

Building Structured Sentencing Data from Federal Criminal Docket Entries: An Ontology and Entity Extraction Study

David Guecha Aanan Biswas Charlotte S. Alexander

Georgia Institute of Technology

{dahumada3, abiswas74, calexander97}@gatech.edu

Abstract

The U.S. Sentencing Commission provides substantial data about federal criminal sentences but omits judge- and defendant-specific identifiers, as well as many non-incarceration components of sentences. To support research on these gaps, we construct a domain-specific ontology and an annotated dataset from federal sentencing docket entries. These short, contemporaneous records describe counts, incarceration, supervision, monetary penalties, and other sentencing conditions; however, their formatting varies across clerks, judges, and courts. We compare six approaches under a shared labeled-span evaluation: ontology-guided Gemini 2.5 Flash, CRFsuite, ModernBERT, Legal-BERT, zero-shot GLiNER, and fine-tuned GLiNER. On 100 manually annotated docket entries, fine-tuned GLiNER achieves the highest exact-span micro-F1 at $.703 \pm .009$, followed by CRFsuite at $.655$. Gemini achieves $.463 \pm .038$ exact F1 but the highest relaxed F1 of $.787 \pm .013$ under a diagnostic that credits a prediction only when its label matches and either normalized span text contains the other. The contrast shows that recovering compatible entity text and reproducing ontology-defined boundaries are materially different abilities. The results emphasize the continuing value of domain-specific task definitions, annotation ontologies, and supervised adaptation for legal information extraction.

1 Introduction

Sentencing is an extraordinarily important moment in criminal cases, when a judge decides the type, length, and manner of punishment that the defendant will face. Since its founding in 1984, the U.S. Sentencing Commission (USSC) has made substantial data available on federal criminal sentences, enabling study of patterns and disparities across charge types, defendant characteristics, and courts. The USSC data, while valuable, also has known blind spots. In particular, the USSC’s datasets omit

identifiers for judges and defendants and do not record many non-incarceration components of sentences.

In this study, we present an alternative source of federal criminal sentencing data by mining the text of docket entries, which are short, contemporaneous records of judges’ sentences. Docket entry text often describes counts, incarceration terms, supervision requirements, monetary penalties, and other sentencing conditions. Although defendant and judge identities are available from the case caption and other docket metadata, the sentencing information recorded in individual docket entries is not consistently expressed across clerks, judges, and courts.

Converting these records into structured entity spans enables the generation of new, standardized sentencing data. Yet general purpose named entity recognition (NER) is not well matched to this setting: general English NER systems degrade substantially when transferred to legal filings (Au et al., 2022). The target entities extend beyond people, organizations, and locations to domain-specific legal concepts, quantities, and modifiers. A system may need to distinguish a count reference (e.g., "Count 1," referring to a particular criminal charge) from its term of imprisonment, identify whether a monetary assessment was waived, or recover whether sentences run concurrently or consecutively. These distinctions are boundary sensitive: identifying the right concept at the wrong location in the docket is an error under exact evaluation. Prior work likewise shows that legal entity extraction depends on domain-specific label definitions and annotation rules (Chalkidis et al., 2020; Kalamkar et al., 2022).

This study develops a domain-specific ontology and manually annotated dataset for extracting structured sentencing information from federal docket entries. We then use a shared labeled-span evaluation to compare six approaches: CRFsuite, ModernBERT, Legal-BERT, zero-shot and fine-tuned

GLiNER, and ontology-guided Gemini. The comparison tests how effectively the dataset and ontology support extraction and how sensitive each approach is to span-boundary requirements.

This work makes three contributions:

1. We introduce a manually annotated federal sentencing-docket dataset and a 44-label ontology designed to support structured sentencing analysis.
2. A comparative evaluation examines ontology-guided Gemini extraction, a CRFsuite baseline, ModernBERT, Legal-BERT, and both zero-shot and fine-tuned GLiNER under a shared evaluation interface.
3. An error analysis evaluates exact span reconstruction, uses relaxed containment as a boundary diagnostic, and categorizes label, boundary, and unmatched extraction errors.

2 Related Work

Sentencing data beyond USSC releases. Recent projects illustrate both the value and difficulty of collecting sentencing data outside USSC releases. The Ohio State University Interdiction Lab and Lissner describe constructing a defendant-level dataset from PACER and Bureau of Prisons sources (The Ohio State University Interdiction Lab and Lissner, 2025). Wroblewski argues that independent sentencing-data collection can expose gaps in commission research (Wroblewski, 2025). SentencingStats similarly offers judge- and case-level comparative sentencing analysis based on public records (SentencingStats, n.d.). These efforts motivate reliable extraction from docket text, where relevant facts are available but not consistently structured. Related docket-focused work applies information-extraction and machine-learning techniques to motions and orders in U.S. federal dockets (Vacek et al., 2019), while open-justice research documents the financial and technical barriers to making federal court records analytically usable (Pah et al., 2022).

Domain-specific legal extraction. Legal information extraction often requires labels that are more fine-grained than general-purpose NER. E-NER demonstrates degradation when general-English NER models are transferred to legal filings (Au et al., 2022). Corpora for Indian judgments (Kalamkar et al., 2022) and Romanian legal

documents (Pais et al., 2024) likewise introduce jurisdiction- and domain-specific entity inventories. LEEC extracts legal elements from Chinese criminal judgments using 155 legal and extra-legal labels for sentencing and judicial-fairness analysis (Xue et al., 2024). These resources motivate evaluation against an explicit ontology rather than only broad entity categories, as in off-the-shelf NER. Earlier contract-element extraction introduced a fine-grained annotated benchmark and compared rules, linear classifiers, and a hybrid approach (Chalkidis et al., 2017). Labor Lex similarly pairs a domain corpus with supervised and in-context extraction experiments for Brazilian labor-law text (Quinta de Castro and Da Silva, 2025). Recent NLLP work also introduces NER and multi-label datasets for planning court cases and compares general-purpose and legal-domain models (Mali et al., 2024).

Legal ontologies and annotation schemas. LKIF Core formalizes reusable legal concepts in OWL to support legal knowledge interchange (Hoekstra et al., 2007). LegalNERo links a legal NER corpus to semantic resources (Pais et al., 2024). Our contribution has a narrower scope: a sentencing-specific ontology used as an annotation and extraction schema with operational span boundaries. Its label taxonomy and JSON annotation rules do not implement an OWL ontology or mappings to these resources. Similar terms alone do not establish equivalent semantics or interoperability. Explicit concept mappings and validation of their jurisdictional and boundary assumptions remain future work.

Domain adaptation and flexible NER. LEGAL-BERT compares general encoders, continued legal-domain pretraining, and pretraining from scratch, finding that adaptation guidance does not transfer uniformly to legal tasks (Chalkidis et al., 2020). Multi-task domain adaptation for legal NER produces modest aggregate gains with larger recall changes (Smădu et al., 2022). ModernBERT provides a newer general-purpose encoder designed for efficient long-context finetuning (Warner et al., 2025). GLiNER predicts entity types supplied at inference and reports strong zero-shot NER performance (Zaratiana et al., 2024).

Prompt-guided extraction. Large language models can perform legal extraction from instructions and examples, but can still produce inconsistent and over-generalized output (Kwak et al.,

Group	Role
Anchor	References such as counts or external court sentences.
Metadata	Defendant, interpreter, redaction, or amendment information.
Event	Outcomes, obligations, recommendations, and conditions.
Connector	Concurrent or consecutive sentencing indicators.
Attribute	Durations, amounts, deadlines, dismissal, or waiver.

Table 1: Sentencing ontology groups.

2023). GuideNER finds that explicit annotation guidelines can be more effective than demonstrations for in-context NER (Huang et al., 2025). Our ontology-guided Gemini condition supplies label definitions, span rules, illustrative annotated snippets, and a two-step entity-and-relation instruction; its returned entity markup is checked against the source (Section 4.2).

Our study focuses on federal criminal sentencing docket entries from U.S. courts. All six approaches pass through the same span-processing and scoring pipeline. We separate exact reconstruction from a relaxed, rule-based character-matching diagnostic, showing whether an approach reproduces the annotated entity at its document location or merely returns compatible text.

3 Task and Dataset

3.1 Task Definition

The input is one criminal sentencing docket entry. The output identifies each ontology-defined entity by its label and by its start and end characters in the entry. Figure 1 illustrates this span-extraction task and the two scoring criteria; Figure 3 shows a redacted example of the docket-entry format.

The labels identify character spans, not document-level topics. The labeling guidelines organize entities into five groups (Table 1) and separately specify the meaning and boundary conventions for every label.

Ontology. Our sentencing ontology provides a domain-specific annotation and extraction schema: a label taxonomy, definitions, and span-boundary rules. We use “ontology” throughout to refer to this operational task definition. Its machine-readable JSON rules provide general notes and, for each label, its name, category, definition, and span rules (Appendix F); the prompt separately specifies

generic directed relations between entities. Those relations are not scored here, and we provide no formal interoperability mappings or logical axioms. The ontology represents components of a federal criminal sentence observable in a docket entry. Anchors connect a sentence to counts or other proceedings; metadata identifies relevant case context; events describe outcomes, requirements, recommendations, and conditions; connectors indicate concurrent or consecutive conditions; and attributes express durations, amounts, deadlines, dismissal, or waiver. A phrase may contain adjacent but distinct elements, for example, an incarceration event and its duration attribute (2 years imprisonment). Therefore, the annotation guidelines define both what each label means and which characters in the docket text belong to it. The complete 44-label inventory and development and evaluation support appear in Appendix D, Table 9; the definitions and span rules appear in Appendix F. For example, Immigration consequences is an event label whose rule tags only the immigration-consequence phrase; Amendment or correction is metadata whose rule tags the minimal phrase indicating that the judgment or entry was amended or corrected. The latter marks a change to the record, not a type of punishment.

3.2 Data and Annotation

The dataset consists of manually annotated criminal sentencing docket entries obtained through SCALES (Pah et al., 2022; SCALES-OKN, n.d.). The source records originate in PACER (Public Access to Court Electronic Records), the electronic public-access system for U.S. federal courts (United States Courts, n.d.); SCALES provides a research data platform built from court records, rather than being the court filing system itself. The final supervised-development collection contains 163 entries and 1,386 reference spans; the separate evaluation collection contains 100 entries and 1,049 reference spans. The two collections are disjoint under normalized-text and docket-identity matching. The retained court identifiers cover 68 federal district courts (58 in development and 44 in evaluation). We preserve the docket text and represent entities using character offsets. Each entry is identified by its case number and row number on the full docket sheet.

