Do Language Models Learn about Legal Entity Types during Pretraining?

Claire Barale and **Michael Rovatsos**
School of Informatics
The University of Edinburgh
{claire.barale,michael.rovatsos}@ed.ac.uk

Nehal Bhuta
School of Law
The University of Edinburgh
nehal.bhuta@ed.ac.uk

Abstract

Language Models (LMs) have proven their ability to acquire diverse linguistic knowledge during the pretraining phase, potentially serving as a valuable source of incidental supervision for downstream tasks. However, there has been limited research conducted on the retrieval of domain-specific knowledge, and specifically legal knowledge. We propose to explore the task of Entity Typing, serving as a proxy for evaluating legal knowledge as an essential aspect of text comprehension, and a foundational task to numerous downstream legal NLP applications. Through systematic evaluation and analysis and two types of prompting (cloze sentences and QA-based templates) and to clarify the nature of these acquired cues, we compare diverse types and lengths of entities both general and domain-specific entities, semantics or syntax signals, and different LM pretraining corpus (generic and legal-oriented) and architectures (encoder BERT-based and decoder-only with Llama2). We show that (1) Llama2 performs well on certain entities and exhibits potential for substantial improvement with optimized prompt templates, (2) law-oriented LMs show inconsistent performance, possibly due to variations in their training corpus, (3) LMs demonstrate the ability to type entities even in the case of multi-token entities, (4) all models struggle with entities belonging to sub-domains of the law (5) Llama2 appears to frequently overlook syntactic cues, a shortcoming less present in BERT-based architectures. The code of the experiments is available at https://github.com/clairebarale/probing_legal_entity_types.

1 Introduction

During the initial phase of pretraining, language models (LMs) are exposed to an extensive corpus of textual data, allowing them to acquire the capacity to represent the probabilistic structure of language. In this process, it has been theorized that they incidentally learn various linguistic signals and patterns, both syntactic and semantic. Work by Petroni et al. (2019) and subsequent studies (Jiang et al., 2020b) make the hypothesis that a side effect of the pretraining stage is that LMs also learn factual knowledge. On the other hand, Gururangan et al. (2020) research demonstrates the significance of both model pretraining and task-specific pretraining; pretraining a model with a specific focus on a particular task or a limited domain corpus yields notable advantages in enhancing model performance and adaptability.

Entity typing and extraction are crucial tasks for a range of use cases including named-entity recognition (NER), relation extraction, summarization, structuring raw data, and most specifically in law, legal search, and past cases retrieval. To gain more insights into entity typing and extraction, entity probing tasks have been designed for bidirectional LSTM conditional random field models (Augenstein et al., 2017), masked language models (Petroni et al., 2019; Jiang et al., 2020b) and autoregressive LMs (Epure and Hennequin, 2022), using GPT-2.

Conversely, one notable bottleneck of the application of NLP within the legal domain is the lack of resources and annotated datasets. Thereby, it is of particular interest to explore the extent to which LMs, during their pretraining phase, acquire a sufficient **understanding of legal entities**, serving as a surrogate for legal knowledge. Ultimately, LMs could be exploited as a source of weak and indirect supervision in downstream tasks such as legal NER or question answering (QA), as they constitute a good proxy to use natural text incidentally thanks to their pretraining stage. Indeed, humans do not exclusively rely on exhaustive supervision but instead make use of occasional feedback and learn from incidental signals originating from various sources. This approach holds potential for

Proceedings of the Natural Legal Language Processing Workshop 2023, pages 25–37

	Pretrained Language Model	Pretraining corpus	# Parameters	# Tokens	Corpus size	# Vocab
Legal	**CaseHOLD** (Zheng et al., 2021)	Harvard Case Law	110M	43B	37 GB	32K
	Pile of law (Henderson et al., 2022)	US, Canadian, ECthR	340M	130B	256 GB	32K
	LexLM (Chalkidis et al., 2023)	US, Canada, EU, UK, India	125	2T + 256B	175 GB	50K
Generic		BookCorpus (Zhu et al., 2015)		-	16GB	-
		CC_news (Nagel, 2016)		-	76GB	-
		OpenWebText (Radford et al., 2019)		-	38GB	-
		Stories (Trinh and Le, 2018)		-	31GB	-
	RoBERTa(Liu et al., 2019)		125M	2T	160GB	50K
	DeBERTa (He et al., 2023)		86M	2T	160GB	128K
	Llama 2 (Touvron et al., 2023)	Data from publicly available sources	7B	2T	-	32K

