Received: 23 February 2026 | Revised: 29 June 2026 | Accepted: 23 July 2026 | Published online: 10 August 2026

RESEARCH ARTICLE

Journal of Computational Law and Legal Technology
2026, Vol. 1, A9, 12 Pages
DOI: 10.47852/bonviewJCLLT62029445

BON VIEW PUBLISHING

Legal NER: Evaluating the Impact of LLM-Generated Annotations on NER Performance for Administrative Decisions

Harry Nan[1,*], Samaneh Khoshrou[2] and Johan Wolswinkel[1]

[1] *Department of Public Law and Governance, Tilburg University, The Netherlands*

[2] *Department of Intelligent Systems, Tilburg University, The Netherlands*

Abstract: Named entity recognition (NER) is a core information extraction (IE) task dependent on high-quality annotated data that are expensive and time-intensive to produce. Large language models (LLMs) offer a promising alternative through LLM-generated pseudo-annotations, yet their reliability in domain-specific legal settings remains insufficiently studied. This study investigates the use of LLM-generated annotations to expand the training set for supervised NER models applied to sentences from Dutch administrative decisions as a low-resource domain and language. To this end, LLM-based annotations of predefined legal entities are created using a schema-driven few-shot prompt, which are evaluated against a human-annotated dataset. Next, two different NER architectures are trained—a token-level NER model and a span-based NER–RE model (joint NER and relationship extraction (RE))—under three training settings: (1) human annotations only, (2) LLM-generated annotations only, and (3) models trained on LLM-generated annotations further fine-tuned on human annotations. The results indicate that LLMs can accurately generate annotations for legal entities that are explicitly defined in legislation but generate less reliable annotations for other legal entities that require deeper contextual understanding beyond what is explicitly stated in the text or prompt. Furthermore, fine-tuning a NER model trained on these LLM-generated annotations with human annotations turns out to slightly outperform models trained on human-annotated data only. Our findings highlight the potential of hybrid supervision strategies to scale low-resource legal NER tasks while maintaining human-level accuracy.

Keywords: named entity recognition, information extraction, relationship extraction, large language models, annotation

1. Introduction

Named entity recognition (NER) is considered a foundational technique for information extraction (IE) [1, 2]. By identifying and classifying relevant entities (e.g., persons, organizations, or dates), NER allows for a structured representation of unstructured texts. While large language models (LLMs) have already shown promising results for various tasks in natural language processing (NLP), such as text classification and question answering [3, 4], their application to IE tasks is more limited, as the generative nature of LLMs does not naturally align with the token-level sequence labeling required for IE tasks such as NER [5]. As a result, supervised transformer-based models like NER have remained the dominant approach within IE in domain-specific settings so far [1, 5].

Legal entity recognition (LER) specifically deals with the extraction of relevant "legal entities" from legal texts. Unlike domain-agnostic NER tasks that mostly focus on isolated entities (e.g., names, dates, or locations), these legal entities are determined by the (domain-specific) relationships that characterize the specific role of an entity (e.g., the recipient of a legal decision or the legal basis of that decision). LER is therefore not only a task of identifying (domain-agnostic) entities (e.g., persons) but also of assigning specific capacities among the identified entities (e.g., the specific person receiving a decision).

Considering legal information extraction (LIE) in general, transformer-based architectures for NER have already shown strong multilingual performance [6–8], but this performance typically relies on large volumes of high-quality annotated data. In many low-resource domains, the acquisition of such annotations is both costly and labor-intensive. Recent studies have therefore explored the use of LLMs to generate pseudo-annotations for NER tasks [9–11]. However, since these studies focus primarily on general-purpose or high-resource NER datasets, they provide limited guidance on how to handle domain-specific structures and relationships between entities.

In this work, we make two main contributions to the field of legal NLP in general and legal information extraction in particular. First, we introduce a new annotated dataset specifically designed for legal entity extraction in the low-resource domain of Dutch administrative decisions[1]. This dataset provides a valuable

***Corresponding author:** Harry Nan, Department of Public Law and Governance, Tilburg University, The Netherlands. Email: H.Nan@tilburguniversity.edu

[1] GitHub link to the created LEAD-NL dataset

© The Author(s) 2026. Published by BON VIEW PUBLISHING. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

benchmark for research in IE from legal texts, thereby addressing a notable gap in available Dutch-language resources. Second, we investigate the potential of LLMs to generate pseudo-annotations that can augment and enhance the performance of supervised NER models. Through a systematic evaluation, we assess both the accuracy of LLM-generated pseudo-annotations and the resulting improvements in NER performance when these annotations are incorporated into training data for extracting legal entities. This dual focus offers insights into how LLMs can be effectively leveraged to scale low-resource datasets and to advance hybrid annotation strategies that combine human expertise with automated labeling.

Specifically, this study addresses two research questions:

1) To what extent can LLMs generate accurate pseudo-annotations, compared to human-labeled annotations?
2) To what extent does the inclusion of LLM-generated pseudo-annotations improve the performance of supervised NER models?

This study shows that although LLMs can generate large volumes of pseudo-annotations, their reliability is uneven. Nevertheless, augmenting human annotations with LLM-generated annotations can increase the performance of NER tasks by expanding training coverage in a scalable way, while the inclusion of human-annotated data ensures the accuracy and reliability of the NER model.

2. Related Work

2.1. Legal information extraction

LIE aims to automatically identify and structure legally relevant information from legal texts, thereby supporting applications such as case retrieval and legal analytics [2, 12], compliance and contract analysis [12], and decision support in legal and administrative contexts [13, 14]. Within LIE, a range of LER techniques have been explored, including rule-based approaches [12, 14], traditional machine-learning methods [14, 15], and more recently transformer-based models that learn contextual representations of legal language [1, 16]. While these techniques have demonstrated strong performance, existing research in LIE has predominantly focused on certain types of legal documents where annotated benchmarks are more readily available [2], particularly court judgments, but also legal contracts [12] and wills [17]. However, this predominant orientation toward certain types of legal documents raises questions about the transferability of existing LIE methods to other types of legal documents, such as administrative decisions, which may differ substantially in structure and legal context [13].

Given that most downstream tasks in LIE depend on accurate identification of entities that are legally relevant, NER constitutes a key challenge here [1]. Recent work has increasingly applied transformer-based NER models to low-resource legal domains and underrepresented languages, including Portuguese [6], German [7], and Turkish [8], where annotated corpora are scarce. However, limited data availability across multilingual and domain-specific settings continues to constrain NER performance [18, 19]. Such bottlenecks in entity recognition directly affect subsequent downstream tasks and overall system performance.

One prominent example of such a downstream task is Relationship Extraction (RE), which aims to identify semantic relationships between recognized entities [18, 20]. To better capture the structural dependencies inherent in legal texts, a new line of research has explored joint modeling approaches that combine NER and RE. These models jointly optimize both entity and relationship prediction, penalizing errors in both tasks and thereby explicitly modeling interdependencies between entities and their relationships. Although joint approaches have demonstrated more consistent improvements for RE [21], their impact on NER performance remains inconclusive: while some studies report modest gains, particularly in domains with strong structural dependencies [22, 23], others find little or no advantage over baseline NER models [21]. This highlights an unresolved question in legal IE research: how can domain-specific structures and relationships be most effectively leveraged to improve NER performance?