Table 2 describes the analysis-ready evaluation records. Only the docket text is supplied to an evaluated system; identifiers and normalized text

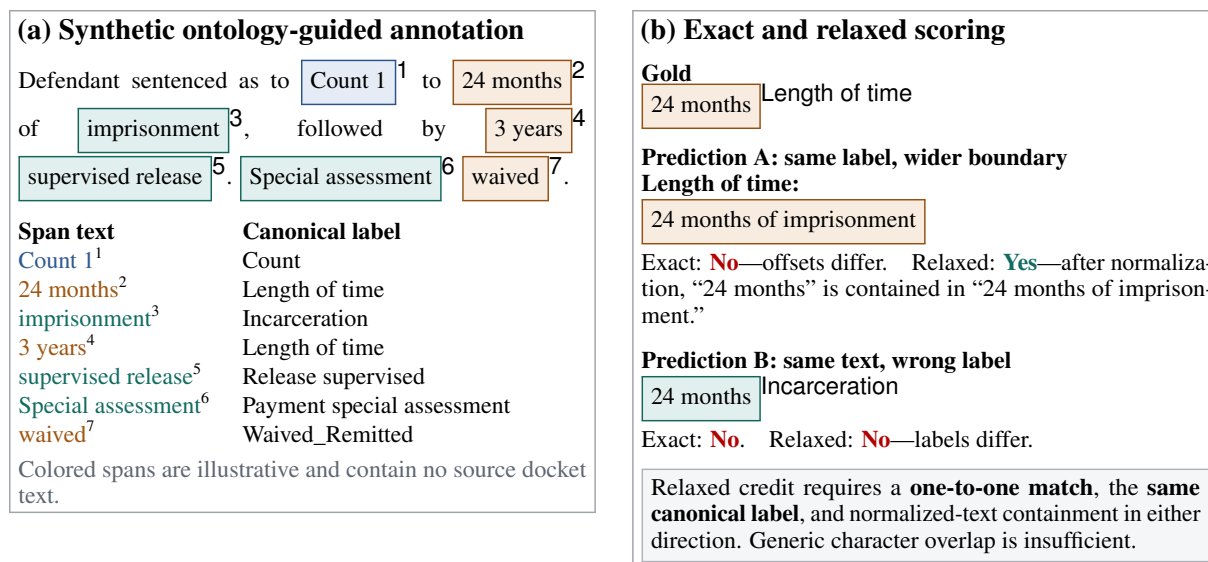


Figure 1: Task and evaluation illustration. Panel (a) shows a synthetic docket excerpt and its canonical entity spans; superscript numbers link each colored span to its label in the table. Panel (b) distinguishes exact document-offset scoring from the normalized-text containment sensitivity analysis.

support record alignment, and reference fields are accessed only during scoring.

Four manual labelers prepared the development and evaluation datasets. To assess agreement, each independently annotated the same sample of 10 docket entries, selected to vary in length and complexity. The agreement calculation used 417 span annotations across the four labelers. Mean F1 over the six annotator pairs was 88.1% when a one-to-one match required the same canonical label on overlapping text. Under the primary scorer’s stricter requirement of identical start, end, and label, mean pairwise F1 was 80.3%. This gap illustrates the difficulty of applying ontology-specific boundary rules even when annotators identify overlapping legal language—for example, whether a duration includes an adjacent modifier. A fifth reviewer subsequently examined every entry in the final datasets and adjudicated the identified annotation disagreements. Human labelers followed the same annotation guidelines later supplied to the prompted model. This review was prediction-blind: no output from Gemini or any other evaluated approach was used by human annotators to create or adjudicate the reference labels. Per-label inter-annotator agreement is unavailable, so we cannot determine whether labels with low extraction performance also produced greater disagreement among annotators.

Availability. The dataset is publicly available as training and evaluation CSV files, including docket

text, annotations, and record identifiers.¹ The released `ucid` and `row_ordinal` fields associate annotations with a case and its entry on the docket sheet. Researchers can also access source court data through the SCALES platform (SCALES-OKN, n.d.).

3.3 Preprocessing and Splits

All modeling approaches begin with unchanged docket text and a shared document-level span representation. Each approach may tokenize or window the text internally, but its predictions are mapped back to the original character coordinates. Invalid, unparseable, or ungrounded outputs are recorded rather than silently rewritten or omitted. Appendix A provides the model-specific preprocessing and adapter details.

The 100-entry evaluation set is excluded from fitting and selection. Supervised approaches use a fixed, grouped (by docket similarity) five-fold split of the 163 development entries and are refit on all development entries only after configuration selection. Model-specific transformations are documented in Appendix A.

¹Public dataset download (Dropbox).

Column	Description	Role
ucid	Case-level identifier used to associate an entry with its docket.	Metadata
row_ordinal	Position of the entry on the full docket sheet.	Metadata
docket_text	Original docket-entry text supplied to every evaluated system.	Model input
docket_text_norm	Deterministically normalized text used for record matching and consistency checks.	Alignment
ground_truth_annotated_text	Curated entity representation from which canonical labels and document-level character spans are recovered.	Entity reference
ground_truth_relations	Curated relation representation retained in the data but not evaluated in this study.	Not scored

Table 2: Columns in each analysis-ready evaluation record. Only docket_text from the record is provided as model input.

4 Methodology

4.1 Shared Span Formulation

Let $x = (c_1, \dots, c_N)$ be the character sequence of a docket. A system predicts

$$\hat{Y} = \{(s_j, e_j, \ell_j)\}_{j=1}^m,$$

where $0 \leq s_j < e_j \leq N$ are half-open document offsets and ℓ_j is a canonical label. Character offsets form the shared interface because systems use different tokenizers and output formats. BIO predictions are reconstructed into document spans, Gemini annotations are parsed and grounded to source text, and GLiNER window-local offsets are restored to document coordinates.

Figure 2 shows the complete comparison path and distinguishes shared ingestion from each model’s preprocessing and output adapter.

4.2 Compared Approaches

The ontology defines the common annotation and evaluation targets, but the systems receive different representations of it. CRFsuite and the two encoders learn from labeled development spans expressed as BIO tags; they do not receive the natural language annotation rules as input. Both GLiNER conditions receive label-name strings, and fine-tuned GLiNER additionally learns from labeled spans. Gemini alone receives the full rules JSON and the prompt instructions.

CRF. CRFsuite (Okazaki, 2007) provides a feature-based structured prediction baseline for recurring docket language. It converts human-annotated reference (“gold”) spans to BIO tags over deterministic offset-preserving tokens. A linear-chain CRF uses lexical, orthographic, sentencing-domain, and neighboring-token features. Each development entry is predicted out of fold, and the

final model is refit on all 163 eligible entries. Appendix A gives the tokenizer, features, and optimization settings.

Gemini. Gemini 2.5 Flash (Gemini Team, Google, 2025) provides a generative, instruction-guided comparison without task-specific weight updates. It uses the provider identifier gemini-2.5-flash and receives the docket, full rules JSON, and span instructions. The fixed prompt contains short annotated guideline examples, including time-served and drug-testing snippets, but no full-docket demonstrations or retrieved examples. Responses are parsed into entity spans and checked against the source text. Empty outputs or recorded request errors prevent final scoring until resolved. Responses that fail source reconstruction remain in evaluation under the error-handling procedure described in Appendix A. The exact prompt template and the identity of its rules input are documented in Appendix E, and configuration, retry, and parsing details appear in Appendix A.

Zero-shot GLiNER. GLiNER (Zaratiana et al., 2024) jointly encodes label names and text with a bidirectional transformer. It constructs candidate span representations from their endpoints and scores compatibility with entity-type representations using a dot product and sigmoid. This is direct span prediction, unlike BIO token classification or Gemini’s generated markup. The frozen condition tests transfer from a general label-conditioned extractor. We use the gliner_medium-v2.1 checkpoint without fine-tuning: it is zero-shot with respect to model parameters, but not blind to the task label names. At inference it receives the 44 canonical ontology label names but no annotated docket examples, and its weights remain unchanged. Its

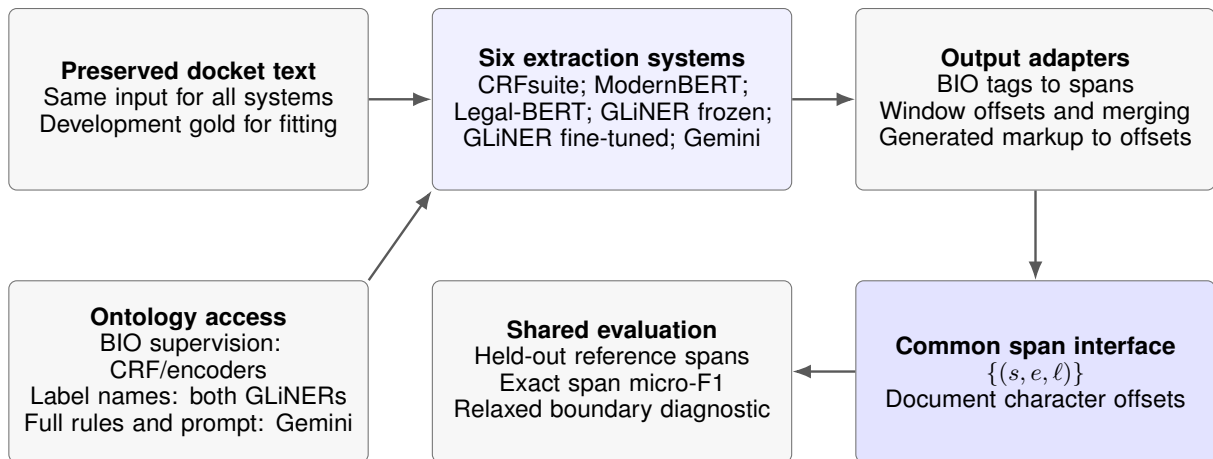


Figure 2: Unified comparison architecture. Supervised fitting uses development spans only; the held-out reference spans enter evaluation. Natural-language rules are supplied to Gemini, label names to GLiNER, and labeled BIO targets to CRFsuite and the encoders. Outputs share the same character-span scoring interface.

threshold, windowing, and span-merging rules appear in Appendix A.