Table 1: Overview of the models used. The table reports the description of the pretraining corpora, the number of parameters, the total number of tokens, the size of the corpus, and the vocabulary size

increased flexibility in terms of entity types, in contrast to supervised methods, and presents an alternative to existing automated annotation extraction approaches (Tedeschi and Navigli, 2022; Savelka, 2023) which hold limitations in the set of entity types. It presents several advantages: it does not require human annotation, it can be easily combined with other sources of supervision such as legal knowledge bases, and it would support an open set of entities and user queries. It would offer the advantage of seamless and fast application to new datasets while facilitating transfers of knowledge between datasets and even potentially between different domains. In this paper, we study the **intersection of entity knowledge and legal knowledge embedded within LMs**, evaluated on a ***AsyLex***, a dataset of Canadian Refugee Decisions.

1.1 Research questions

We are interested in evaluating the quality of the entity knowledge learned during pretraining in *off-the-shelf* LMs, specifically domain-specific entities, such as those pertinent to the legal field. **How proficient are Language Models at acquiring knowledge about domain-specific entities like legal entities during pretraining?** Can this acquired knowledge be considered sufficiently reliable for tasks such as annotating new datasets or serving as an indirect source of supervision? How does the choice of prompt type impact the results obtained from knowledge queries? To what extent does the variation in acquired knowledge differ across entity types? What categories of factual knowledge can LMs retrieve, and in what instances do they make errors? Does domain-specific pretraining and jurisdiction-specific pretraining enhance the amount of factual knowledge compared to generic pretraining? To what degree does knowledge acquisition in the legal domain overlap with that of general language models?

1.2 Contributions

Differing from the research objectives of Petroni et al. (2019), which focuses on relation extraction, and Chalkidis et al. (2023) which investigates eight distinct legal knowledge probing tasks with a focus on legislation and legal terminology, we focus on **Legal Entity Types**. To be clear, we ask the LM to predict the entity type, similarly to Epure and Hennequin (2022), and not the actual entity. For example, in the prompt <*Mask*> *is the capital of Germany*, we expect the answer to be *City* or *Location* and not *Berlin*.

In addition, we adopt a comprehensive interpretation of entity types, aligned with the work of Barale et al. (2023), encompassing both essential factual knowledge (e.g., location and dates) and more abstract legal concepts, such as the credibility of a claimant and the rationale behind a judgment. Moreover, our approach diverges by allowing longer entities to be masked (Figure 4), where previous work was limited either to single token (Petroni et al., 2019) or 2-tokens entities (Jiang et al., 2020a; Chalkidis et al., 2023). Where most previous work focuses on masked language modeling objective models (MLM), we introduce the use of autoregressive LM (Llama2) in a zero-shot setting, similar to the approach employed in Epure and Hennequin (2022).

We make the hypothesis that pretrained LMs inherently contain structured knowledge about specific domains, which could be leveraged to generate incidental training instances. We seek to investigate the depth of a model's knowledge, its nature, and whether it predominantly acquires knowledge from semantic or syntactic cues.

We first conduct in section 3 an analysis of the pretraining corpus of selected models both generic and legal LMs. We then prompt the LM with two different styles of prompts, cloze text and question-based, for the task of Entity Typing (section 5). After evaluating the experimental results in section 6, we analyze the type of failure cases (6.5) to highlight the strengths and weaknesses of the learning process and to draft directions for future work.

Our contributions are as follows:

- We propose two new experiments on the task of Legal Entity Typing in a zero-shot setting on a large set of entity types: ***Experiment MLM*** that evaluates generic and legal BERT-based LM on cloze sentences, and ***Experiment Llama2*** which evaluates Llama2 on QA-style prompts.