Apart from data scarcity, the complex structure of legal texts strongly affects annotation quality. Legal texts often contain domain-specific characteristics such as specialized vocabulary and formal syntax [14]. In addition to linguistic structure, visual and layout-related information can be valuable for LIE: Document Layout Analysis (DLA) has been applied to incorporate visual cues (such as headings of documents) into NER tasks, improving performance in document-centric settings [24]. Similarly, document segmentation techniques restrict entity extraction to predefined zones of legal texts when entity locations and structural patterns are predictable [14]. In the absence of such visual cues or predefined zones, however, textual indicators are needed to limit the amount of text from which legal entities are to be extracted [15].

2.2. LLM-driven NER

To mitigate the scarcity of available annotated data, recent work has focused on developing few-shot or limited-data approaches [10, 11] and weakly supervised or other limited supervision techniques [9, 25] to reduce the reliance on large human-annotated datasets while still enabling supervised models to achieve strong and reliable performance [1, 5]. To counter this bottleneck of scarce annotated datasets, LLMs have been used to reduce the required work needed to obtain high-quality annotated data. Some studies have applied LLMs to assist legal researchers by improving annotation efficiency and annotator agreement [26] or applied (hybrid) methods to perform IE tasks directly [13, 17]. Other studies have shown that LLM-generated annotations, when combined with human annotations, can achieve human-level performance in domain-specific settings [9, 10, 20].

Since domain specificity poses an additional challenge for legal NER, adapting transformer-based NER models, such as Bidirectional Encoder Representations from Transformers (BERT), through domain-specific pretraining has been shown to enhance performance in both legal classification and LER tasks [16]. When applying LLMs, these structural and stylistic cues, while often intuitive to human readers with legal expertise, are not always explicitly encoded in the text. As a result, LLMs may struggle to infer implicit cues that signal legal meaning, which can limit their performance on legal NER tasks [27]. To address this limitation, some studies have incorporated domain-specific relationships directly into prompts, thereby providing additional structural guidance and improving LLM performance on NER tasks [20].

Ensuring accuracy and reliability in LLMs requires more than just scale, as it depends critically on the quality of supervision. Recent studies show that models fine-tuned on human-curated datasets consistently achieve superior or at least comparable performance to those trained solely on LLM-generated labels [28]. Moreover, even a relatively small proportion of human-labeled examples can improve the accuracy and reasoning ability of LLMs

[29]. Human-labeled data are therefore particularly effective in improving correctness, depth of reasoning, and robustness to ambiguity, while synthetic data such as pseudo-annotations provide unmatched scalability. Consequently, the most effective strategy seems to be a hybrid one, where human annotations ensure quality and reliability, and synthetic data are employed to broaden coverage in a cost-efficient manner.

A common strategy for using LLMs in annotation generation is to frame NER as a generative or question-answering (QA) task [4, 5]. Research shows that explicit label definitions, fixed output formats, and decomposed QA strategies improve NER performance [4]. Additionally, integrating entity extraction with RE has been found to be effective [11, 20]. Few- and zero-shot prompting techniques have also been applied to NER [4], showing that LLMs can achieve reliable entity labels in low-resource settings, with few-shot techniques often resulting in higher performance [5]. However, most research focuses on general-purpose datasets and largely ignores domain-specific contexts, relationships between entities, and the impact of pseudo-annotations on supervised NER model performance, as these are often used to replace human annotations [1, 9, 18]. However, such relationships are of utmost importance in the legal domain, where legal entities derive their meaning from their specific roles and the relationships defining them.

In summary, prior research highlights two key challenges: (1) supervised methods remain dominant for NER but are costly to apply in low-resource legal domains, and (2) while LLMs offer promising strategies for reducing annotation burden, most work has focused so far on general-purpose datasets rather than domain-specific (legal) corpora. Together, these gaps call for a systematic evaluation of the question of whether LLM-generated pseudo-annotations, enriched with domain-specific relationships, can improve NER performance in legal texts. Our work addresses both challenges by evaluating LLM-generated annotations in a low-resource domain and language, namely, Dutch administrative decisions, and examining the impact of incorporating LLM-generated annotations into two different supervised NER models.

3. Methodology

Figure 1 shows an overview of the methodology of this paper. Section 3.1 characterizes administrative decisions as a specific type of legal document and explains which legal entities can be extracted from these documents. Section 3.2 demonstrates how the dataset is obtained and how candidate sentences for IE are selected from full documents, resulting in the dataset used in this study. Sections 3.3 and 3.4 describe the annotation process for human and LLM-generated annotations, respectively. Section 3.5 explains the experimental setup used in this study, followed by a description of the two different NER models in Section 3.6.

3.1. Legal entities in administrative decisions

Following the Model Rules on EU Administrative Procedure[2], an administrative decision can be understood as a specific type of "administrative action addressed to one or more individualized public or private persons which is adopted unilaterally by a [public authority] to determine one or more concrete cases with a legally binding effect." In the same vein, though less explicit about its constitutive elements, the Dutch General Administrative Law Act defines an administrative decision as a type of administrative action that is not of a general nature[3].

From the legal definition of administrative decision provided for in the Model Rules, five key entities of administrative decisions can be derived, which should be present in every administrative decision, irrespective of its substance. See Table 1. First, two types of legal subjects [30] are involved in every administrative decision: (1) a "person" receiving the decision (to one or more individualized persons) (*recipient*) and (2) a "public authority" issuing the decision (*authority*). Next, every decision has a "legally binding effect," which means that certain rights or obligations are conferred on the recipient (legal act [30]). In particular, such a legal effect consists of the combination of some form of (4) *legal action* (e.g., to grant or to revoke) (3) related to some *legal object* (e.g., a license or a fine). Finally, since an administrative decision is adopted "unilaterally," that is, without the consent of the recipient, it should always have (5) a *legal basis* ([30]): a statutory provision that confers competence to a specific public authority to issue a specific type of administrative decision. While other legal entities can also be present in certain types of administrative decisions (such as breach of a legal provision in case of an enforcement decision), these five entities can be assumed to be present in all types of administrative decisions.