Fine-tuned GLiNER. The sixth condition fine-tunes the same GLiNER checkpoint (Zaratiana et al., 2024) on the 163 development records, testing the value of local supervised adaptation within the same model family. Final models are independently refit and reported across three seeds. The selected training settings appear in Appendix A.

Encoder baselines. We fine-tune ModernBERT-base (Warner et al., 2025) and Legal-BERT-base-uncased (Chalkidis et al., 2020) for BIO classification with token cross-entropy. They provide general-purpose and legal-domain pretrained encoder baselines, respectively; their other architectural differences prevent a controlled attribution to pretraining domain alone. Their exact checkpoint identifiers appear in Appendix A. Their token predictions are restored to document character offsets before scoring. Window sizes, ownership, masking, and span-merging rules appear in Appendix A.

4.3 Training and Inference

Supervised approaches use the fixed grouped five-fold split for model and configuration selection, then are refit on all 163 development entries. The evaluation collection is not used for selection. Gemini and zero-shot GLiNER require no task-specific fitting; their prompt or checkpoint, label source, threshold, and preprocessing are fixed before evaluation.

The selected settings, candidate configurations,

and adapter details are reported in Appendix A.

4.4 Evaluation

The primary metric is exact document character span micro-F1. A true positive requires identical start, end, and canonical label, so predicting the same text at a different occurrence in a document is incorrect. We also report precision, recall, per-label scores, reference support (the number of annotated evaluation spans), and secondary macro-F1 over labels with reference support. Of the 44 ontology labels, 41 occur in the evaluation collection; Redacted, Suspended sentence, and No contact or restraining order have zero evaluation support; the latter two are also unattested in the development collection. We retain these labels for compatibility with the original label inventory and to preserve the ontology’s intended coverage as the corpus expands. We therefore compute macro-F1 over the 41 labels with evaluation support.

The secondary relaxed scorer is a boundary-sensitivity diagnostic, not a measure of semantic correctness. It makes a one-to-one match when labels agree and one normalized span string contains the other; for example, supervised release can match 1 year of supervised release. Because this can credit boundaries that violate the ontology, relaxed scoring does not replace exact scoring. The normalization details appear in Appendix A.

Table 3 gives two examples from the official evaluation. In both records, the supervised model had a document-level exact F1 higher than the Gemini pass shown. The systems found overlapping text with the same label, but only CRFsuite or fine-

tuned GLiNER reproduced the annotated boundary.

Each trained seed variant and each Gemini provider pass is evaluated separately on all 100 records. For ModernBERT, Legal-BERT, and fine-tuned GLiNER, we report the arithmetic mean and sample standard deviation of the resulting metrics across three training seeds. For Gemini, we report the corresponding summary across five API passes conducted with the same prompt and configuration. CRFsuite and zero-shot GLiNER are each evaluated once and are therefore reported as single values without a standard deviation. We summarize complete evaluation results rather than pooling predictions across repetitions or selecting the best-performing repetition.

5 Results

5.1 Main Results

Table 4 summarizes exact and relaxed extraction performance for all six systems.

Across the six systems, the evaluation contains 16 complete results: one each for CRFsuite and zero-shot GLiNER, three seeded results each for fine-tuned GLiNER, ModernBERT, and Legal-BERT, and five Gemini provider passes. Each is scored on the same 100 docket entries against all 1,049 reference spans; no document, failed prediction, or label with gold support is excluded.

Fine-tuned GLiNER has the highest primary exact F1 at $.703 \pm .009$, with balanced precision and recall. CRFsuite ranks second at .655 and has the highest exact precision, .742. Legal-BERT obtains $.569 \pm .010$ exact F1 and ModernBERT $.485 \pm .005$. Their difference is descriptive: the models also differ in architecture, tokenizer, context length, and checkpoint, so this comparison does not isolate the effect of legal-domain pretraining.

Fine-tuning changes GLiNER substantially within the evaluated family. The zero-shot condition obtains .345 exact F1, compared with $.703 \pm .009$ after task-specific fine-tuning. The three fine-tuned runs (.696, .712, and .701) show that the mean is not driven by one favorable refit.

The exact and relaxed metrics yield different practical interpretations. Gemini has an exact F1 of $.463 \pm .038$ but the highest relaxed F1 at $.787 \pm .013$, an average gap of .324. ModernBERT has the next-largest gap, increasing from $.485 \pm .005$ to $.645 \pm .014$. Gemini had 24–37 source-reconstruction failures per pass (Appendix A); its gap therefore reflects reconstruction

failures as well as boundary sensitivity. Receiving relaxed credit because the label matches and one normalized span contains the other does not show that the prediction has the meaning or boundary required by the ontology.

5.2 Per-Label Results

Across the 41 evaluation-supported labels, fine-tuned GLiNER also has the highest exact macro-F1 ($.453 \pm .008$). Gemini’s exact macro-F1 is $.375 \pm .030$, followed by CRFsuite (.309), zero-shot GLiNER (.262), ModernBERT ($.210 \pm .010$), and Legal-BERT ($.195 \pm .003$). Macro-F1 and micro-F1 therefore emphasize different behavior: CRFsuite is strong on frequent patterns, while Gemini and the GLiNER variants cover a broader portion of the ontology.

All approaches perform better on labels represented by more than 20 spans in the 163-entry development collection than on less frequent labels. Fine-tuned GLiNER reaches .794 exact micro-F1 in the highest-support bin and remains nonzero for support-one (.417) and support-2–5 (.398) labels. CRFsuite reaches .754 in the highest-support bin but .077 on support-one labels. The two encoders have no exact true positives in the lowest support bins. These descriptive results identify minority-label representation as an important target for future data collection and analysis.

Figures 4–5 show exact-span *recall*, not F1. Immigration consequences has only 3 development spans but 24 evaluation spans: exact recall is .125 for CRFsuite, .042 for Gemini and zero-shot GLiNER, and zero for both encoders and fine-tuned GLiNER. Amendment or correction has 1 development and 4 evaluation spans and zero exact recall for every system. Conversely, fine-tuned GLiNER’s perfect recall and F1 for Community service concern just one evaluation span (2 development spans), not strong evidence of generalization. Scarcity is one plausible factor, but these descriptive comparisons do not isolate its effects from boundary rules, wording, or model design.

5.3 Run-to-Run Variation

Table 5 reports mean exact F1 and sample standard deviation across repeated runs.

Variation is modest for the three trained neural systems and larger for Gemini’s five provider passes. These sample standard deviations describe the observed checkpoints or passes; they are not confidence intervals over the evaluation population.

Label	Reference and exact output		Gemini output		Scoring consequence
Release supervised	supervised	release (CRF-suite)	1 year of supervised release		Relaxed match; exact mismatch
Length of time	TWENTY-FOUR (fine-tuned GLiNER)	(24) MONTHS	*TWENTY-FOUR (24) MONTHS		Relaxed match; exact mismatch

Table 3: Redacted boundary examples from official predictions. Exact scoring requires the reference start and end; relaxed scoring credits these same-label containment cases.

Method	Exact P	Exact R	Exact F1	Relaxed P	Relaxed R	Relaxed F1
Gemini 2.5 Flash	.419 $\pm$.039	.518 $\pm$.035	.463 $\pm$.038	.711 $\pm$.023	.881 $\pm$.008	.787 $\pm$.013
CRFsuite	.742	.586	.655	.801	.633	.707
ModernBERT	.457 $\pm$.018	.517 $\pm$.012	.485 $\pm$.005	.608 $\pm$.029	.688 $\pm$.018	.645 $\pm$.014
Legal-BERT	.569 $\pm$.014	.568 $\pm$.015	.569 $\pm$.010	.646 $\pm$.012	.645 $\pm$.013	.645 $\pm$.005
GLiNER zero-shot	.320	.373	.345	.402	.467	.432
GLiNER fine-tuned	.698 $\pm$.014	.708 $\pm$.005	.703 $\pm$.009	.756 $\pm$.016	.768 $\pm$.006	.762 $\pm$.010

Table 4: Results on all 100 manually annotated entries. Repeated systems report arithmetic mean $\pm$ sample SD across three seeds or five Gemini passes. Exact span F1 is primary; relaxed matching is a secondary boundary sensitivity.

System	Runs	Exact F1	Sample SD
Gemini 2.5 Flash	5	.463	.038
ModernBERT	3	.485	.005
Legal-BERT	3	.569	.010
GLiNER fine-tuned	3	.703	.009

Table 5: Variation in final exact F1 across genuinely repeated runs.

6 Error Analysis

We deterministically pair exact false negatives and false positives. A wrong-label pair covers exactly the same characters but uses different labels. A boundary mismatch uses the same label and overlapping text but different start or end characters. An unmatched false positive is a prediction with neither type of pair; an unmatched false negative is an unpaired reference span. Each paired error accounts for one false positive and one false negative (Table 6). Thus total FP equals wrong-label pairs plus boundary pairs plus unmatched FP; total FN uses the same paired counts plus unmatched FN. Fractional entries arise from averaging integer run-level counts, not fractional spans.

ModernBERT has the largest number of paired same-label boundary mismatches, consistent with its exact-to-relaxed increase. Gemini also has frequent boundary mismatches, but its larger problem is unmatched output: it averages 656.4 unmatched false positives and 405.4 unmatched false negatives. Its high relaxed recall, therefore, reflects broad recovery of candidate text, not uniformly controlled extraction.

CRFsuite has the fewest unmatched false positives but substantially more unmatched false negatives, matching its high precision, lower recall profile. Fine-tuned GLiNER has the fewest unmatched false negatives and a comparatively low unmatched false-positive burden. Zero-shot GLiNER has the largest same boundary wrong label count; fine-tuning reduces this from 86.0 to 24.7 while also reducing unmatched errors. This within-family change is consistent with local examples improving both category distinctions and docket-specific boundary patterns.

7 Discussion

General-purpose linguistic capability does not guarantee ontology-exact extraction on this task. Fine-tuned GLiNER obtains the highest observed exact-span F1, while the much simpler CRFsuite model remains second and achieves the highest precision. Both results illustrate the value of aligning a model with recurring docket patterns and explicit boundary rules.

Gemini illustrates a complementary strength and weakness. One possible contributor is prediction mechanism: the discriminative approaches decode spans directly, whereas Gemini generates marked-up text that is grounded to document offsets after generation. The experiments do not isolate this design difference, but it is consistent with Gemini’s substantially higher relaxed than exact score. It often finds the relevant words, producing the highest relaxed F1 and recall, but its exact F1 is much lower.