- We report the results for both experiments and show that Llama2 exhibits good performance on specific entities and has the potential for improvement with optimized prompts. However, law-oriented LMs display inconsistent results, likely influenced by training corpus variations and struggle with Refugee Law-specific vocabulary.

- We propose an in-depth analysis of the failure modes of the models on this task, opening the way for future work.

2 Background and related work

2.1 Legal NLP and Legal LMs

A range of tasks and use cases have been investigated in legal NLP (Zhong et al., 2018), including summarization, information retrieval, and extraction, or question answering. It is worth emphasizing that entity typing is foundational for many of these tasks.

The legal domain presents numerous challenges for self-supervised learning, primarily due to the specificity of legal language in contrast to ordinary language. This can lead to ambiguity in contextual meaning (that we aim to assess in this paper), potential implicit meanings, and variations in the significance of a term. A term that may be decisive in a legal context, such as "appeal," might not carry the same weight in a generic domain.

Given these specific challenges and the demonstrated benefits of pretraining LM on legal text to achieve better performance on downstream tasks

Figure 1: Vocabulary overlap (%) between the pretraining corpora. *Gen* stands for *Generic* and is sampled from sources similar to RoBERTa's pretraining corpus, presented in Table 1. Vocabularies are created with the top 10K most frequent tokens in a sample of 50K documents per model

(Barale et al., 2023), there has been interest in pretraining models on legal texts (Zhong et al., 2018; Xiao et al., 2021). These models typically use an encoder-only architecture based on the BERT architecture: LegalBERT (Chalkidis et al., 2020), CaseHOLD (Zheng et al., 2021), Pile of Law (Henderson et al., 2022), and LexLM (Chalkidis et al., 2023), that we use in our first experiment (details in Table 1). To the best of our knowledge, there is no decoder-only legal LM, which is why we limit our second experiment to a Llama2.

2.2 Probing LMs for Entity Typing

The idea of latent language representations derived from pre-trained LMs holds promise as a source of structured knowledge. Similar to human learning, LMs accumulate domain-specific and linguistic knowledge, along with the development of general pattern recognition capabilities through their pretraining experiences (Brown et al., 2020). As noted in the introduction, our work follows Petroni et al. (2019)'s LAnguage Models Analysis framework (LAMA) and LegalLAMA (Chalkidis et al., 2023). Several probing methods have been investigated (Yin et al., 2023), evaluating multilingual extraction (Jiang et al., 2020a) as well as effective prompting for factual knowledge extraction (Haviv et al., 2021; Qin and Eisner, 2021; Blevins et al., 2023). Various types of tasks have been targeted by these works, including relation extraction, NER, or entity typing (Shen et al., 2023), our task of interest. Concurrently, there have been efforts

to enhance entity typing pipelines, particularly to expand the range of entities beyond traditional categories like location or dates (Choi et al., 2018; Dai et al., 2021) or to entities unseen during training (Epure and Hennequin, 2022; Lin et al., 2020), and approaches leveraging supervision from other tasks such as QA (Zhang et al., 2022). However, to the best of our knowledge, there has been no work conducted in the legal domain specifically addressing entity typing, and no prior research on entity typing in this domain has made use of prompts in the form of questions.

3 Pretraining corpus analysis

3.1 Vocabulary

To understand the difference between pretraining corpora across LMs, we conduct an exploratory vocabulary analysis inspired by Gururangan et al. (2020) that investigates the impact of domain-specific pretraining on a range of downstream tasks. This preliminary study is destined to clarify and offer insights that will help explain the results of the experiments presented in section 5. We select a total of fifty thousand documents for each language model, perform basic cleaning, tokenize the text, and remove stopwords, which gives us a list of tokens per LM. From this list, we select the most common ten thousand tokens, that constitute the final vocabulary for a given LM.

For the three legal LMs, as the datasets used for pretraining are directly available, we randomly select the fifty thousand documents. To construct a generic pretraining corpus, we reconstitute a vocabulary based on the RoBERTa and DeBERTa corpus as indicated in Table 1. As for the other models, we gathered fifty thousand entries, selected proportionally to the size of each corpus. That is to say, we select 5,000 documents from *BookCorpus* which constitutes 10% of RoBERTa pretraining data, 23,750 entries from *CC_news*, 11,875 entries from *OpenWebText* (using the open source version: (Gokaslan and Cohen, 2019)), and 9,688 entries from *CC_stories*. Given our limited knowledge of the precise composition of Llama2's pretraining corpus, we propose utilizing the generic vocabularies of RoBERTa and DeBERTa as suitable proxies for our analysis.