3.2. Dataset creation

To answer the research questions and to ensure generalizable results, a highly diverse set of administrative decisions is created. This diversity covers both variety in administrative decisions (such

[2]Article III-2(1) and (2) of the ReNEUAL Model Rules on EU Administrative Procedure

[3]Article 1:3(2) of the Algemene wet bestuursrecht

Figure 1
Flowchart of the methodology

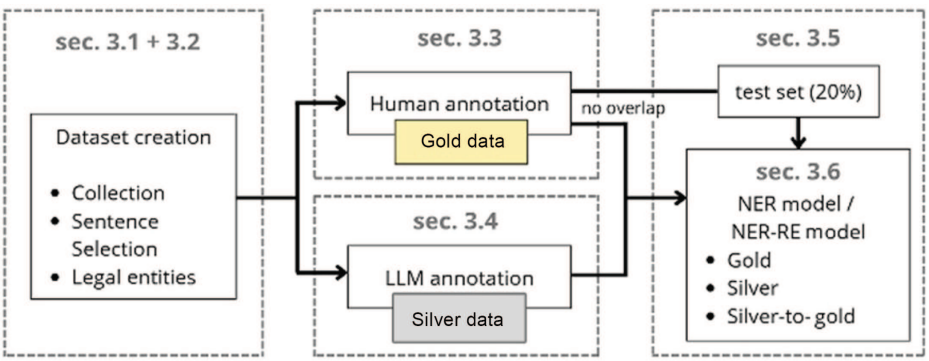

Table 1
Five constitutive entities of administrative decisions

Entity	Explanation	Example (English)
Recipient	The natural or legal person to whom the decision is addressed	ABC Construction B.V
Authority	The public authority competent to take the decision	Minister of Economic Affairs
Legal object	The legal right or obligation affected by the decision (noun)	Permit, sanction, penalty
Legal action	The act with regard to a certain legal object (verb)	Grants, imposes, withdraws
Legal basis	The statutory provision authorizing the authority to take a decision	Article 2a, first paragraph, under a, of the Gas Act

as permits and sanctions) and variety in public authorities within the Netherlands. In total, 20,798 administrative decisions publicly accessible and available in PDF format were scraped from the different websites of seven public authorities and were saved in txt format using pdfplumber where possible. Table 2 shows the dataset statistics of the administrative decisions obtained from these seven different authorities.

Legal documents such as administrative decisions can be quite long [14], but the five legal entities identified in Section 3.1 are usually present together in only a few individual sentences in the text (usually referred to as the "operative part" or *dictum* of the decision). However, unlike judicial decisions, administrative decisions do not follow a predefined or harmonized structure, giving public authorities considerable flexibility in how they compose their decisions. Consequently, for IE tasks, it is not always possible to focus on certain predefined segments or zones of the text of an administrative decision containing its *dictum*. Therefore, we adopt a sentence-level approach rather than a segment or zone-based approach [7, 12]. Specifically, individual sentences are selected from the full text of the decision and the legal entities will be extracted from them. To select those sentences containing the operative part of the decision, we identify the sentences in which the legal effect of the decision is expressed using keyword pair matching. To that end, we have defined pairs of legal action (verbs) and legal object (nouns) based on the applicable legislation.[4] Since the consistent use of terminology used for legal action and legal object is prescribed by (general) legislation, such as the *Algemene wet bestuursrecht*, and additional government-wide instructions (e.g., Drafting instructions for legislation (*Aanwijzingen voor de Regelgeving*), these pairs of legal action and legal object can be assumed to consistently and almost exhaustively express the legal effect of a decision (e.g., the pair of *grant* and *permit*).

The texts of the decisions were split into sentences using SpaCy[5], after which regular expressions were applied to identify whether one of these pairs of legal action and legal object was present in a sentence. If such a pair was identified, this sentence was stored for subsequent processing as it might contain one or more of the five legal entities defined above. As an example, a legal effect (a combination of legal action and legal object) always applies to a specific person (recipient), and that legal action always supposes a legal actor (authority), which is often stated in the same sentence. A total of 14,793 documents were found that contained this sentence pattern at least once, which resulted in a total of 38,766 candidate sentences. See Table 2. Documents in which no such sentence pattern was found were not machine readable or were not administrative decisions (noise).

3.3. Human annotation

After selecting these sentences, two law students (one third-year bachelor and one master) annotated the predefined five legal entities in these sentences (recipient, authority, legal basis, legal object, and legal action) using Lawnotation. The annotations were based on an annotation protocol[6] drafted for this study and tested beforehand. One third of the sentences were annotated by both annotators to measure inter-annotator agreement, which resulted in an overall Cohen's kappa score of 0.47 (with a raw agreement of 0.78 (recipient), 0.58 (authority), 0.51 (legal object), 0.46 (legal action), 0.39 (legal basis) for the individual features), indicating moderate agreement. Among the entity classes, disagreement occurred most frequently for legal basis (303

[4]See GitHub for the action–object pairs and their derivation from the *Algemene wet bestuursrecht*.

[5]SpaCy's pretrained model "nl_core_news_lg"

[6]GitHub link to annotation protocol

Table 2
Overview of documents and extracted sentence statistics per authority and document type

Public authority				Documents			Extracted sentences			Extracted sentence characteristics (in length of tokens)			
Name	Type	Type of decision		Total number	With extracted sentences	%	Total number	Mean per doc.	Mean per extracted doc.	Mean	Med.	Min	Max
ACM	Supervisory	Sanction	License	4725	1764	37.3	10,767	1.19	6.01	31.70	27.0	7	192
ANVS	Supervisory	–	License	10,948	9975	91.1	16,703	1.53	1.67	29.53	29.0	4	80
DNB	Supervisory	Sanction	–	729	317	43.5	2589	2.85	8.17	31.68	30.0	8	75
Rotterdam (ROT)	Municipality	–	License	3791	2,261	59.6	6702	1.65	2.96	35.17	38.0	18	44
KSA	Supervisory	Sanction	–	390	239	61.3	1524	3.90	6.38	29.85	28.0	5	76
Drenthe (DRE)	Province	–	Subsidy	74	72	97.3	203	2.74	2.82	16.49	18.0	6	35
Rijksoverheid (RIJ)	Central Government (Ministries)	–	Subsidy	141	138	97.9	278	1.97	2.01	23.39	21.0	7	65
Total				20,798	14,793	71.1	38,766	1.82	4.00	28.14	25.0	4	192

token-level disagreements), followed by legal object (140), authority (113), legal action (76), and recipient (75). This suggests that disagreements were often caused by boundary mismatches (such as "*Schiphol Nederland B.V.*" versus "*Schiphol Nederland*") or by cases where one annotator marked a span as an entity while the other left it unannotated. Qualitative analysis shows that the moderate agreement can largely be attributed to ambiguities in determining whether a sentence expressed the operative part of a decision or merely referred to, for example, a legal object in another part of the text of that decision. This indicates limitations in the sentence selection techniques used, which might nonetheless be inevitable given the lack of a harmonized structure of administrative decisions. For example, references to legal effects in precedents, legal reasoning, or hypothetical scenarios may resemble operative legal effects linguistically, even though they do not express the actual legal effect of the decision.

To address the observed level of inter-annotator agreement and to ensure the legal correctness and internal consistency of the annotated data, all annotations were subsequently reviewed and validated by a legal expert (law professor). The expert served as an adjudicator in cases of disagreement, resolving conflicts and correcting annotations where necessary to ensure consistency and legal validity, thus acting as an (expert) reviewer [31]. The adjudication process did not systematically favor one annotator: in disagreement cases, the final ground truth followed the first annotator in 334 token-level cases and the second annotator in 364 cases, while only 6 cases matched neither annotator. For sentences that were annotated by a single annotator, 809 out of 20,303 tokens were changed during expert review, corresponding to 3.98%. A further entity-level analysis shows that most entity spans remained unchanged: 2093 entities exactly matched the final ground truth, while 424 entities were removed and 334 new entities were introduced. Boundary corrections were comparatively rare, with 58 small boundary corrections (≤ 3 tokens) and only 6 larger boundary corrections (> 3 tokens). Overall, these results show that the final dataset was not a mechanical merge of the raw annotations, but the result of expert adjudication aimed at improving legal validity and annotation consistency. Table 3 shows the results thereof, demonstrating the number of entities found per authority, and the average length of each type of entity in tokens. The human-annotated dataset is available on GitHub.

