-NER — an annotated named entity recognition corpus of legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 246–255, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2017a. A low-cost, high-coverage legal named entity recognizer, classifier and linker. In *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law*, pages 9–18.

Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2017b. Legal NERC with ontologies, Wikipedia and curriculum learning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 254–259, Valencia, Spain. Association for Computational Linguistics.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2025. LAiW: A Chinese legal large language models benchmark. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10738–10766, Abu Dhabi, UAE. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rinke Hoekstra, Joost Breuker, Marcello Di Bello, and Alexander Boer. 2009. Lkif core: Principled ontology development for the legal domain. In *Law, ontologies and the semantic web*, pages 21–52. IOS Press.

Rinke Hoekstra, Joost Breuker, Marcello Di Bello, Alexander Boer, et al. 2007. The lkif core ontology of basic legal concepts. *LOAIT*, 321:43–63.

Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022. Named entity recognition in Indian court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2022*,

pages 184–193, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Dilek Küçük and Fazli Can. 2025. Computational law: Datasets, benchmarks, and ontologies.

Yanis Labrak, Mickael Rouvier, and Richard Dufour. 2024. A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2049–2066, Torino, Italia. ELRA and ICCL.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2020. A dataset of German legal documents for named entity recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4478–4485, Marseille, France. European Language Resources Association.

Constantine Lignos, Maya Kruse, and Andrew Rueda. 2023. Improving NER research workflows with SeqScore. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 147–152, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Pedro H. Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In *International Conference on the Computational Processing of Portuguese (PROPOR)*, Lecture Notes on Computer Science (LNCS), pages 313–323, Canela, RS, Brazil. Springer.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9340–9366, Miami, Florida, USA. Association for Computational Linguistics.

Chester Palen-Michel, Nolan Holley, and Constantine Lignos. 2021. SeqScore: Addressing barriers to reproducible named entity recognition evaluation. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 40–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2020. Fast wordpiece tokenization. *arXiv preprint arXiv:2012.15524*.

Bipesh Subedi, Sunil Regmi, Bal Krishna Bal, and Praveen Acharya. 2024. Exploring the potential of large language models (LLMs) for low-resource languages: A study on named-entity recognition (NER) and part-of-speech (POS) tagging for Nepali language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6974–6979, Torino, Italia. ELRA and ICCL.

Fabian M Suchanek, Mehwish Alam, Thomas Bonald, Lihu Chen, Pierre-Henri Paris, and Jules

Soria. 2024. Yago 4.5: A large and clean knowledge base with a rich taxonomy. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 131–140.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

Qwen Team. 2025. Qwen3 technical report.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

# A. Prompt Templates

Each entity type uses a shared system prompt structure with type-specific user prompts. We show the common system prompt once (§A.1), followed by representative user prompts for each category.

## A.1. Shared System Prompt

**System Prompt**

```
Your input fields are:
1. `mention` (str): the mention to extract the type
from
2. `context` (str): the context surrounding the
mention
3. `possible_tags` (list[str]): list of possible
level-2 tags

Your output fields are:
1. `reasoning` (str)
2. `tag` (str): the type of mention. MUST BE ONE OF THE
   POSSIBLE TAGS PROVIDED.

All interactions will be structured in the following
way, with the appropriate values filled in.

[[ ## mention ## ]]
{mention}

[[ ## context ## ]]
{context}

[[ ## possible_tags ## ]]
{possible_tags}

[[ ## reasoning ## ]]
{reasoning}

[[ ## tag ## ]]
{tag}

[[ ## completed ## ]]

In adhering to this structure, your objective is:
Extract contiguous tokens referring to members of
congress, titles, or simple names, if any, from a
list of string tokens. Output a list of tokens.
```

## A.2. Abstraction

**User Prompt**

```
[[ ## mention ## ]]
Children

[[ ## context ## ]]
2023 Mr . Cardin ( for himself and Ms . Stabenow )
introduced the following bill ; which was read twice
and referred to the Committee on Finance A BILL To
amend title XXI of the Social Security Act to prohibit
lifetime or annual limits on dental coverage under the
Children ' s Health Insurance Program , and to require
wraparound coverage of dental services for certain
children under such program . Be it enacted by the
Senate and House of Representatives of the United
States of America in Congress assembled , SECTION 1 .
SHORT TITLE . This Act may

[[ ## possible_tags ## ]]
["Doctrine", "Fund", "Infrastructure", "Misc",
 "Program", "Session", "Specification", "System"]

Respond with the corresponding output fields, starting
with the field [[ ## reasoning ## ]], then
[[ ## tag ## ]], and then ending with the marker for
[[ ## completed ## ]].
```