3.2 Vocabulary Overlap

The vocabulary overlap is represented in percentage in the matrix in Figure 1. As anticipated, the legal LMs exhibit a greater overlap in vocabulary compared to their counterparts with generic training. However, significant disparities emerge among the legal LMs. For example, CaseHOLD shares only 44.5% of its vocabulary with LexLM. This observation may be attributed to the more extensive and more diverse set of jurisdictions included in the LexLM pretraining corpus. This aligns with the fact that LexLM shows a higher percentage of vocabulary overlap with Pile of Law, which, in contrast to CaseHOLD which is limited to the United States, also includes legal documents from a broader range of jurisdictions.

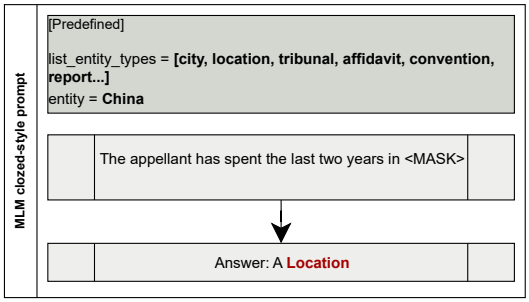

Figure 2: *Experiment MLM* prompt example

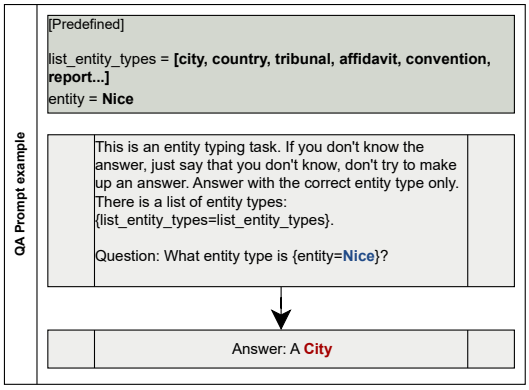

Figure 3: *Experiment Llama2* QA prompt example

4 Dataset

We use ***AsyLex***, a dataset of refugee decisions from Canada curated for entity typing and extraction[1]. This publicly available dataset comprises 19,115 human annotated instances, encompassing 20 distinct categories of entities that hold legal relevance as explained in Barale et al. (2023). These categories have been identified as categories of interest with the collaboration of legal professionals and experts in the field of refugee law. ***AsyLex*** comprises 59,112 historical decision documents,

[1] https://huggingface.co/datasets/clairebarale/AsyLex

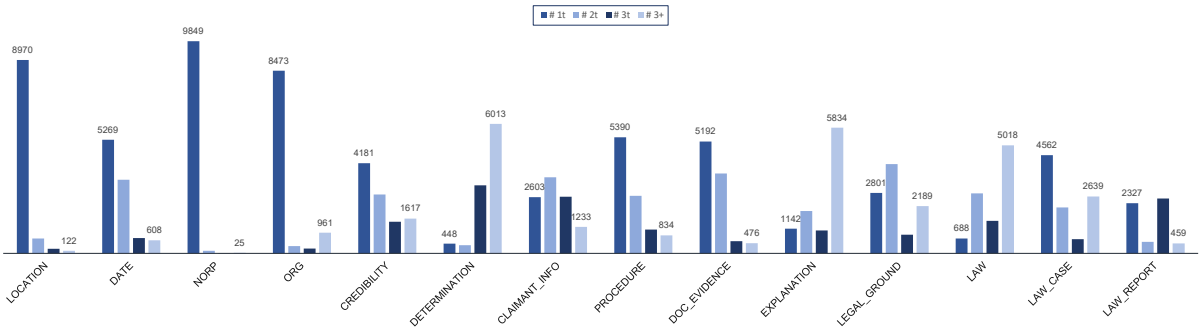

Figure 4: Length of the target masked entities, in number of tokens, for all entity types