3.4. Pseudo-labeling using GPT

GPT-5[7] was used for this study to expand the training data for the NER model. Based on recent advances in LLM-based NER, a schema-driven few-shot prompt[8] was designed. The prompt was based on recent work for prompt engineering, as explained in Section 2.2. Earlier work has conceptualized NER tasks for LLMs as a prompting task, for example, by applying schema-driven instructions to guide the LLM to generate entities in a structured manner. Ideas from previous studies were implemented in the prompt by treating the NER task at stake as a structured generation task aligned with domain-specific definitions, as seen in Table 4. Following prior work in legal IE that validates automated pipelines using statistically representative randomly selected subsets rather than exhaustively annotating entire corpora [7], we use the human-annotated dataset as a benchmark (n = 1101) to evaluate LLM-generated annotations before scaling to a larger dataset. The LLM-generated annotations were evaluated using precision (P), recall (R), and F1 by comparing them with the human-annotated dataset. Afterward, a total of n = 7228 additional sentences were annotated using GPT-5, resulting in the Silver dataset. The LLM-annotated dataset is available on GitHub.

3.5. Experimental setup

This study trains two different NER models on three different subsets of data, which results in six different experimental settings. A stratified random 20% sample of the human-annotated data was used as a test set to evaluate each model, with stratification applied to preserve the representation of documents across

[7] gpt-5-2025-08-07

[8] GitHub link to full prompt

Table 3
Entity counts (c) and mean token length (m) per authority and entity type for the complete and test dataset

Authority	Set	No	Recipient		Authority		Legal object		Legal action		Legal basis	
			c	m	c	m	c	m	c	m	c	m
ACM	All	293	76	3.00	105	2.96	99	1.25	101	1.00	58	8.24
	Test	53	13	3.08	24	3.21	19	1.47	19	1.00	13	7.92
ANVS	All	149	98	4.39	3	1.00	112	1.00	112	1.01	4	14.0
	Test	25	17	4.88	0	–	21	1.00	21	1.00	0	–
DNB	All	149	31	1.55	28	1.18	32	1.88	33	1.00	14	6.79
	Test	37	5	1.20	11	1.00	6	1.83	6	1.00	4	7.25
ROT	All	60	36	1.00	36	1.00	41	1.00	41	1.00	30	8.53
	Test	6	5	1.00	5	1.00	5	1.00	5	1.00	5	9.00
KSA	All	150	28	2.25	46	3.96	37	2.16	37	1.00	25	6.52
	Test	27	4	2.75	7	3.71	8	2.00	8	1.00	6	7.50
DRE	All	150	71	1.00	80	1.05	82	1.01	82	1.00	3	5.67
	Test	15	7	1.00	10	1.00	10	1.00	10	1.00	1	12.0
RIJ	All	150	50	1.06	64	1.09	88	1.04	88	1.00	22	11.3
	Test	21	8	1.12	10	1.00	13	1.15	13	1.00	3	12.0
Total	All	1101	390	2.38	362	1.99	487	1.21	494	1.00	156	8.42
	Test	184	59	2.73	67	2.07	82	1.29	82	1.00	32	8.44

Table 4
Structured overview of the prompt

Prompt structure	Implementation	Idea
Introduction	Introduce the role of the LLM and frame the task as a QA task	Generative task for entity extraction [4]
Entity descriptions	Define entities individually in a schema-driven way and include an explanation, with possible relations to other legal entities with entity examples	Schema-driven prompting [5] and highlighting entity relations [20]
Normalization	Explain rules for normalization (deduplication, abstention, span/boundary), and show expected output with an example JSON structure	Normalization and fixed output [4]
Examples	Provide three examples	Few-shot prompting

administrative authorities. The remaining 80% served as training data for the NER models trained on human annotations. Because sentences expressing the operative clause of the decision frequently use standardized legal terminology, similar sentences may appear multiple times. To prevent data leakage, sentences that occurred in both the training set and the test set were removed from the test set, resulting in a test set containing 184 sentences.

After training, model performance was evaluated using token-level annotations from this test set (n = 184). Model output was transformed to token-level predictions to align with the test set where necessary. Precision (P), recall (R), and F1 were computed at the token level. Scores were calculated per class and subsequently micro-averaged to obtain an overall performance score across all classes. The evaluation followed a strict matching criterion, where partial overlaps between LLM-annotated and human-annotated entities at the token level were considered incorrect.

3.6. NER models

For this study, two different NER architectures were trained under three different training settings, each resulting in three models: Gold, Silver, and Silver-to-Gold. All models were optimized using AdamW with a learning rate of 5e-5, batch size 8, weight decay 0.01, and linear learning rate decay. Training was conducted for a maximum of 10 epochs with early stopping (patience = 3) based on validation F1. See the GitHub repository for the full training settings.

First, a token-level BERT NER model based on the Dutch BERT variant (BERTje) [32] was trained on three data subsets. BERTje was chosen because its Dutch-specific pretraining enables state-of-the-art performance on Dutch NLP tasks, making it a reliable choice for token-level NER in Dutch legal texts [32]. The model assigns an entity label to each token using its context-aware representation. The dataset was tokenized with the BERTje tokenizer [32]. Three training settings were applied:

1) Gold model: Trained on the manually annotated dataset (n = 883; Section 3.3), of which 20% was used as a validation set.
2) Silver model: Trained on the LLM-generated annotations (n = 7228; Section 3.4), with 20% used for validation.
3) Silver-to-Gold model: Initialized from the Silver model and subsequently fine-tuned with human-annotated data (20% of human-annotated data used for validation).

Second, a span-based joint NER–RE model was trained, based on BERTje [32]. This is a more sophisticated NER architecture that combines NER with RE. Inspired by spERT [21], this span-based model represents candidate spans with contextual embeddings, labels them as entities, and simultaneously predicts relationships between spans.

Since relationships between entities were not annotated by humans, these relationships were constructed automatically by pairing, where possible, entities present in a sentence according to a pre-set schema. Four relationship types were defined for this project based on the legal definition of administrative decisions (as described in Section 3.1): (1) *recipient* receives *legal object*, (2) *legal action* concerns *legal object*, (3) *legal basis* authorizes *authority*, and (4) *authority* performs *legal action.* See Figure 2 for an example of this combined NER and RE task. This automated construction of relationships could introduce some noise. However, based on a qualitative analysis of 200 randomly selected sentences with at least one entity present, only five samples were found to contain incorrect or missing relationship labels (2.5%). Four of these were caused by misclassified or missing entities, and one was caused by the automated pairing procedure itself: the sentence contained two different legal actions and two different legal objects, resulting in too many relationships, since each legal action should have been paired with only one legal object. The joint NER–RE model was trained with a combined loss

[9]Translation: "*The Dutch Authority for Consumers and Markets decide, pursuant to Article 2a(1)(a) of the Gas Act, to grant an exemption from the obligation to appoint a grid operator to Schiphol Nederland B.V., for the [. . .]*"

Figure 2
Example of NER (highlighted) and RE (red arrows; automatically constructed)[9]



function, combining the entity classification loss with the relationship classification loss. This span-based model was trained under three different training settings, identically to the token-level NER model, resulting in a Gold, Silver, and Silver-to-Gold NER–RE model.