System	Wrong label	Boundary mismatch	Unmatched FP	Unmatched FN
Gemini 2.5 Flash	13.6	86.6	656.4	405.4
CRFsuite	15.0	42.0	157.0	377.0
ModernBERT	15.7	171.3	460.3	319.7
Legal-BERT	44.7	77.0	329.3	331.3
GLiNER zero-shot	86.0	92.0	651.0	480.0
GLiNER fine-tuned	24.7	58.0	239.7	223.3

Table 6: Exact-span errors per 100-entry evaluation run. Wrong label and Boundary mismatch count paired prediction/reference errors; each pair contributes one FP and one FN. Unmatched FP counts remaining predicted spans; Unmatched FN counts remaining reference spans. Means over five Gemini passes or three neural refits are fractional; CRFsuite and zero-shot GLiNER are single runs. Counts are not percentages.

This distinction matters when predictions populate an auditable database: including or omitting a modifier, duration, count, or surrounding phrase can change the represented entity even when the prediction appears textually close. Relaxed matching is useful for diagnosing boundary sensitivity, but exact character-span evaluation remains the primary measure.

The two GLiNER conditions provide the clearest within-family methodological contrast. Supplying label names to the evaluated frozen extractor yields substantially lower exact F1 than adapting it to local examples and the task ontology. This supports the practical claim that domain-driven task definitions, carefully developed ontologies, and task-specific adaptation remain important even when powerful general models are available. Future work should evaluate relation extraction, explore development-only aggregation strategies, and test whether other comparatively simple structured models can exploit recurring docket language. The current ontology and dataset remain fixed for this study; broader sentencing and fairness analyses will require separately versioned expansion and validation.

Scalability. The evaluated inputs are short docket entries. Longer, heterogeneous judgments or briefs can spread relevant facts and their modifiers across windows. Our current window adapters restore offsets and resolve overlaps; they do not establish cross-window reasoning. A larger label inventory and longer guidelines would also increase Gemini’s input size and potential inference cost, and may increase label confusion. Hierarchical label selection or staged extraction could address these constraints, but were not tested. Transfer across courts and document types would require new annotation, explicit ontology versions, and independent validation of coverage, boundary rules, and accuracy, rather than

reuse of the present scores as an assurance of performance.

8 Conclusion

We present a domain-specific ontology and manually annotated dataset for extracting structured sentencing information from federal criminal docket entries, together with a comparative evaluation of six approaches. On 100 evaluation entries, fine-tuned GLiNER obtains the highest observed exact-span F1 ($.703 \pm .009$), compared with Gemini 2.5 Flash ($.463 \pm .038$) and zero-shot GLiNER (.345). Gemini recovers a large share of compatible text under relaxed matching, but its exact score shows that finding relevant language is not the same as producing ontology-correct spans. CRFsuite’s competitive .655 exact F1 and highest exact precision (.742) show that a comparatively simple structured model can remain competitive on this formulaic docket task when boundary rules are explicit. Together, the findings reinforce the importance of domain-driven task definitions, ontology development, and task-specific adaptation for this legal information-extraction objective. The resulting structured spans are intended to support future sentencing research; fairness or disparity outcomes are not evaluated here.

9 Limitations

The study evaluates 100 held-out docket entries containing 1,049 spans and uses a separate 163-entry pool containing 1,386 spans for supervised development (selection and final refitting). These samples are too small to establish cross-court, temporal, or cross-jurisdictional generalization. Many ontology labels are rare. Labels absent from development have no positive supervised training examples; labels absent from evaluation cannot support an empirical recall estimate. The ontology retains

all 44 labels, while macro-F1 averages the 41 with evaluation support. Table 9 reports both collections’ complete span counts. The label inventory and development/evaluation splits are fixed across systems.

The systems differ in more than just model family. ModernBERT and Legal-BERT use different architectures, tokenizers, context lengths, and checkpoints, so their comparison does not isolate the effect of legal-domain pretraining. The relaxed scorer measures same-label normalized substring containment rather than semantic correctness. Exact scoring also treats identical text at a different occurrence as incorrect; this is intentional for location-specific structured data but can magnify ambiguity in repetitive docket language. Annotation agreement drops from 88.1% under same-label overlap to 80.3% under exact boundaries, underscoring that the primary metric depends on difficult boundary judgments, even after independent adjudication.

We evaluate one fixed ontology-guided Gemini prompt containing definitions, boundary and suppression rules, annotated illustrative snippets, and sequential entity-tagging and relation-generation instructions. We did not ablate these components or vary the examples, retrieval, prompt, or decoding controls; their individual contributions cannot be inferred from the comparison. Provider-side changes may also be mixed with run-to-run generation variability. The prompt also requires an unscored relation graph alongside entity tags, so this condition does not isolate entity-only generation. This study does not evaluate relation extraction. Some labels are absent from development despite appearing in evaluation; for example, *Interpreter Present* has zero development spans and 26 evaluation spans.

Docket entries almost never contain all the detail available in the underlying sentencing judgment. Our scope reflects the availability and convenience of the SCALES docket collection and the cost of acquiring additional documents, not an assumption that judgments are less uniform or harder to parse. PACER’s electronic access schedule specifies \$0.10 per page, subject to caps and exemptions (United States Courts, n.d.). Extracting and comparing information from the judgments remains future work.

10 Ethical Considerations

Federal docket records are publicly accessible, but they can contain names and details about punishment, health, treatment, finances, immigration status, and other sensitive matters. Public access does not remove privacy or re-identification risk; work on large open legal corpora likewise emphasizes that such data may contain private or harmful material (Henderson et al., 2022). During this study, source dockets and record-level model outputs were kept in access-controlled storage; the paper presents aggregate results and reviewed redacted or explicitly synthetic examples. The public dataset release includes docket text, annotations, and case/entry identifiers, as described in the availability statement. Its public availability does not remove the privacy risks of linking, aggregating, or reusing these records.

Automated extraction can make court data easier to analyze, but it can also make sensitive information easier to aggregate, search, and use outside its original context, creating privacy risks from data processing and linkage (National Institute of Standards and Technology, 2020). Errors may omit a condition, attach a duration or penalty to the wrong event, or turn an ambiguous phrase into an apparently authoritative database field. These possible downstream harms motivate contextual validation and human oversight (Tabassi, 2023). The systems are research tools, not sources of legal advice or substitutes for reviewing the underlying docket. Their outputs should not be used without human verification for decisions about individuals, surveillance, risk assessment, or eligibility for legal relief.

The dataset also reflects the availability and conventions of its sampled records and the judgments embedded in the ontology. Uneven PACER coverage, inconsistent clerk language, and minority-label scarcity may create systematic blind spots. Reporting exact errors and label support makes these limitations more visible, but does not eliminate them. Dataset documentation helps delimit claims about generalization across populations (Bender and Friedman, 2018); neither that documentation nor aggregate performance demonstrates fairness in downstream sentencing uses.

Acknowledgments

Computing resources were provided by the Georgia Institute of Technology in Atlanta, Georgia, USA, through access to the Partnership for an Advanced

Computing Environment (PACE) (Partnership for an Advanced Computing Environment, 2026).

References

- Ting Wai Terence Au, Vasileios Lampsos, 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. Association for Computational Linguistics.
- Emily M. Bender and Batya Friedman. 2018. [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2017. [Extracting contract elements](#). In *Proceedings of the 16th International Conference on Artificial Intelligence and Law*, pages 19–28. Association for Computing Machinery.
- Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. [LEGAL-BERT: The muppets straight out of law school](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904. Association for Computational Linguistics.
- Gemini Team, Google. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). Technical report, Google.
- Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. [Pile of law: Learning responsible data filtering from the law and a 256GB open-source legal dataset](#). In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc.
- Rinke Hoekstra, Joost Breuker, Marcello Di Bello, and Alexander Boer. 2007. [The LKIF core ontology of basic legal concepts](#). In *Proceedings of the 2nd Workshop on Legal Ontologies and Artificial Intelligence Techniques*, volume 321 of *CEUR Workshop Proceedings*, pages 43–63.
- Shizhou Huang, Bo Xu, Yang Yu, Changqun Li, and Xin Alex Lin. 2025. [GuideNER: Annotation guidelines are better than examples for in-context named entity recognition](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24159–24166.
- 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. Association for Computational Linguistics.
- Alice Saebom Kwak, Cheonkam Jeong, Gaetano Vincent Forte, Derek E. Bambauer, Clayton T. Morrison, and Mihai Surdeanu. 2023. [Information extraction from legal wills: How well does GPT-4 do?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4336–4353. Association for Computational Linguistics.
- Drish Mali, Rubash Mali, and Claire Barale. 2024. [Information extraction for planning court cases](#). In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 97–114. Association for Computational Linguistics.
- National Institute of Standards and Technology. 2020. [NIST privacy framework: A tool for improving privacy through enterprise risk management, version 1.0](#). Technical report, National Institute of Standards and Technology.
- Naoaki Okazaki. 2007. [CRFsuite: A fast implementation of conditional random fields \(CRFs\)](#).
- Adam R. Pah, David L. Schwartz, Sarath Sanga, Charlotte S. Alexander, Kristian J. Hammond, Luís A. N. Amaral, and SCALES OKN Consortium. 2022. [The promise of AI in an open justice system](#). *AI Magazine*, 43(1):69–74.
- Vasile Pais, Maria Mitrofan, Carol Luca Gasan, Alexandru Ianov, Corvin Ghita, Vlad Silviu Coneschi, and Andrei Onut. 2024. [LegalNERo: A linked corpus for named entity recognition in the Romanian legal domain](#). *Semantic Web*, 15(3):831–844.
- Partnership for an Advanced Computing Environment. 2026. [PACE: Partnership for an advanced computing environment](#). <https://pace.gatech.edu/>.
- Pedro Vitor Quinta de Castro and Nádia Félix Felipe Da Silva. 2025. [Labor lex: A new Portuguese corpus and pipeline for information extraction in Brazilian legal texts](#). In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 184–206. Association for Computational Linguistics.
- SCALES-OKN. n.d. [Using SCALES’ platform](#). Accessed 2026-09-25.
- SentencingStats. n.d. [Federal judge sentencing analyzer](#). Accessed 2026-09-25.
- Răzvan-Alexandru Smădu, Ion-Robert Dinică, Andrei-Marius Avram, Dumitru-Clementin Cercel, Florin Pop, and Mihaela-Claudia Cercel. 2022. [Legal named entity recognition with multi-task domain adaptation](#). In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 305–321. Association for Computational Linguistics.
- Elham Tabassi. 2023. [Artificial intelligence risk management framework \(AI RMF 1.0\)](#). Technical Report NIST AI 100-1, National Institute of Standards and Technology.