spanning from 1996 to 2022. These documents are derived from the online repository of cases of the Canadian Legal Information Institute[2]. The documents encompass both initial determinations and subsequent appeals on whether the claimant is granted refugee status or not. It is important to note that the dataset contains entities of varying generality. Some entities, such as geographical locations, possess broad applicability and could be pertinent to any text (generic entity types: `location`, `date`, `norp`). Others are more specialized within the legal domain, such as procedural steps (generic legal entity types: `org`, `law`, `claimant_info`, `procedure`, `doc_evidence`, `law_case`). Finally, certain entities are highly specific to refugee law, such as the assessment of credibility, which frequently determines the acceptance or rejection of a refugee claim (Refugee Law entity types: `credibility`, `determination`, `explanation`, `legal_ground`, `law_report`). This diversity in entity scope presents an opportunity for assessing the impact of pretraining, particularly in scenarios where entity types have received various exposures during the pretraining phase.

4.1 Legal Entity Types

The selection of legal entity types within this dataset is intended to encapsulate characteristics that can reflect similarities among various refugee cases (see Appendix A for an exhaustive description of the types) and for which we have precise gold-standard annotations (Barale et al., 2023). The set of 14 entity types is pre-defined and closed for both experiments. To extend the coverage of each entity type and extract specific entities, we use a synonym generator to give synonyms for each of our 14 entity types. As a result, when prompted, the model would have to choose between a total of **151 entity types**, increasing the difficulty but also the interest of the task. For example, `location` accepts *city* or *country*. The complete list of synonyms generated per entity type is available in Appendix B. In our evaluation process, we assess predictions across the 14 entity types. For instance, if a prediction is *country*, it will be categorized as `location` and evaluated against a gold answer that specifies `location`. Contrary to previous work, we do not limit the entities' length to a single token (Petroni et al., 2019) or to entities spanning only two tokens (Jiang et al., 2020a). On the contrary, one of the objectives is to use entity types in a broader way for extracting information from text. Thereby we are interested in identifying multi-token entities that have short spans of text and can be longer than two tokens, which is often the case for explaining a decision for instance (entity type: `explanation`). The length of the entity per entity type is presented in Figure 4, which provides explicit numerical values for both single-token entities and entities longer than three tokens.

5 Proposed Entity Typing Methodology

5.1 Task description

The goal of this task is to classify legal entities mentioned in text documents or sentences into specific types. Legal entities can include various organizations, companies, government bodies (`org`), or more abstract concepts such as the credibility assessment made in the context of a refugee claim (`credibility`). The task involves extracting and categorizing these entities based on their attributes or context within the text. As input, we use text documents split by sentences that contain mentions of legal entities. We then categorized legal entity types for each mention found in the input text.

Let E be the set of possible entity types: $E = \{e_1, e_2, \ldots, e_n\}$, S the set of sentences or text seg-

[2]Canlii: https://www.canlii.org/en/

ments, T the set of masked tokens within the sentences and $P(e_i|t_j, s_k)$ represents the conditional probability that masked token t_j in sentence s_k belongs to entity type e_i. The goal is to find the entity type e_i that maximizes the conditional probability for each masked token t_j in each sentence s_k:

$$e_i^* = \arg\max_{e_i \in E} P(e_i|t_j, s_k)$$

In other words, the objective is to find the entity type that is most likely for each masked token in each sentence.

5.2 Language models used

For the first experiment, *Experiment MLM* with BERT-based LMs, we experiment with two generic models optimized for MLM, RoBERTa, and DeBERTa, and three legal-oriented LMs (see Table 1). For the second experiment, *Experiment Llama2* we use the open-source model Llama2, optimized for dialogue use cases. Both tasks take the list of entity types as an input argument, making it a multiple-choice task.