## A.3. Act

**User Prompt**

```
[[ ## mention ## ]]
Defending Domestic Produce Production Act of 2023

[[ ## context ## ]]
seasonal industries affected by antidumping or
countervailing duty investigations , and for other
purposes . Be it enacted by the Senate and House of
Representatives of the United States of America in
Congress assembled , SECTION 1 . SHORT TITLE . This
Act may be cited as the `` Defending Domestic Produce
Production Act of 2023 '' . SEC . 2 . DEFINITIONS .
( a ) Core Seasonal Industry . -- Section 771 of the
Tariff Act of 1930 ( 19 U.S.C. 1677 ) is amended by
adding at the end the following : `` ( 37

[[ ## possible_tags ## ]]
["Amendment", "PublicAct"]

Respond with the corresponding output fields, starting
with the field [[ ## reasoning ## ]], then
[[ ## tag ## ]], and then ending with the marker for
[[ ## completed ## ]].
```

### A.4. Class

**User Prompt**

```
[[ ## mention ## ]]
foster youth

[[ ## context ## ]]
Congress ] [ From the U.S. Government Publishing
Office ] [ S. 102 Introduced in ( IS ) ] 118th
CONGRESS 1st Session S. 102 To amend title IV of the
Social Security Act to establish a demonstration grant
program to provide emergency relief to foster youth
and improve pre-placement services offered by foster
care stabilization agencies , and for other purposes .
IN THE SENATE OF THE UNITED STATES January 26 , 2023
Mrs . Fischer ( for herself and Mr . Hickenlooper )
introduced the following bill ; which was read twice
and

[[ ## possible_tags ## ]]
["Non-ProtectedClass", "ProtectedClass"]

Respond with the corresponding output fields, starting
with the field [[ ## reasoning ## ]], then
[[ ## tag ## ]], and then ending with the marker for
[[ ## completed ## ]].
```

### A.5. Document

**User Prompt**

```
[[ ## mention ## ]]
S. 104

[[ ## context ## ]]
[ Congressional Bills 118th Congress ] [ From the
U.S. Government Publishing Office ] [ S. 104
Introduced in Senate ( IS ) ] 118th CONGRESS 1st
Session S. 104 To amend title VII of the Tariff Act
of 1930 to provide for the treatment of core seasonal
industries affected by antidumping or countervailing
duty investigations , and for other purposes . IN THE
SENATE OF

[[ ## possible_tags ## ]]
["Bill", "Code", "Parenthetical", "Reference",
 "Report"]

Respond with the corresponding output fields, starting
with the field [[ ## reasoning ## ]], then
[[ ## tag ## ]], and then ending with the marker for
[[ ## completed ## ]].
```

### A.6. Organization

**User Prompt**

```
[[ ## mention ## ]]
Senate

[[ ## context ## ]]
Committee on Finance A BILL To amend title IV of the
Social Security Act to establish a demonstration grant
program to provide emergency relief to foster youth
and improve pre-placement services offered by foster
care stabilization agencies , and for other purposes .
Be it enacted by the Senate and House of
Representatives of the United States of America in
Congress assembled , SECTION 1 . SHORT TITLE . This
Act may be cited as the `` Foster Care Stabilization
Act of 2023 '' . SEC . 2 . GRANTS TO IMPROVE PRE -
PLACEMENT SERVICES FOR

[[ ## possible_tags ## ]]
["Agency", "Association", "Committee",
 "InternationalInstitution", "LegislativeBody",
 "Locality", "Nation", "State"]

Respond with the corresponding output fields, starting
with the field [[ ## reasoning ## ]], then
[[ ## tag ## ]], and then ending with the marker for
[[ ## completed ## ]].
```