- The Ohio State University Interdiction Lab and Michael Lissner. 2025. [Sentencing research without USSC data: Strategies and lessons learned](#). *Federal Sentencing Reporter*, 37(2):123–128.
- United States Courts. n.d. [Electronic public access fee schedule](#). Accessed 2026-09-25.
- Thomas Vacek, Dezhao Song, Hugo Molina-Salgado, Ronald Teo, Conner Cowling, and Frank Schilder. 2019. [Litigation analytics: Extracting and querying motions and orders from US federal courts](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 116–121. Association for Computational Linguistics.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547. Association for Computational Linguistics.
- Jonathan J. Wroblewski. 2025. [Sentencing “boat defendants”: Breaking the U.S. Sentencing Commission’s monopoly on gathering data on federal sentencing practices, and why it matters](#). *Federal Sentencing Reporter*, 37(2):147–151.
- Zongyue Xue, Huanghai Liu, Yiran Hu, Yuliang Qian, Yajing Wang, Kangle Kong, Chenlu Wang, Yun Liu, and Weixing Shen. 2024. [LEEC for judicial fairness: A legal element extraction dataset with extensive extra-legal labels](#). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 7527–7535.
- Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. [GLiNER: Generalist model for named entity recognition using bidirectional transformer](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376. Association for Computational Linguistics.

A Docket Example and Implementation Details

Figure 3 shows a randomly selected docket entry with names, dates, and identifying case information removed.

PACER Criminal Docket Entry Case: [REDACTED]
Minute Entry for proceedings held before Judge [REDACTED]: Sentencing held on [DATE] for [REDACTED], Count(s) 3, BOP 61 Months / 15 years supervised release / standard, special and mandatory conditions of release / no fine / \$100 special assessment / \$20,000 restitution. (Court Reporter [REDACTED].) (Entered: [DATE])
Randomly selected source entry; names, dates, case metadata, and other identifying information are redacted.

Figure 3: Redacted PACER sentencing docket entry. The compact record combines count references, sentencing events, durations, and monetary conditions in one semi-structured text field.

Configuration selection. Supervised-model configurations and GLiNER thresholds were selected using pooled out-of-fold predictions from grouped five-fold cross-validation on the 163 development entries. The held-out evaluation set was not used for selection. Stochastic configurations were ranked by mean pooled exact-span F1 across seeds 13, 37, and 73. Selected supervised models were refit on all development entries. Gemini’s prompt was fixed. Table 7 lists the search ranges and selected settings. The accompanying manuscript source package contains configuration/experiment-settings.json and configuration/model-environment.lock, recording seeds and the archived model-environment dependency versions.

Gemini inference and parsing. We ran five passes with models/gemini-2.5-flash through the Google Gen AI SDK in August 2026. Requests specified only the model identifier and prompt; generation settings and structured-output controls were not explicitly set. Appendix E reproduces the prompt template, omitting separator-only lines for readability. The prompt requests inline entity tags and relations, but only entity spans are evaluated.

The parser accepts text blocks or JSON, converts entity tags to character offsets, maps legacy label names to canonical labels, and checks that the annotated text reconstructs the source. Empty outputs or recorded request errors cause the final scoring preflight to fail. Outputs that cannot reconstruct

the source remain in evaluation: parsed entities receive out-of-document offsets, making them false positives under exact scoring and leaving all reference spans unmatched. Relaxed scoring instead compares their parsed entity text. In the reported Gemini passes, source reconstruction failed for 29, 24, 37, 34, and 30 of the 100 responses, respectively (154 of 500 responses across passes). These are repeated responses to the same 100 entries, not 500 distinct entries. The scored outputs contained no empty responses or recorded request errors.

CRFsuite. The tokenizer identifies currency and numeric expressions before general alphanumeric tokens and treats hyphens as separate tokens. Features include original and lowercase token text, prefixes, suffixes, token shape, capitalization, digit and punctuation indicators, sentencing lexicons, and the two preceding and following tokens.

GLiNER training and prediction. Both GLiNER conditions used urchade/gliner_medium-v2.1. Each overlap is divided at its midpoint; a prediction is retained only when its midpoint belongs to that window’s assigned region. Predictions were then mapped to document offsets. For duplicate spans, we retained the highest-scoring prediction. Remaining candidates were ranked by score, with ties resolved by earlier start, longer span, and label; overlapping lower-ranked predictions were removed.

Encoder training and prediction. We used answerdotai/ModernBERT-base and nlpauub/legal-bert-base-uncased. For training, each reference span was assigned to the window that contained it fully with the most surrounding context. Tokens belonging to partial or duplicate copies of that span in other windows were excluded from the loss. If a token overlapped multiple reference spans, all tokens intersecting those spans were excluded from the loss. All original reference spans remained in evaluation.

At inference, duplicate spans were resolved using the highest mean token probability. Other overlapping predictions were resolved by descending score, and retained spans were sorted into document order.

Relaxed matching. Span text was lowercased, whitespace was collapsed, and punctuation other than dollar signs was removed. The scorer then

System	Tested settings	Selected settings
CRFsuite	$c_1 \in \{0, .05, .1\}; c_2 \in \{.01, .05, .1\}$	$c_1 = c_2 = .1$; L-BFGS; 200 iterations; all transitions
Encoders	Learning rate $\{1, 2\} \times 10^{-5}$	2×10^{-5} ; 6 epochs; batch 2; weight decay .01; warmup .1; gradient norm 1.0
GLiNER zero-shot	Threshold .2–.6, step .1	Threshold .4
GLiNER fine-tuned	Encoder/other rates ($10^{-5}, 5 \times 10^{-5}$) or ($2 \times 10^{-5}, 10^{-4}$); 3 or 6 epochs; thresholds .3, .4, .5	($2 \times 10^{-5}, 10^{-4}$); 6 epochs; threshold .5; batch 4/device; accumulation 2; weight decay .01; warmup .1; cosine schedule
Windows	Fixed processing settings	GLiNER: 384 words, overlap 64; ModernBERT: 1,024 tokens; Legal-BERT: 512 tokens; encoder overlap: 128 content tokens
Gemini	Fixed prompt; 5 passes	1-second spacing; at most 5 attempts; exponential backoff from 10 seconds

Table 7: Configuration search and selected implementation settings. Both encoders and fine-tuned GLiNER use seeds 13, 37, and 73.

greedily matched predictions and references one-to-one when their labels agreed and either normalized text contained the other. References are visited in annotation order (document order for source-aligned markup); each is paired with the first eligible, unused prediction in prediction-markup order. Matches are not ranked by confidence or overlap length.

B Per-Label Results

Figures 4 and 5 report per-label recall.

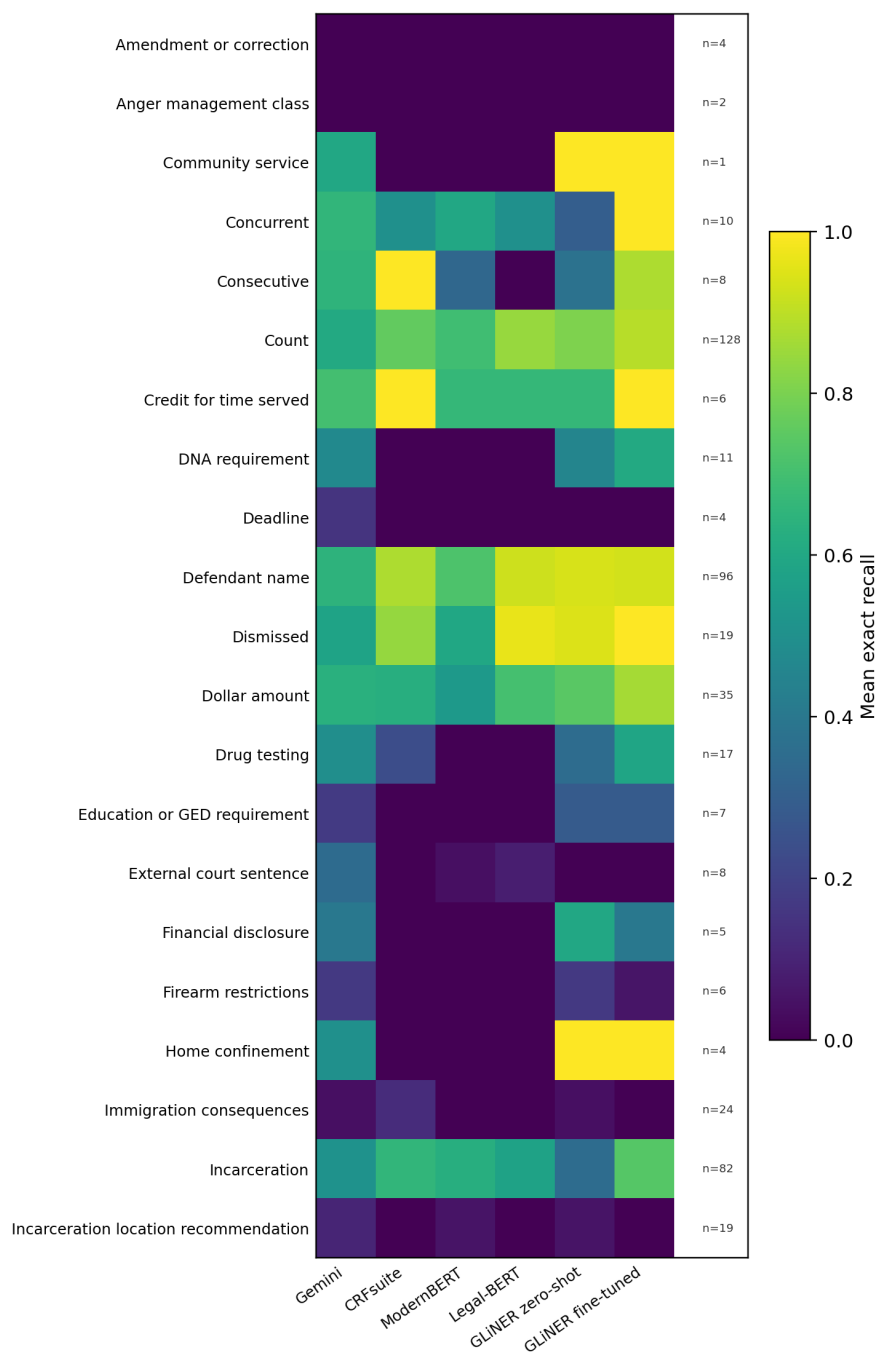


Figure 4: Exact-span recall by label on the 100-record evaluation set (panel 1 of 2). Row annotations give evaluation reference support. Only labels with reference support are shown; low-support rows should be interpreted descriptively.