5.3 Cloze prompts with BERT-based models

For the first experimental setting, we use cloze-style prompts that perfectly fit masked language models (MLM). We replace the entities in the sentences with a masked token and use BERT-based models with an MLM objective. Multi-token entities are substituted with a single masked token. If multiple entities appear in the same sentence, only the initial entity occurring in the sentence is considered. The model's answers are limited to the predefined list of 151 entities. We do not provide more context than what is contained in the input sentence. We randomly select ten thousand sentences per entity type, for which we have ground truth annotations (the actual number of prompts after cleaning is given in the column *# prompts* in Table 2). An example of a cloze-style prompt is given in Figure 2. With this *Experiment MLM*, our objective is to assess whether the models can make predictions about the type of entity to expect based on contextual and syntactic cues. For instance, in the example presented in Figure 2, we assume that a human reader could deduce from the context that a location is the expected entity to fill the masked portion. Can an LM do the same?

5.4 QA prompts with Llama2

For the second experimental setting, *Experiment Llama2*, we use a template that briefly explains the task to the model and we input the predefined list of 151 entity types. To provide a simple task framing, we prompt the language model according to the following template: "*What entity types is {entity}*?", to which the model is asked to answer with the most probable entity type. Because of the format of the prompt, we use a text generation objective with an open-source, state-of-the-art auto-regressive LM, Llama2. We use the smallest available version of the model (7B parameters, to spare computing resources) and its fine-tuned version Llama2-chat, which is optimized for dialogue use cases. An example of this QA-style prompt is presented in Figure 3. In that experiment, the prompt explicitly mentions the entity, for example, here the question is *"What entity type is Nice?"* which makes it a simpler task compared to the task of *Experiment MLM*.

6 Experimental Results

We evaluate the results in terms of recall since we want to ensure capturing as many true positives as possible, and F1 score to assess the overall performance on the task. In this section, we compare the results in terms of LM used, length of the input entity, prompt type, and entity type, before conducting an error analysis in the section 6.5. The results of *Experiment MLM* are presented in Table 2 and the results of *Experiment Llama2* in Table 3.

6.1 Language Models Comparison

Given the high difficulty of the task, the choice between 151 entity types when accounting for the synonyms list, and the lack of description of the entities and extra context given, it is no surprise that the scores are relatively low. However, the goal of this work is not to reach the best accuracy, but rather to explore where the models succeed or fail. On *Experiment MLM*, results are generally lower than in *Experiment Llama2* which is firstly explained by the greater difficulty of the task of *Experiment MLM* and the relative lack of context provided for this task. In this experiment, Pile of Law is the model that performs the best on average, in terms of F1, retrieving 16.36% of entity types, with 9.47% in recall. The second best performing model is CaseHOLD with 15.29% average F1 and 8.58% average recall. This is despite LexLM's big-

Type \ Model	RoBERTa		DeBERTa		CaseHOLD		PoL		LexLM		# prompts
	R	F1	R	F1	R	F1	R	F1	R	F1	
LOCATION	0.058	0.110	0.108	0.194	0.070	0.131	0.336	**0.503**	0.055	0.104	9,913
DATE	0.036	0.069	0.100	**0.183**	0.034	0.065	0.025	0.048	0.071	0.133	9,442
NORP	0.037	0.072	0.035	0.067	0.031	0.060	0.032	0.062	0.065	**0.122**	9,986
ORG	0.088	**0.161**	0.066	0.123	0.081	0.149	0.018	0.036	0.074	0.138	9,947
CREDIBILITY	0.028	0.054	0.026	0.051	0.123	**0.219**	0.056	0.106	0.028	0.055	9,527
DETERMINATION	0.384	**0.555**	0.079	0.147	0.070	0.131	0.142	0.249	0.071	0.133	7,242
CLAIMANT_INFO	0.080	0.149	0.060	0.114	0.081	**0.150**	0.079	0.147	0.045	0.085	9,666
PROCEDURE	0.128	0.227	0.080	0.148	0.078	0.145	0.207	**0.344**	0.228	0.129	9,716
DOC_EVIDENCE	0.128	0.228	0.125	0.223	0.188	**0.317**	0.048	0.092	0.056	0.105	9,814
EXPLANATION	0.013	0.026	0.013	0.026	0.009	0.018	0.088	**0.161**	0.008	0.015	8,825
LEGAL_GROUND	0.029	0.056	0.048	0.091	0.045	0.087	0.041	0.079	0.061	**0.116**	9,640
LAW	0.093	0.170	0.203	0.337	0.237	**0.383**	0.061	0.115	0.066	0.124	9,128
LAW_CASE	0.079	0.146	0.071	0.133	0.058	0.109	0.106	**0.191**	0.091	0.167	9,290
LAW_REPORT	0.057	0.107	0.075	0.140	0.098	**0.178**	0.087	0.160	0.089	0.164	8,601