4. Results

4.1. LLM pseudo-annotation performance

Table 5 shows the quality of the pseudo-annotations generated by GPT-5, by comparing them with human annotations (n = 1101). Relatively high F1-scores can be observed for the entities of "authority" (F1 = 0.69) and "legal object" (F1 = 0.59). These high scores can be explained by the uniform way in which these entities are defined in legislation and by the structure of these entities being relatively consistent across administrative decisions, making it easy for LLMs to capture these. These scores are lower, however, for authorities that impose sanctions (ACM, DNB, KSA). This can be explained by the more complex nature of sanctioning decisions, which often contain lengthy reasoning preceding the relevant operative part of the decision (see also Table 2). Additionally, in our dataset, the average token length of the authority imposing a sanction is often longer than that of authorities imposing licensing or subsidizing decisions, as Table 3 shows. As an example of these longer token lengths, enforcement decisions (such as a fine) often display the legal object with an adjectival premodifier, an adjective that appears before a noun to tailor it (such as *administrative fine*). The LLM failed to capture these premodifiers, suggesting systematic boundary annotation errors.

The model shows moderate performance for the entity of "legal action" (F1 = 0.51), which, despite being formally defined in legislation, exhibits considerable variation in form and often depends on the context. For example, legal actions may be confused with auxiliary or procedural verbs (e.g., *decide to grant*) and can be expressed through different grammatical forms (e.g., *grant, granted, granting*) as well as a wide range of action-specific verbs (e.g., *grant, approve, amend, revoke*). In contrast, the entity "legal object" (F1 = 0.59), although also formally defined in legislation, is generally expressed through a more limited and stable vocabulary (e.g., *permit, fine, exemption*). These findings suggest that the LLM has greater difficulty identifying the boundaries and meaning of legal actions than those of legal objects.

In contrast, the entities of "recipient" and "legal basis" yield lower scores (recipient: F1 = 0.42; legal basis: F1 = 0.38). These entities are typically longer, more heterogeneous in form (e.g., the recipient can be any person), and more complex from a legal-linguistic perspective, which makes it difficult for LLMs to identify them consistently. Performance for "recipient" is higher among authorities that frequently issue their decisions in a pre-structured letter format (e.g., province of Drenthe (DRE), Central Government (Rijksoverheid, RIJ), municipality of Rotterdam (ROT)) where the recipient is often expressed through short and direct referential tokens such as "you." This drafting style in letter format differs from more formally drafted decisions ready for publication. Interestingly, precision scores for legal basis are relatively high (p = 0.78) compared to its recall (R = 0.25), suggesting that while LLMs can correctly capture clearly identifiable patterns, they fail to identify less explicit instances or relevant legal provisions that they have not seen before.

The results reveal a key limitation of using LLMs for data annotation and NER tasks. Although LLMs reliably identify enti-

Table 5
LLM pseudo-label quality per authority in precision (P), recall (R), F1, and total TP + FP + FN count (#) per entity

	Recipient			Authority				Legal object				Legal action				Legal basis			
	P	R	F1	#	P	R	F1	#	P	R	F1	#	P	R	F1	#	P	R	F1	#
ACM	0.05	0.05	0.05	150	0.51	0.74	0.60	178	0.36	0.90	0.51	261	0.38	0.75	0.51	223	0.00	0.00	0.00	59
ANVS	0.07	0.09	0.08	213	0.60	1.0	0.75	5	0.80	0.98	0.88	140	0.71	0.60	0.65	140	0.00	0.00	0.00	6
DNB	0.17	0.07	0.09	41	0.29	0.96	0.45	93	0.01	0.03	0.01	138	0.27	0.82	0.41	105	0.55	0.43	0.48	19
ROT	0.69	1.0	0.82	52	0.66	1.0	0.79	55	0.68	1.0	0.81	60	1.0	0.56	0.72	41	1.0	0.67	0.80	30
KSA	0.13	0.11	0.12	48	0.36	0.94	0.52	122	0.23	0.81	0.36	139	0.26	0.76	0.39	116	0.80	0.48	0.60	28
DRE	0.96	1.0	0.98	74	0.94	0.99	0.96	85	0.46	0.96	0.62	176	0.41	0.94	0.57	194	0.00	0.00	0.00	3
RIJ	0.77	0.94	0.85	64	0.80	0.95	0.87	79	0.67	0.98	0.80	124	0.33	0.30	0.31	141	1.0	0.05	0.09	22
Combined	0.41	0.44	0.42	642	0.56	0.90	0.69	617	0.44	0.89	0.59	1038	0.41	0.66	0.51	960	0.78	0.25	0.38	167

ties expressed through clear and explicit patterns, they perform less consistently when entities are novel relative to the prompt, require deeper contextual inference, or rely on common-sense knowledge that is not explicitly encoded in the text.

4.2. NER performance

Table 6 shows the results of the token-based NER and span-based joint NER–RE models, with precision, recall, and F1-scores for each of the three subsets of annotated data: Gold (human annotations only), Silver (LLM annotations only), and Silver-to-Gold (Silver model fine-tuned on human annotations).[10] Additional information about the test set is reported in Table 3.

These results show that both Gold models achieve high micro average F1-scores (NER: 0.87; NER–RE: 0.95), providing a strong baseline and demonstrating good performance on human annotations. In contrast, the Silver models, trained exclusively on LLM-generated pseudo-labels, perform worse (NER: 0.50; NER–RE: 0.84), largely due to inconsistencies in LLM-generated annotations for complex or implicit entities such as "recipient" and "legal basis." Importantly, when the Silver models are subsequently fine-tuned with human annotations (Silver-to-Gold), their performance surpasses the baselines of the Gold models, improving the micro average F1-scores by 0.02 and 0.04, respectively, for the NER and NER–RE models.

Using McNemar's exact test, with the prediction of the model (binary: correct or incorrect based on human baseline) as the dependent variable and the model type (Gold or Silver-to-Gold) as the independent variable, no statistically significant difference was found for the NER models. For the NER–RE models, the individual entity-level comparisons were also not found to be statistically significant. However, a pooled analysis across all entities together showed a statistically significant difference between the Gold and Silver-to-Gold model ($p = 0.019$) for overall model performance for NER–RE, although this result should be interpreted cautiously because the pooled entity-level decisions are not fully independent. A further qualitative analysis shows that improvements in the Silver-to-Gold models compared to Gold models primarily stem from more accurate span boundaries and better coverage of alternative expressions for entities (e.g., abbreviations or referential phrases). This finding indicates that combining imperfect but large-scale LLM-generated annotations with smaller sets of more reliable human annotations can effectively expand coverage and performance in low-resource legal NER tasks.