### A.7. Person

**User Prompt**

```
[[ ## mention ## ]]
Secretary

[[ ## context ## ]]
this subsection shall have 3 years to spend funds
awarded by the grant and return any unused grant funds
to the Secretary . `` ( 3 ) Application . -- A foster
care stabilization agency that desires to receive a
grant under this subsection shall submit to the
Secretary an application at such time , in such
manner , and containing such information as the
Secretary may require , that shall include the
following : `` ( A ) A description of how grant funds
will be used to provide emergency relief to foster
youth by the foster

[[ ## possible_tags ## ]]
["Member", "Name", "Title"]

Respond with the corresponding output fields, starting
with the field [[ ## reasoning ## ]], then
[[ ## tag ## ]], and then ending with the marker for
[[ ## completed ## ]].
```

## B. Annotation Guidelines

Our guidelines consist of a set of rules with examples as well as definitions for tags at both levels.

### B.1. Rules

1. No Nesting

   (a) A token may be part of one tag at maximum

   (b) A token should be part of the largest possible tag and only that tag

   | | |
   |---|---|
   | Secretary of Veterans' Affairs TTL | YES |
   | Secretary TTL of Veterans' Affairs AGC | NO |
   | Secretary TTL of Veterans' Affairs AGC TTL | NO |

2. Person tags do not include titles

   Mr. Biggs MBC
   Administrator Sanders IND

3. Person tags include shortened named references

   the Speaker TTL

4. ONLY use Document tags outside of quotations

   Amend Section 1710F REF as, "1710F: . . ."

| Top Tag | Subclass | Description | Examples |
|---|---|---|---|
| **Person** | Member **MBC** | A member of Congress, typically the sponsor or co-sponsor of the legislation. | Mr. Biggs MBC |
| | Title **TTL** | Mentions of individuals by their official position or office. | Speaker TTL ; President TTL |
| | Individual **IND** | Any specific named person who is neither a Member of Congress nor a Title. | John Doe IND |

Table 7: Tag Definitions: PERSON Subclasses

| Top Tag | Subclass | Description | Examples |
|---|---|---|---|
| **Organization** | Nation **NAT** | National-level geopolitical entities (includes American Indian nations). | United States NAT |
| | State **STA** | State-level geopolitical entities within the U.S. | Illinois STA |
| | Locality **LOC** | Local geopolitical entities, such as cities, counties, or municipalities. | Cook County LOC |
| | Committee **COM** | Named Congressional committees and subcommittees. | Ways and Means COM |
| | Agency **AGC** | Federal departments, bureaus, or executive administrations. | Dept. of Justice AGC |
| | Leg. Body **LEG** | Primary legislative chambers (Senate, House, Congress). | Senate LEG |
| | International **INT** | International organizations and multinational institutions. | United Nations INT |
| | Association **ASC** | Catch-all for other named organizations, nonprofits, or private entities. | Planned Parenthood ASC |

Table 8: Tag Definitions: ORGANIZATION Subclasses

5. ONLY use Document tags if the Act name or full reference is mentioned in the text, not when only a section or paragraph is mentioned

   paragraph one of the twenty-seventh article of amendment to the Constitution of the United States REF — YES

   section (1)(b) REF — NO

   section (1)(b) of US Code3 ... REF — YES

6. ONLY Document tags span over "and" and "or"

   sections (a) and (b) of US Code .... REF — YES

   John and Michael IND — NO

   John IND and Michael IND — YES

7. Administrative actions in bills, including "Clerical Amendment", and "Date of Effect" should not be tagged

8. If any information below the Act title is included, tag the mention as Document (Title, section, paragraph, part, etc.)

   Title IX of the Farm Security and Rural Investment Act of 2002 REF

   Farm Security and Rural Investment Act of 2002 PA

9. Tag parenthesized mentions of Documents separately

   Title IX of the Farm Security and Rural Investment Act of 2002 REF (7 U.S.C. 8101 et seq.) PAR

10. Do not tag dates, currency, or any other numbers

| Top Tag | Subclass | Description | Examples |
|---|---|---|---|
| **Document** | Code **CODE** | Foundational legal codes and structured statutory bodies. | U.S. Code **CODE** |
| | Bill **BILL** | Specific bill identifiers and tracking numbers. | H.R. 189 **BILL** |
| | Reference **REF** | Structural subdivisions of a legal text (sections, titles, paragraphs). | section 102 **REF** |
| | Parenthetical **PAR** | Legal citations provided in parenthetical format. | (42 U.S.C. 4332) **PAR** |
| | Treaty **TRE** | Named international treaties and conventions. | Geneva Convention **TRE** |
| | Report **REP** | Official reports, budget documents, or findings. | President's Budget **REP** |