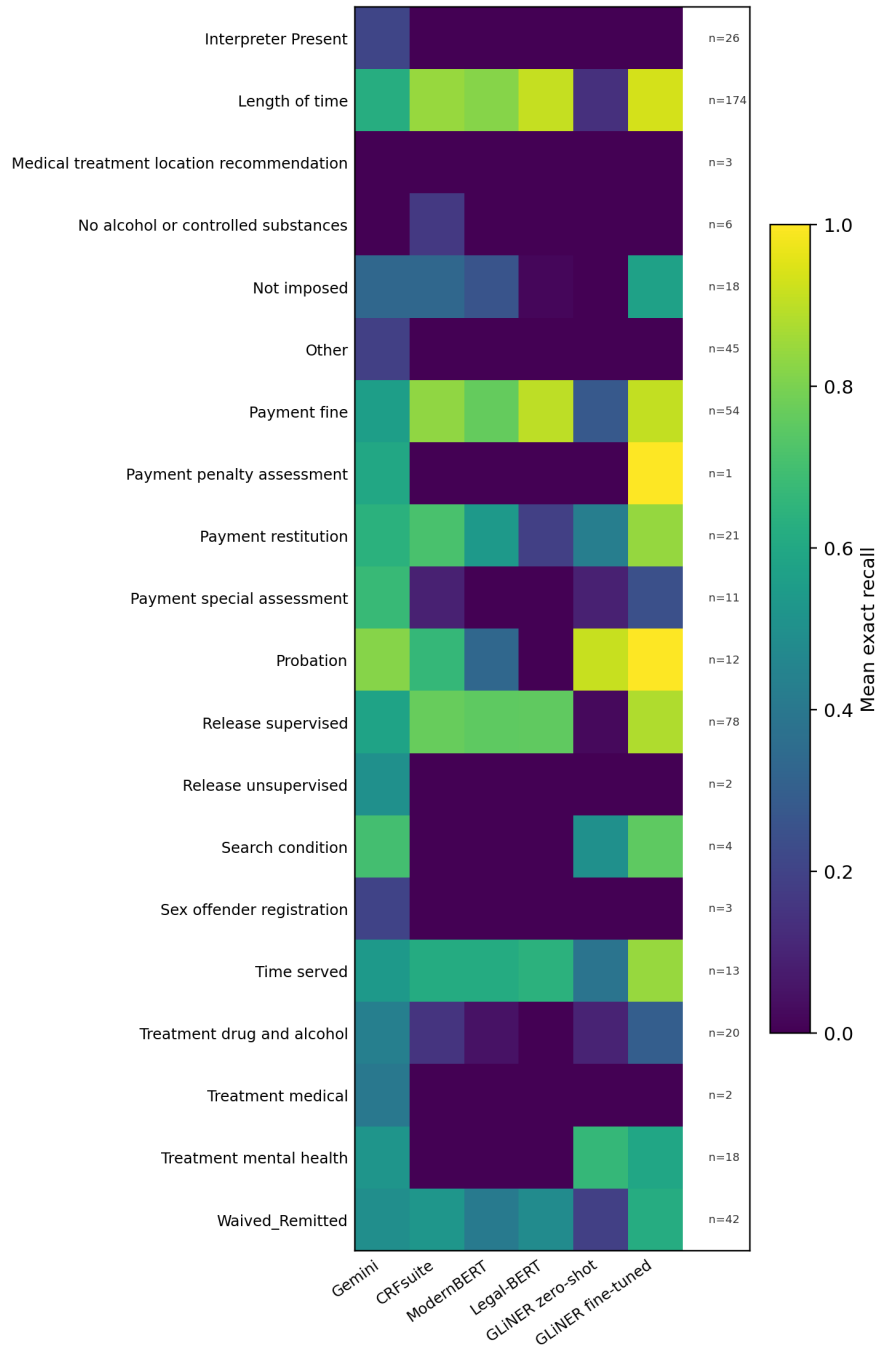


Figure 5: Exact-span recall by label on the 100-record evaluation set (panel 2 of 2). Row annotations give evaluation reference support. Only labels with reference support are shown; low-support rows should be interpreted descriptively.

C Run-Level Results

System	Run	Exact F1	Relaxed F1
Gemini 2.5 Flash	pass 1	.491	.802
Gemini 2.5 Flash	pass 2	.504	.790
Gemini 2.5 Flash	pass 3	.428	.794
Gemini 2.5 Flash	pass 4	.419	.771
Gemini 2.5 Flash	pass 5	.474	.777
CRFsuite	fixed	.655	.707
ModernBERT	run 1	.490	.659
ModernBERT	run 2	.482	.631
ModernBERT	run 3	.482	.645
Legal-BERT	run 1	.565	.644
Legal-BERT	run 2	.580	.651
Legal-BERT	run 3	.561	.642
GLiNER zero-shot	fixed	.345	.432
GLiNER fine-tuned	run 1	.696	.754
GLiNER fine-tuned	run 2	.712	.773
GLiNER fine-tuned	run 3	.701	.758

Table 8: Every independently scored provider pass, stochastic refit, and deterministic run. No best run is selected.

D Complete Ontology and Support

Table 9 gives all 44 labels and their reference-span support. Names retain the experimental wording. Full definitions and boundary guidance appear in Appendix F.

Canonical label	Group	Dev.	Eval.
Count	Anchor	219	128
Defendant name	Metadata	154	96
Interpreter Present	Metadata	0	26
Redacted	Metadata	1	0
Amendment or correction	Metadata	1	4
Incarceration	Event	129	82
Time served	Event	23	13
Credit for time served	Attribute	13	6
Suspended sentence	Event	0	0
Home confinement	Event	1	4
Release supervised	Event	126	78
Release unsupervised	Event	3	2
Probation	Event	15	12
External court sentence	Anchor	17	8
Incarceration location recommendation	Event	7	19

Table 9: Complete label inventory and span support in the 163-entry development pool (Dev.) and 100-entry evaluation set (Eval.).

Canonical label	Group	Dev.	Eval.
Medical treatment location recommendation	Event	0	3
Payment fine	Event	58	54
Payment restitution	Event	24	21
Payment special assessment	Event	13	11
Payment penalty assessment	Event	1	1
Anger management class	Event	0	2
Education or GED requirement	Event	3	7
Community service	Event	2	1
DNA requirement	Event	3	11
Financial disclosure	Event	3	5
Firearm restrictions	Event	1	6
Immigration consequences	Event	3	24
No alcohol or controlled substances	Event	1	6
No contact or restraining order	Event	0	0
Search condition	Event	1	4

Table 9 (continued).

Canonical label	Group	Dev.	Eval.
Sex offender registration	Event	0	3
Drug testing	Event	6	17
Treatment drug and alcohol	Event	9	20
Treatment medical	Event	0	2
Treatment mental health	Event	2	18
Other	Event	7	45
Concurrent	Connector	46	10
Consecutive	Connector	17	8
Dismissed	Attribute	45	19
Dollar amount	Attribute	68	35
Length of time	Attribute	291	174
Not imposed	Attribute	27	18
Waived Remitted	Attribute	46	42
Deadline	Attribute	0	4

Table 9 (continued).

E Gemini Prompt Template

The template below was used for all five Gemini passes. The placeholders {rules} and {text} were filled with the annotation rules and the docket entry, respectively. Appendix F reproduces the complete rules JSON; Appendix D reports label support. The prompt includes illustrative annotated snippets and requests both entity markup and relations; only entity spans are evaluated. Separator-only lines are omitted from this listing for readability.

You have two inputs:
 1) LabelingRules.json – the label set and span-tagging guidance: {rules}
 2) An unlabeled criminal sentencing docket entry: {text}

Your task has TWO steps.

STEP 1 – ENTITY TAGGING (INLINE)

Using LabelingRules.json, identify and tag entity spans in the docket entry text.

Mark entities in-text exactly as:
[Entity text][{{Label}}]

Span rules (strict):

- Tag only the smallest span that still makes the label unambiguous.
- For {{Length of time}}, tag the full duration expression (number + unit), e.g., "120 months", "5 years".
- Do NOT invent labels that are not present in LabelingRules.json.
- Use {{Other}} as a catch-all for events/conditions not captured by the label set, but keep the span minimal.
- If multiple counts are presented together (e.g., "Count(s) 2, 3" or "Count(s) 1-8"), tag them as ONE grouped Count span (a count block).
- If counts are mentioned separately in different parts of the text, tag them as separate {{Count}} spans.
- Even if a grouped block is described as concurrent, keep it grouped as one Count span.

Time-served rules:

- When "time served" or "time-served" is itself the sentence imposed, tag only those words as {{Time served}}.
- In that situation, {{Time served}} replaces the {{Incarceration}} anchor for that sentence.
- Do NOT additionally tag "sentenced," "sentenced to," or the same time-served text as {{Incarceration}}.
- Do NOT include attached dates, parenthetical day counts, or other durations in the {{Time served}} span.
- Do NOT separately tag attached dates, day counts, or durations as {{Length of time}}.
- When prior custody is credited against a separately stated incarceration sentence, tag only the words expressing the credit as {{Credit for time served}}.
- {{Credit for time served}} does not replace the separately stated {{Incarceration}} sentence. Tag that incarceration sentence and its duration normally.
- Do NOT include attached dates, day counts, or durations in the {{Credit for time served}} span.
- Do NOT separately tag attached dates, day counts, or durations as {{Length of time}}.

Examples:

- "sentenced to [TIME-SERVED][{{Time served}}] (241 days)"
- "[84 months][{{Length of time}} [imprisonment][{{Incarceration}} with [credit for time served][{{Credit for time served}}] from 10/21/2015 to 2/11/2016"

Redundancy suppression (try to avoid duplicate tags):

- Do NOT repeatedly tag the same {{Defendant name}} if it is restated; tag it once (prefer first mention).
- Do NOT repeatedly tag the same {{Count}} reference if it is restated; tag it once, unless a later mention is necessary to support a new event→count edge that would otherwise be ambiguous.