Across different experimental setups, the models show a clear performance divide: entities expressed through standardized and explicit legal phrasing, such as public authority, legal object, and legal action, are identified most reliably, whereas more context-dependent and linguistically variable entities, particularly recipient and legal basis, remain more challenging. Similarly, sentences that contain the legal clause of an entitlement letter decision, which are thus more harmonized, achieve better results compared to decisions where these entities are not part of such harmonized components. An example thereof is the entity of recipient, which scores lower for documents that are not written in a letter format and thus lack structured sections for recipient (see footnote 7). These observations confirm that model performance depends strongly on both the linguistic structure of the texts and the degree of standardization in the way an administrative decision is drafted.

Interestingly, the Silver NER–RE model performs better than expected given the quality of its LLM-generated training labels (Table 5). This indicates that the model may be able to generalize beyond the noise present in the pseudo-labels. The joint NER–RE architecture likely encourages structurally coherent predictions by linking entities through predefined relationships. In addition, the large volume of pseudo-labeled data increases the model's exposure to diverse linguistic contexts, effectively broadening its training coverage. These aspects suggest that the model benefits from denoising effects, which may explain the high precision and F1-scores observed despite the noisy supervision.

[10] GitHub link to more detailed results per authority

Table 6
Per-label performance (P, R, F1) of three approaches for both BERT-based NER and spERT-based joint NER-RE models

	NER								
	Gold			Silver			Silver-to-Gold		
	P	R	F1	P	R	F1	P	R	F1
Recipient	0.77	0.86	0.82	0.22	0.34	0.26	0.83	0.92	0.87
Auth.	0.80	0.90	0.85	0.58	0.84	0.69	0.83	0.90	0.86
Legal obj.	0.97	0.85	0.91	0.43	0.85	0.57	0.95	0.88	0.91
Legal act.	0.96	0.89	0.92	0.41	0.61	0.49	0.96	0.89	0.92
Legal basis	0.76	0.91	0.83	0.53	0.25	0.34	0.83	0.94	0.88
Total	0.87	0.88	0.87	0.41	0.63	0.50	0.89	0.90	0.89
	Joint-NER-RE								
Recipient	1.0	0.97	0.98	0.95	0.32	0.48	1.0	0.98	0.99
Auth.	0.98	0.93	0.95	0.98	0.91	0.95	1.0	0.97	0.99
Legal obj.	0.96	0.94	0.95	0.99	0.85	0.91	0.99	0.98	0.98
Legal act.	0.96	0.94	0.95	0.96	0.92	0.94	1.0	1.0	1.0
Legal basis	0.94	0.91	0.92	1.0	0.41	0.58	1.0	0.94	0.97
Total	0.97	0.94	0.95	0.98	0.74	0.84	1.0	0.98	0.99

5. Discussion

5.1. Pseudo-annotations by LLMs

This study shows that LLMs are relatively consistent in identifying entities with clear patterns and little linguistic variety but act less consistently with regard to entities that have a lot of linguistic variety and may require more contextual understanding or common sense in a less consistent way. To illustrate this, legal entities such as the legal object of decision or the public authority issuing the decision are defined explicitly in legislation and require specific domain knowledge. The (legally) formalized nature of administrative decisions means that there is little variation in terminology and syntactic structure for features that are explicitly defined in legislation. In contrast, features such as the recipient, which are not necessarily defined in legislation and do not require specific domain knowledge (as any citizen can be the recipient of an administrative decision), score relatively low. Since the latter features rely more on common-sense reasoning, LLMs struggle to identify these entities correctly, as these features require understanding of everyday knowledge or contextual inference beyond what is explicitly stated in legislation or elsewhere. However, when these entities are written concisely or uniformly, the scores for extracting these entities correctly improve.

These results are in line with other studies insofar as LLMs demonstrate limitations in abstract common-sense reasoning, often failing to grasp relationships or contexts that humans intuitively understand [27]. Additionally, the findings are also in line with other studies [11], indicating that general-purpose LLMs show promise for NER in few-shot settings but continue to face challenges related to boundary errors, false positives, and incomplete recall. More broadly, the results highlight the limitations of LLMs in handling contextual variation across certain entity types, suggesting that tailoring models to legal language may be necessary to achieve more reliable performance. Our findings align with the literature, indicating that prompt design has a strong impact on performance and that a more optimal formulation could further improve results [27]. All in all, LLMs can provide useful annotations and improve coverage in a cost-efficient way, specifically for more standardized legal entities. However, for more ambiguous entities, which require deeper common-sense understanding, LLM annotations are less reliable, highlighting the importance of expert prompt optimization or the inclusion of expert annotated data.

5.2. Impact of pseudo-annotations on supervised NER

Gold models, trained solely on human annotations, achieve strong results and form a reliable baseline for NER. This is consistent with prior work showing the superiority of supervised methods over LLM-only approaches [1, 5]. As expected, the Silver models perform considerably worse due to incorrect and noisy annotations in the LLM-generated labels. However, when the Silver models are fine-tuned on human annotations (Silver-to-Gold), their performance slightly exceeds that of the Gold models, indicating that weakly supervised LLM-generated annotations can broaden training coverage and better represent the linguistic and legal variation in the data for low-resource LER tasks, albeit with modest gains. Even though the observed improvements are small in absolute terms, they are consistent across both models (NER and NER–RE), indicating that the gains are not confined to a single model configuration or entity type. This pattern suggests that the improvements reflect enhanced generalization rather than overfitting or random variation.

The observed improvement of the Silver-to-Gold models over the Gold-only baselines could be explained by the complementary roles of pseudo-annotations and human annotations during training. While LLM-generated pseudo-annotations are noisier, they increase the diversity and coverage of entities observed during training and thereby expose the model to a broader range of lexical, syntactic, and contextual variations. While qualitative analysis shows that errors made only by Silver-to-Gold models were likely introduced by noise in the LLM-annotated data, the opposite was also true: when cases are classified correctly by the Silver-to-Gold models only, the pseudo-annotations likely have contributed to a broader representation of the dataset and contain more diverse examples from which the model can learn. This can, for example, be observed for the specific entity of a legal basis, as the Silver-to-Gold models identify legal provisions better, as these are better represented in additional examples. As a result, the Silver-to-Gold models achieve slightly improved performance compared to models trained solely on limited human-annotated data.

Within the joint NER–RE setting, precision is generally higher than recall. In contrast, within the token-based NER setting, recall generally exceeds precision, even though the absolute scores are lower compared to the joint setting. This behavior can also be observed in previous work [21] and might be explained by the structural constraints introduced by RE: entity predictions are no longer independent but must participate in valid relationships defined by the domain-specific schema. As a result, weakly supported entities are suppressed, reducing false positives and increasing precision. At the same time, this constraint leads to the omission of some correct but atypical or implicit entities that fail to form explicit relationships, resulting in lower recall. In contrast, the NER model relies solely on local token-level cues, leading to higher recall but also more false positives. To support this finding, qualitative analysis shows that in some cases NER models misclassified entities, while these mistakes were not found for NER–RE models. This indicates that the addition of the RE task could help reduce misclassification in such cases by filtering out or correcting these false positives.