Table 9: Tag Definitions: DOCUMENT Subclasses

| Top Tag | Subclass | Description | Examples |
|---|---|---|---|
| **Act** | Public Act **PA** | Named acts or legislation referred to by their popular title. | Green New Deal **PA** |
| | Amendment **AMD** | Specific amendments to the U.S. Constitution. | Second Amendment **AMD** |

Table 10: Tag Definitions: ACT Subclasses

| Top Tag | Subclass | Description | Examples |
|---|---|---|---|
| **Abstraction** | Program **PRO** | Named government-managed programs and initiatives. | Medicare **PRO** |
| | System **SYS** | Intangible government frameworks, databases, or digital networks. | Forest System **SYS** |
| | Infrastructure **INF** | Tangible or physical frameworks established by government action. | Interceptor **INF** |
| | Fund **FUND** | Specific, named accounts or monetary trust funds. | General Fund **FUND** |
| | Doctrine **DTR** | Abstract legal policies, principles, or constitutional rights. | Free Speech **DTR** |
| | Case **CASE** | Court cases and judicial precedents. | McCulloch v. MD **CASE** |
| | Specification **SPC** | Specific legal classifications, modifiers, or schedules. | 340B **SPC** |
| | Session **SES** | Named intervals, meetings, or sessions of a legislative body. | 118th Congress **SES** |

Table 11: Tag Definitions: ABSTRACTION Subclasses

| Top Tag | Subclass | Description | Examples |
|---|---|---|---|
| **Class** | Protected **PC** | Groups defined by law for special protections (Race, Disability, etc.). | veterans **PC** |
| | Non-Prot. **NPC** | Other groups that are the subject of legislation but not protected. | refugees **NPC** |

Table 12: Tag Definitions: CLASS Subclasses

## C. Summary Statistics

| Type | | Train | Dev | Test | Total |
|---|---|---|---|---|---|
| Abstraction | | 217 | 91 | 79 | 387 |
| | Case | 0 | 0 | 0 | 0 |
| | Doctrine | 8 | 0 | 1 | 9 |
| | Fund | 39 | 11 | 1 | 51 |
| | Infrastructure | 3 | 0 | 21 | 24 |
| | Misc | 4 | 0 | 0 | 4 |
| | Program | 30 | 39 | 18 | 87 |
| | Session | 87 | 21 | 31 | 139 |
| | Specification | 14 | 17 | 0 | 31 |
| | System | 32 | 3 | 7 | 41 |
| Act | | 197 | 40 | 80 | 317 |
| | Amendment | 3 | 1 | 1 | 5 |
| | Public Act | 194 | 39 | 79 | 312 |
| Class | | 208 | 57 | 67 | 332 |
| | Non-Protected Class | 194 | 45 | 59 | 298 |
| | Protected Class | 14 | 12 | 8 | 34 |
| Document | | 774 | 214 | 350 | 1,338 |
| | Bill | 162 | 42 | 60 | 264 |
| | Code | 30 | 2 | 5 | 37 |
| | Parenthetical | 204 | 32 | 112 | 348 |
| | Reference | 361 | 131 | 171 | 663 |
| | Report | 17 | 7 | 2 | 26 |
| | Treaty | 0 | 0 | 0 | 0 |
| Organization | | 1261 | 379 | 579 | 2,219 |
| | Agency | 249 | 78 | 191 | 518 |
| | Association | 28 | 38 | 38 | 104 |
| | Committee | 158 | 44 | 78 | 280 |
| | International Institution | 1 | 0 | 3 | 4 |
| | Legislative Body | 566 | 150 | 194 | 910 |
| | Locality | 20 | 7 | 28 | 55 |
| | Nation | 178 | 35 | 120 | 333 |
| | State | 61 | 27 | 17 | 105 |
| Person | | 731 | 248 | 238 | 1,217 |
| | Member | 339 | 106 | 121 | 566 |
| | Name | 4 | 0 | 0 | 4 |
| | Title | 388 | 142 | 117 | 647 |

Table 13: Counts of total level-one and level-two mentions by split