- More generally, avoid tagging repeated or summary restatements of the same entity when one tag already supports all required relations.
- Do NOT repeatedly tag the same {{Interpreter Present}} if it is restated; tag once.

Canonical incarceration anchor rule:

- Create an {{Incarceration}} anchor for each distinct count-specific imprisonment TERM when the terms differ.
Example: "120 months as to Count 1 and 60 months as to Count 11" requires TWO {{Incarceration}} anchors.
- If a single imprisonment term clearly applies to multiple counts with the SAME duration, use ONE {{Incarceration}} anchor connected to each applicable {{Count}}.
- If multiple words could serve as the incarceration anchor for the SAME term, tag only ONE span, preferably the span closest to the duration or count reference.
- All duration, count, and concurrency/consecutiveness relations must attach to the correct {{Incarceration}} anchor(s).
- Exception: when "time served" is itself the sentence imposed, use {{Time served}} instead of creating an {{Incarceration}} anchor for that sentence.

Drug-testing rules:

- {{Drug testing}} is separate from {{Treatment drug and alcohol}}.
- Tag only the minimal drug- or alcohol-testing phrase as {{Drug testing}}.
- Do NOT include modifiers such as "mandatory" in the {{Drug testing}} span.
- Tag treatment, counseling, evaluation, or rehabilitation as {{Treatment drug and alcohol}}.
- When testing and treatment are stated together, tag the testing words and treatment words as separate entities whenever the wording permits.
- When a drug-testing requirement is suspended, tag "suspended" or "suspends" as {{Not imposed}} and connect it to the {{Drug testing}} event.
- Do NOT use {{Drug testing}} for unrelated testing, such as polygraph, DNA, medical, or psychological testing.

Examples:

- "submit to mandatory [drug testing][{{Drug testing}}]"
- "participate in [drug testing][{{Drug testing}}] and [treatment][{{Treatment drug and alcohol}}]"
- "[suspends][{{Not imposed}}] the mandatory [drug testing requirement][{{Drug testing}}]"

Assessment suppression rule:

- Do NOT tag {{Payment special assessment}} or {{Payment penalty assessment}} unless the docket explicitly states that the assessment was waived, remitted, or not imposed.
- If either type of assessment is imposed normally, do NOT tag the assessment or its associated {{Dollar amount}}.
- Keep the historical label distinction: use {{Payment special assessment}} for "special assessment," "assessment," SPA, SA, or MSA, and use {{Payment penalty assessment}} for "penalty assessment" or "special penalty assessment."

Supervised release concurrency assumption:

- Supervised release terms are assumed to run concurrently by default.
- Do NOT tag `{{Concurrent}}` or `{{Consecutive}}` for supervised release.
- Only the LONGEST supervised release term matters.
- If multiple supervised release terms are listed, even for different counts, tag ONLY the supervised release term with the greatest duration and its `{{Length of time}}`.
- Do not tag the shorter supervised release terms.
- Supervised release does NOT need to be linked to a `{{Count}}`. It may float without an event-to-count edge.

STEP 2 – RELATIONS (GRAPH EDGES)

After tagging entities inline, output relations as a list of generic, unlabeled directed edges between tagged items.

Relations have NO names.

Output format:

- AnnotatedText: the full docket entry with inline tags
- Relations: a JSON array of 2-element arrays, each [FROM, TO], where FROM and TO must be exact substrings copied from AnnotatedText INCLUDING their `{{Label}}`.

Example edge:

```
[ "[264 months]{{Length of time}}",
  "[CBOP]{{Incarceration}}"]
```

RELATION RULES (STRICT)

A) Attributes must attach to an event

- 1) Every `[...{{Length of time}}` must connect to the sentencing event it modifies.
- 2) Every `[...{{Dollar amount}}` must connect to the payment or sentencing event it modifies.
- 3) Every status attribute, such as `[waived]{{Waived_Remitted}}`, `[remitted]{{Waived_Remitted}}`, `[not imposed]{{Not imposed}}`, or `[dismissed]{{Dismissed}}`, must connect to the event it modifies.
- 4) Attributes cannot stand alone. If you cannot find the modified event, do not tag the attribute.
- 5) Every `[...{{Deadline}}` must connect to the payment event it modifies.

Time-served relation exception:

- `{{Time served}}` and `{{Credit for time served}}` do not require relation edges for this experiment.
- Do not create an `{{Incarceration}}` annotation solely to provide a relation endpoint for either label.
- Use Time served when it is the sentence imposed, and use Credit for time served only when prior custody is credited against a separately stated incarceration term.

Duration and amount de-duplication:

- If the same duration or amount is restated or summarized for the same event, create only ONE attribute edge for that event.

B) Sentencing events may connect to a Count

- If the docket entry explicitly specifies the count for an event, connect the event to the `{{Count}}`.
- If not specified, do NOT guess; the event may float.

Priority:

- `{{Incarceration}}` and payment events, including fine, restitution, penalty assessment, and special assessment when waived, remitted, or not imposed, should be linked to `{{Count}}` when specified.

C) Concurrent / Consecutive connectors

- Tag only when explicitly stated.
- MUST connect to `{{Incarceration}}`.
- MUST also connect to the comparison target:
 - Prefer another `{{Incarceration}}`
 - Otherwise `{{External court sentence}}`
 - Fallback: `{{Count}}`

D) No relation names

- Output edges only as [FROM, TO].

FINAL INSTRUCTION

Process the docket entry and return the output in the specified format.

F Complete Annotation Rules

The following JSON specifies the complete experimental label inventory, definitions, and span guidance supplied to Gemini. Relation instructions appear in [Appendix E](#).

```
{
  "version": "2.8-experiment",
  "notes": [
    "This file defines the label set and span-tagging guidance only.",
    "All relation and graph-construction logic is specified in the prompt.",
    "This JSON follows the prompt label logic exactly where the prompt and error-category notes differ.",
    "If multiple counts are presented together, they must be tagged as ONE grouped Count span.",
    "Even if a grouped block is described as concurrent, keep it grouped as one Count span.",
    "Supervised release terms are assumed to run concurrently by default.",
    "Do NOT tag Concurrent or Consecutive for supervised release.",
    "If multiple supervised release terms are listed, tag ONLY the longest supervised release term and its Length of time.",
```

```

"Payment special assessment and Payment
penalty assessment should only be tagged when
the docket explicitly states that the
assessment was waived, remitted, or not
imposed. Normally imposed assessments and
their dollar amounts should not be tagged.",
"Drug testing is a separate event from drug
or alcohol treatment. Tag testing and
treatment separately when both are stated.",
"Time served and Credit for time served are
restored for compatibility with the labeled
experimental data.",
"When time served is itself the sentence
imposed, tag only the time-served expression
as Time served; do not additionally tag
'sentenced', 'sentenced to', or the same span
as Incarceration.",
"When prior custody is credited against a
separately stated incarceration sentence,
tag only the words expressing the credit as
Credit for time served.",
"Do not include attached dates, day counts,
or durations in Time served or Credit for
time served spans, and do not tag them as
Length of time.",
"For Drug testing, exclude modifiers such as
'mandatory' from the span. If the requirement
is suspended, annotate the suspension word as
Not imposed and connect it to Drug testing."
],
"labels": [
{
"label": "Count",
"category": "anchor",
"definition": "A single distinct charge or
a grouped block of charges referenced in a
docket entry.",
"span_rules": [
"Include the word 'Count' or 'Count(s)'
and the count number(s) or range(s).",
"If multiple counts are presented
together, tag them as ONE grouped Count
span.",
"If counts are mentioned separately in
different parts of the text, tag them as
separate Count spans.",
"Even if a grouped block is described as
concurrent, keep it grouped as one Count
span."
]
},
{
"label": "Defendant name",
"category": "metadata",
"definition": "The defendant's name.",
"span_rules": [
"Tag the defendant's name only.",
"Do not repeatedly tag the same Defendant
name if it is restated; tag it once,
preferably the first mention."
]
},
{
"label": "Interpreter Present",
"category": "metadata",
"definition": "The interpreter present at
the proceeding.",
"span_rules": [
"Tag the interpreter mention only.",
"Do not repeatedly tag the same
Interpreter if it is restated; tag once."
]
},
{
"label": "Redacted",
"category": "metadata",
"definition": "Indicates that information
has been redacted.",
"span_rules": [
"Tag the minimal redaction indicator
phrase."
]
},
{
"label": "Amendment or correction",
"category": "metadata",
"definition": "Indicates that the judgment
or entry was amended or corrected.",
"span_rules": [
"Tag the minimal amendment or correction
phrase when it clearly refers to the
judgment or entry."
]
},
{
"label": "Incarceration",
"category": "event",
"definition": "Sentence of imprisonment in
federal prison or Bureau of Prisons
custody.",
"possible_attributes": [
"Length of time"
],
"span_rules": [
"Include only the words that indicate the
prison sentence, such as 'imprisonment',
'custody of the Bureau of Prisons',
'CBOP', or 'BOP'.",
"If present, include the word 'custody'.",
"Create an Incarceration anchor for each
distinct count-specific imprisonment
term when the terms differ.",
"If a single imprisonment term clearly
applies to multiple counts with the same
duration, use ONE Incarceration anchor
connected to each applicable Count.",
"If multiple words could serve as the
incarceration anchor for the same term,
tag only ONE span, preferably the span
closest to the duration or count
reference.",
"Exception: when 'time served' is itself
the sentence imposed, do not tag
'sentenced', 'sentenced to', or the same
time-served span as Incarceration."
]
},
{
"label": "Time served",
"category": "event",
"definition": "Indicates that time already
served is itself the incarceration
sentence imposed.",
"span_rules": [
"Tag only the words expressing 'time
served' or 'time-served'.",
"Do not include 'sentenced', 'sentenced
to', or similar introductory wording.",
"Use Time served instead of an
Incarceration label when time served is
itself the sentence imposed."
]
}
]