In contrast to earlier studies, which found that weakly supervised models generally underperform fully supervised ones [9, 28], our findings show that fine-tuning a Silver model on human-annotated data can slightly surpass human-only baselines. Unlike prior work that simultaneously combined human and pseudo-labels for training NER models [9], our Silver-to-Gold strategy separates learning into two phases. This separation allows the model to first learn broad representational patterns from large-scale pseudo-annotations and to constrain these representations only afterward using reliable human supervision, reducing the risk that noisy labels dominate the learning signal. Nevertheless, inconsistencies in the pseudo-labels can still introduce noise, emphasizing the need for careful integration of LLM-generated data.

5.3. Limitations

Although the results highlight the potential of NER and NER–RE models for extracting legal entities, they also point to several limitations that call for further investigation. First, the methodology can be further refined and strengthened in future research. For example, some sentences were included in

the dataset for NER despite the absence of a legal effect, revealing shortcomings in the sentence selection process. To address this shortcoming in the stage prior to the actual extraction of entities from preselected parts of the legal text, future research could explore structure-aware or visually rich document NER approaches, such as DLA or document segmentation techniques, to ensure a more accurate sentence selection [1, 24].

Second, despite creating a diversified but prevalence-aware dataset, certain rare pairs of legal action and legal object were underrepresented in the dataset (such as *revoking* a *subsidy*), which may lead to scalability issues.

Third, entities expressed as pronouns in letter-style decisions (e.g., "I," "you") may create coreference issues for downstream tasks and may have simplified the NER task.

Fourth, the experimental setup was restricted to a single transformer-based token classification model, one LLM configuration, and one fixed split between training set and test set. Alternative NER architectures such as span-based or transformer-based RobBERT (instead of BERTje), alternative LLM architectures such as LLaMA or Mistral, and multiple-seed settings, for example, through cross-validation, may yield different performances. Moreover, although the experiments were conducted at the sentence level, template leakage cannot be fully excluded. Administrative decisions often contain recurring formulations, and certain formulations may be overrepresented in the dataset, which could allow models to benefit from authority-specific or template-specific patterns rather than generalizable entity recognition. Future work should therefore include document-level evaluation, in which all sentences from a document are assigned exclusively to either the training or test set, or authority-held-out evaluation, in which decisions from one or more administrative authorities are excluded from training and used only for testing. Such evaluation settings would better assess the robustness of the approach and reduce the risk of template leakage.

Fifth, pre-training BERTje on the larger unlabeled administrative decisions could provide a stronger baseline and potentially amplify the hybrid gains obtained by the strategy adopted in this study.

Finally, future research should evaluate the influence of including an RE task on NER performance. Although we observe that the more sophisticated joint span-based NER–RE model outperforms the token-based NER model on the same data, future research should investigate whether these gains indeed stem specifically from the inclusion of RE and thus the inclusion of extra domain knowledge by comparing a span-based NER task with a similar span-based NER–RE task.

6. Conclusion

Low-resource legal domains pose a challenge for NER tasks, as high-quality annotated data are costly and time-consuming to create. To address this problem, this study examined the use of LLM-generated pseudo-annotations to augment supervised NER models in a low-resource legal domain (Dutch administrative decisions) on a sentence level. Specifically, (1) few-shot GPT-5 was used to generate large-scale pseudo-annotations for five predefined legal entities, after which their performance was evaluated, and (2) these LLM-generated annotations were used to expand the training data of two different NER architectures (token-based NER and span-based NER–RE), after which their impact was evaluated by comparing multiple training configurations: Gold (human data only), Silver (LLM data only), and Silver-to-Gold (LLM-trained model fine-tuned with human data).

Overall, the findings highlight the promise and limitations of LLM-generated pseudo-annotations: (1) GPT-5 can produce large-scale annotations, but with uneven quality: legal entities that are defined explicitly in legislation and consistently expressed (e.g., *public authority, legal object*) are captured in a reliable way, while longer or context-dependent entities (e.g., *recipient, legal basis*) remain challenging for LLMs to extract. Furthermore, (2) although NER models trained solely on LLM-generated annotations underperform compared to NER models trained on human annotations, fine-tuning these LLM-based models with a smaller set of human annotations yields performance that slightly surpasses NER models trained only on human-labeled data. This indicates that weakly supervised LLM-generated annotations, although adding some noise, provide broader training coverage and can thereby improve performance in low-resource legal NER tasks. These findings demonstrate promising initial evidence of the value of hybrid annotation strategies that combine scalable LLM-generated labels with targeted human supervision, which can offer a practical approach for developing more scalable and accurate legal IE systems.

Ethics and Data Governance

This study makes use of Dutch administrative decisions to which no copyright restrictions apply and which have been made publicly available in accordance with the applicable legislation on access to government information. Accordingly, some decisions had already been anonymized partially before publication by the issuing public authority. No additional anonymization or masking steps were applied by the authors. Although it cannot be excluded that a (potentially) identifiable natural person is captured as the recipient of the decision in the dataset, this is the immediate consequence of the public release of the decision in accordance with the applicable legislation. When creating LLM-generated pseudo-annotations, a ChatGPT Edu License obtained by Tilburg University was used in accordance with the guidelines issued by Tilburg University to ensure only public (non-sensitive) data were used. The public release of the dataset obtained by this study is limited to the information necessary for scientific transparency and reproducibility, while taking into account that the underlying decisions were already publicly accessible.

Acknowledgment

We are grateful to the two law student annotators, Laura Matthijs and Sanne Rademakers, for their contributions to the annotation process. We also acknowledge the support of Tilburg University.

Funding Support

This publication is part of the project "Administrative Decision-making in Times of Open Government: From Rule to Case Transparency" with file number VI.Vidi.221R.026 of the research program Vidi NWO Talent, which is (partly) financed by the Dutch Research Council (NWO), and part of the Icon project "Case-Inclusive Transparency for a Digital and Open Government (CITaDOG)" by Tilburg University's Digital Sciences for Society (DSFS) program.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available on GitHub at https://github.com/Harry-Nan/legal-ner/tree/main/data.

Author Contribution Statement

Harry Nan: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Samaneh Khoshrou:** Conceptualization, Validation, Resources, Writing – review & editing, Supervision, Project administration. **Johan Wolswinkel:** Conceptualization, Validation, Resources, Data curation, Writing – review & editing, Supervision, Project administration, Funding acquisition.