```

```

        "Do not include attached dates,
        parenthetical day counts, or other
        durations in the Time served span.",
        "Do not separately tag attached dates, day
        counts, or durations as Length of time."
    ]
},
{
    "label": "Credit for time served",
    "category": "attribute",
    "definition": "Indicates that prior custody
    is credited against a separately stated
    incarceration sentence.",
    "span_rules": [
        "Tag only the words expressing the
        credit, ordinarily 'credit for time
        served' or a close variant.",
        "Do not include attached dates, day
        counts, or durations in the Credit for
        time served span.",
        "Do not separately tag attached dates, day
        counts, or durations as Length of time."
    ]
},
{
    "label": "Suspended sentence",
    "category": "event",
    "definition": "A sentence where jail or
    prison time is imposed but suspended.",
    "possible_attributes": [
        "Length of time"
    ],
    "span_rules": [
        "Include only 'suspended sentence' or a
        close proxy."
    ]
},
{
    "label": "Home confinement",
    "category": "event",
    "definition": "Court-ordered confinement to
    residence except approved outings.",
    "possible_attributes": [
        "Length of time"
    ],
    "span_rules": [
        "Tag only the home-confinement indicator
        phrase."
    ]
},
{
    "label": "Release supervised",
    "category": "event",
    "definition": "Supervised release term.",
    "possible_attributes": [
        "Length of time"
    ],
    "span_rules": [
        "Tag only the supervised release phrase.",
        "If multiple supervised release terms are
        listed, tag ONLY the supervised release
        term with the greatest duration.",
        "Do not tag shorter supervised release
        terms."
    ]
},
{
    "label": "Release unsupervised",
    "category": "event",
    "definition": "Unsupervised or
    non-reporting supervised release.",
    "possible_attributes": [
        "Length of time"
    ],
    "span_rules": [
        "Tag the unsupervised or non-reporting
        release phrase."
    ]
},
{
    "label": "Probation",
    "category": "event",
    "definition": "Probation sentence.",
    "possible_attributes": [
        "Length of time"
    ],
    "span_rules": [
        "Tag the probation phrase."
    ]
},
{
    "label": "External court sentence",
    "category": "anchor",
    "definition": "A referenced non-federal
    sentence used for concurrency or
    consecutiveness comparisons.",
    "possible_attributes": [
        "Length of time"
    ],
    "span_rules": [
        "Tag the minimal case or sentence
        reference."
    ]
},
{
    "label": "Incarceration location
    recommendation",
    "category": "event",
    "definition": "Recommendation regarding
    place of incarceration.",
    "span_rules": [
        "Tag only the incarceration-location
        recommendation phrase."
    ]
},
{
    "label": "Medical treatment location
    recommendation",
    "category": "event",
    "definition": "Recommendation regarding
    treatment location or facility.",
    "span_rules": [
        "Tag only the treatment-location
        recommendation phrase."
    ]
},
{
    "label": "Payment fine",
    "category": "event",
    "definition": "A court-imposed fine.",
    "possible_attributes": [
        "Dollar amount",
        "Deadline",
        "Not imposed",
        "Waived/Remitted"
    ],
    "span_rules": [
        "Tag only the fine event phrase."
    ]
},
{
    "label": "Payment restitution",

```

```

"category": "event",
"definition": "A restitution payment obligation.",
"possible_attributes": [
  "Dollar amount",
  "Deadline",
  "Not imposed",
  "Waived_Remitted"
],
"span_rules": [
  "Tag only the restitution event phrase."
]
},
{
  "label": "Payment special assessment",
  "category": "event",
  "definition": "A special assessment payment event.",
  "possible_attributes": [
    "Dollar amount",
    "Not imposed",
    "Waived_Remitted"
  ],
  "span_rules": [
    "Tag only when the docket explicitly states that the special assessment was waived, remitted, or not imposed.",
    "Do not tag a normally imposed special assessment or its associated Dollar amount.",
    "Use Payment special assessment for 'special assessment', 'assessment', SPA, SA, or MSA."
  ]
},
{
  "label": "Payment penalty assessment",
  "category": "event",
  "definition": "Fee ordered because of failure to meet specific obligation.",
  "possible_attributes": [
    "Dollar amount",
    "Not imposed",
    "Waived_Remitted"
  ],
  "span_rules": [
    "Include only the words 'penalty assessment' or 'special penalty assessment'.",
    "Tag only when the docket explicitly states that the penalty assessment was waived, remitted, or not imposed.",
    "Do not tag a normally imposed penalty assessment or its associated Dollar amount.",
    "If it just says 'assessment', SPA, SA, or MSA, use Payment special assessment instead."
  ]
},
{
  "label": "Anger management class",
  "category": "event",
  "definition": "Court-ordered anger management class or program.",
  "possible_attributes": [
    "Length of time"
  ],
  "span_rules": [
    "Tag only the anger-management condition phrase."
  ]
},
{
  "label": "Education or GED requirement",
  "category": "event",
  "definition": "Condition requiring education participation or GED completion.",
  "possible_attributes": [
    "Length of time"
  ],
  "span_rules": [
    "Tag only the education or GED condition phrase."
  ]
},
{
  "label": "Community service",
  "category": "event",
  "definition": "Community-service requirement.",
  "possible_attributes": [
    "Length of time"
  ],
  "span_rules": [
    "Tag only the community-service phrase."
  ]
},
{
  "label": "DNA requirement",
  "category": "event",
  "definition": "Requirement to provide a DNA sample.",
  "span_rules": [
    "Tag only the DNA-requirement phrase."
  ]
},
{
  "label": "Financial disclosure",
  "category": "event",
  "definition": "Requirement to provide financial disclosure or records.",
  "span_rules": [
    "Tag only the financial-disclosure phrase."
  ]
},
{
  "label": "Firearm restrictions",
  "category": "event",
  "definition": "Condition restricting firearm possession or access.",
  "span_rules": [
    "Tag only the firearm-restriction phrase."
  ]
},
{
  "label": "Immigration consequences",
  "category": "event",
  "definition": "Immigration-related consequence or condition.",
  "span_rules": [
    "Tag only the immigration-consequence phrase."
  ]
},
{
  "label": "No alcohol or controlled substances",
  "category": "event",
  "definition": "Condition prohibiting alcohol or controlled substances.",

```

```

    "span_rules": [
      "Tag only the prohibition condition
      phrase."
    ]
  },
  {
    "label": "No contact or restraining order",
    "category": "event",
    "definition": "No-contact or
    restraining-order condition.",
    "span_rules": [
      "Tag only the no-contact or
      restraining-order phrase."
    ]
  },
  {
    "label": "Search condition",
    "category": "event",
    "definition": "Condition allowing or
    requiring search or inspection.",
    "span_rules": [
      "Tag only the search-condition phrase."
    ]
  },
  {
    "label": "Sex offender registration",
    "category": "event",
    "definition": "Requirement to register as a
    sex offender.",
    "span_rules": [
      "Tag only the registration requirement
      phrase."
    ]
  },
  {
    "label": "Drug testing",
    "category": "event",
    "definition": "A requirement to undergo
    drug or alcohol testing, screening, or
    monitoring for substance use.",
    "span_rules": [
      "Tag only the minimal phrase expressing
      the drug- or alcohol-testing
      requirement.",
      "Do not include modifiers such as
      'mandatory' in the Drug testing span.",
      "Do not use Drug testing for drug or
      alcohol treatment, counseling,
      evaluation, or rehabilitation.",
      "When testing and treatment are stated
      together, tag the testing words as Drug
      testing and the treatment words
      separately as Treatment drug and
      alcohol.",
      "When a drug-testing requirement is
      suspended, tag 'suspended' or 'suspends'
      separately as Not imposed and connect it
      to the Drug testing event.",
      "Do not tag unrelated testing, such as
      polygraph, DNA, medical, or
      psychological testing, as Drug testing."
    ]
  },
  {
    "label": "Treatment drug and alcohol",
    "category": "event",
    "definition": "Drug or alcohol treatment
    requirement.",
    "possible_attributes": [
      "Length of time"
    ]
  },

```

```

    "span_rules": [
      "Tag the treatment requirement phrase
      only."
    ]
  },
  {
    "label": "Treatment medical",
    "category": "event",
    "definition": "Medical treatment
    requirement.",
    "possible_attributes": [
      "Length of time"
    ]
  },
  {
    "label": "Treatment mental health",
    "category": "event",
    "definition": "Mental health treatment
    requirement.",
    "possible_attributes": [
      "Length of time"
    ]
  },
  {
    "label": "Other",
    "category": "event",
    "definition": "Catch-all for events or
    conditions not captured by the label set.",
    "span_rules": [
      "Use Other only when no existing label
      fits.",
      "Keep the span minimal."
    ]
  },
  {
    "label": "Concurrent",
    "category": "connector",
    "definition": "Concurrent relationship
    indicator.",
    "span_rules": [
      "Tag 'concurrent' or 'concurrently'.",
      "Tag Concurrent only when the docket
      explicitly uses those words or clear
      variants.",
      "Do not tag Concurrent for supervised
      release."
    ]
  },
  {
    "label": "Consecutive",
    "category": "connector",
    "definition": "Consecutive relationship
    indicator.",
    "span_rules": [
      "Tag 'consecutive' or 'consecutively'.",
      "Tag Consecutive only when the docket
      explicitly uses those words or clear
      variants.",
      "Do not tag Consecutive for supervised
      release."
    ]
  },

```

```

        "label": "Dismissed",
        "category": "attribute",
        "definition": "Status attribute indicating
dismissed.",
        "span_rules": [
            "Tag 'dismissed' when used as a
disposition or status."
        ]
    },
    {
        "label": "Dollar amount",
        "category": "attribute",
        "definition": "A monetary amount.",
        "span_rules": [
            "Tag the full monetary amount including
currency symbol."
        ]
    },
    {
        "label": "Length of time",
        "category": "attribute",
        "definition": "A duration of time.",
        "span_rules": [
            "Tag the full duration expression,
including number and unit.",
            "Do not tag parenthetical day counts,
dates, or other details contained in a
Time served or Credit for time served
expression as Length of time."
        ]
    },
    {
        "label": "Not imposed",
        "category": "attribute",
        "definition": "Status attribute indicating
not imposed.",
        "span_rules": [
            "Tag 'not imposed'."
        ]
    },
    {
        "label": "Waived Remitted",
        "category": "attribute",
        "definition": "Status attribute indicating
waived or remitted.",
        "span_rules": [
            "Tag 'waived' or 'remitted'."
        ]
    },
    {
        "label": "Deadline",
        "category": "attribute",
        "definition": "A deadline by which a
payment must be made.",
        "span_rules": [
            "Tag the full deadline expression, such
as 'within 30 days' or 'by January 1,
2025'."
        ]
    }
]
}

```