References

[1] Seow, W. L., Chaturvedi, I., Hogarth, A., Mao, R., & Cambria, E. (2025). A review of named entity recognition: From learning methods to modelling paradigms and tasks. *Artificial Intelligence Review*, *58*(10), 315. https://doi.org/10.1007/s10462-025-11321-8

[2] Premasiri, D., Ranasinghe, T., Mitkov, R., El-Haj, M., & Frommholz, I. (2025). Survey on legal information extraction: Current status and open challenges. *Knowledge and Information Systems*, *67*(12), 11287–11358. https://doi.org/10.1007/s10115-025-02600-5

[3] Li, J., Wang, J., Zhang, Z., & Zhao, H. (2024). Self-prompting large language models for zero-shot open-domain QA. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 296–310. https://doi.org/10.18653/v1/2024.naacl-long.17

[4] Xie, T., Li, Q., Zhang, J., Zhang, Y., Liu, Z., & Wang, H. (2023). Empirical study of zero-shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7935–7956. https://doi.org/10.18653/v1/2023.emnlp-main.493

[5] Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., ..., & Guo, C. (2025). GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 4257–4275. https://doi.org/10.18653/v1/2025.findings-naacl.239

[6] Costa, M. Z., Robson, D. F., Vieira, T. B. P., Bourguet, J.-R., Guizzardi, G., & Almeida, J. P. A. (2024). Automated semantic annotation pipeline for Brazilian judicial decisions. In *37th Annual Conference on Legal Knowledge and Information Systems*, 226–238. https://doi.org/10.3233/FAIA241248

[7] Darji, H., Mitrović, J., & Granitzer, M. (2023). German BERT model for legal named entity recognition. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*, *3*, 723–728. https://doi.org/10.5220/0011749400003393

[8] Çetindağ, C., Yazıcıoğlu, B., & Koç, A. (2023). Named-entity recognition in Turkish legal texts. *Natural Language Engineering*, *29*(3), 615–642. https://doi.org/10.1017/S1351324922000304

[9] Oliveira, V., Nogueira, G., Faleiros, T., & Marcacini, R. (2025). Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents. *Artificial Intelligence and Law*, *33*(2), 361–381. https://doi.org/10.1007/s10506-023-09388-1

[10] Bogdanov, S., Constantin, A., Bernard, T., Crabbé, B., & Bernard, E. P. (2024). NuNER: Entity recognition encoder pre-training via LLM-annotated data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 11829–11841. https://doi.org/10.18653/v1/2024.emnlp-main.660

[11] Breton, J., Billami, M. M., Chevalier, M., Nguyen, H. T., Satoh, K., Trojahn, C., & Zin, M. M. (2025). Leveraging LLMs for legal terms extraction with limited annotated data. *Artificial Intelligence and Law*. Advance online publication. https://doi.org/10.1007/s10506-025-09448-8

[12] Chalkidis, I., Androutsopoulos, I., & Michos, A. (2017). Extracting contract elements. In *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law*, 19–28. https://doi.org/10.1145/3086512.3086515

[13] Nan, H., Marx, M., & Wolswinkel, J. (2024). Combining rule-based and machine learning methods for efficient information extraction from enforcement decisions. In *The Thirty-seventh Annual Conference on Legal Knowledge and Information Systems* , 321–326. https://doi.org/10.3233/FAIA241262

[14] Zadgaonkar, A. V., & Agrawal, A. J. (2021). An overview of information extraction techniques for legal document analysis and processing. *International Journal of Electrical and Computer Engineering*, *11*(6), 5450–5457. https://doi.org/10.11591/ijece.v11i6.pp5450-5457

[15] Leitner, E., Rehm, G., & Moreno-Schneider, J. (2020). A dataset of German legal documents for named entity recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4478–4485.

[16] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261

[17] Kwak, A., Jeong, C., Forte, G., Bambauer, D., Morrison, C., & Surdeanu, M. (2023). Information extraction from legal wills: How well does GPT-4 do? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 4336–4353. https://doi.org/10.18653/v1/2023.findings-emnlp.287

[18] Abdullah, M. H. A., Aziz, N., Abdulkadir, S. J., Alhussian, H. S. A., & Talpur, N. (2023). Systematic literature review of information extraction from textual data: Recent methods, applications, trends, and challenges. *IEEE Access*, *11*, 10535–10562. https://doi.org/10.1109/ACCESS.2023.3240898

[19] Hartung, D., Katz, D. M., Bommarito, M. J., Gerlach, L., Jana, A., & Soh, J. (2023). *Natural language processing in the legal domain.* arXiv. https://doi.org/10.48550/arXiv.2302.12039

[20] Xiao, L., Xu, Y., & Zhao, J. (2024). *LLM-DER: A named entity recognition method based on large language models for*

Chinese coal chemical domain. arXiv. https://doi.org/10.48550/arXiv.2409.10077

[21] Eberts, M., & Ulges, A. (2020). Span-based joint entity and relation extraction with transformer pre-training. In *24th European Conference on Artificial Intelligence*, 2006–2013. https://doi.org/10.3233/FAIA200321

[22] Wadden, D., Wennberg, U., Luan, Y., & Hajishirzi, H. (2019). Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5784–5789. https://doi.org/10.18653/v1/D19-1585

[23] Sui, D., Zeng, X., Chen, Y., Liu, K., & Zhao, J. (2024). Joint entity and relation extraction with set prediction networks. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(9), 12784–12795. https://doi.org/10.1109/TNNLS.2023.3264735

[24] Yang, H.-W., & Agrawal, A. (2023). Extracting complex named entities in legal documents via weakly supervised object detection. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3349–3353. https://doi.org/10.1145/3539618.3591852

[25] Xie, T., Li, Q., Zhang, Y., Liu, Z., & Wang, H. (2024). Self-improving for zero-shot named entity recognition with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *2*, 583–593. https://doi.org/10.18653/v1/2024.naacl-short.49

[26] Gray, M., Savelka, J., Oliver, W., & Ashley, K. (2023). Can GPT alleviate the burden of annotation? *In The Thirty-sixth Annual Conference on Legal Knowledge and Information Systems* , 157–166. https://doi.org/10.3233/FAIA230961

[27] Gawin, C., Sun, Y., & Kejriwal, M. (2025). Navigating semantic relations: Challenges for language models in abstract common-sense reasoning. In *Companion Proceedings of the ACM on Web Conference 2025*, 971–975. https://doi.org/10.1145/3701716.3715472

[28] Møller, A. G., Pera, A., Dalsgaard, J., & Aiello, L. (2024). The parrot dilemma: Human-labeled vs LLM-augmented data in classification tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, *2,* 179–192. https://doi.org/10.18653/v1/2024.eacl-short.17

[29] Ashok, D., & May, J. (2025). A little human data goes a long way. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, *2,* 381–413. https://doi.org/10.18653/v1/2025.acl-short.30

[30] Hage, J. (2017). Basic concepts of law. In J. Hage, A. Waltermann, & B. Akkermans (Eds.), *Introduction to law* (pp. 33–52). Springer. https://doi.org/10.1007/978-3-319-57252-9_3

[31] Braun, D. (2024). I beg to differ: How disagreement is handled in the annotation of legal machine learning data sets. *Artificial Intelligence and Law*, *32*(3), 839–862. https://doi.org/10.1007/s10506-023-09369-4

[32] de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). *BERTje: A Dutch BERT model*. arXiv.https://doi.org/10.48550/arXiv.1912.09582 .

How to Cite: Nan, H., Khoshrou, S., & Wolswinkel, J. (2026). Legal NER: Evaluating the Impact of LLM-Generated Annotations on NER Performance for Administrative Decisions. *Journal of Computational Law and Legal Technology*, *1*, A9. https://doi.org/10.47852/bonviewJCLLT62029445