Table 2: Entity type prediction scores in a zero-shot setting, on cloze sentences, measured in Recall and F1 score across 2 generic LMs (RoBERTa and DeBERTa-V3), and 3 legal LMs (CaseHOLD, Pile of Law and LexLM)

Type	R	F1
LOCATION	**0.956**	**0.916**
DATE	0.730	0.575
NORP	0.211	0.118
ORG	0.098	0.051
LAW	0.100	0.053
CREDIBILITY	0.219	0.123
DETERMINATION	0.357	0.217
CLAIMANT_INFO	0.627	0.456
PROCEDURE	0.259	0.149
DOC_EVIDENCE	0.653	0.485
EXPLANATION	0.006	0.003
LEGAL_GROUND	0.022	0.011
LAW_CASE	0.034	0.017
LAW_REPORT	0.048	0.025

Table 3: Entity type prediction scores in a zero-shot setting, on QA-style prompts, measured in Recall and F1 score, with Llama2, on 10K prompts per entity

ger size, LexLM being the model that performs the worst across all, being outperformed by generic LMs RoBERTa and DeBERTa. For all entities except one, the model that achieved the best recall also achieved the best F1, highlighting the consistency in the precision. The only exception is the type `procedure` for which the best F1 is reached with Pile of Law and the best recall with LexLM.

6.2 Single-token vs Multi-token

An interesting point is that we did not impose any restrictions on the length of entities; the entities that tend to be longer are typically more abstract and closer to a piece of legal commonsense knowledge and reasoning, for example, `explanation`, `determination`, `credibilty` and `legal ground`. Interestingly the best overall F1 score in *Experiment MLM* is achieved for the type `determination`, reaching an F1 score of 55.5% (RoBERTa). For instance, an entity flagged as `determination` can as long as: *claimants are not convention refugees and not persons in need of protection*. While the other models achieve lower scores for this entity type, it is to note that the disparity between these relatively lengthy multi-token entities and those that are typically single tokens is not substantial (refer to Figure 4). This may be due to the nature of the task, which may mitigate such disparities compared to tasks like NER where the model has to retrieve the actual entity. In *Experiment Llama2*, shorter entities (that are also the most generic ones) are well recognized (`location`, `date`), with also good scores achieved on the types `determination, claimant_info, procedure`. Overall for both experiments and certainly due to the nature of the task of entity typing, it seems that the length of the initial entity to categorize does not have an impact on the results.

6.3 Prompt Templates Comparison

The scores are on average higher in the *Experiment Llama2* with a total average F1 score of 30.86% when *Experiment MLM* reaches an average of 14.46% across all types. However, as noted in

	Error Type	Prompt example	Prediction	Gold	%
MLM	Random Prediction	under <mask> of the Republic of China, they cannot take on a second citizenship	lawsuit	`law`	70.71
	Contextually Accurate	the applicant has not returned to <mask> since 2008	employment	`location`	12.43
	Closely Related	my colleague relied on this <mask> in her conclusion	ngo report	`doc_evidence`	16.86
Llama2	Random Prediction	What is *Subsection 648*?	country	`law`	22.22
	Closely Related	What is *vietnamese*?	country	nationality (`norp`)	18.52
	False Negative	What is *female claimant*?	female claimant	gender (`claimant_info`)	33.33
	Prompt Error	What is *removal order*?	It is a type of judicial decision.	`procedure`	25.93

Table 4: Error cases and the ratio of the different error types for both experiments, across all tested models

	Gen	CH	PoL	LexLM	Llama2
Generic	11.59	8.51	**20.42**	11.97	**63.26**
Gen Legal	17.98	**20.89**	15.40	12.48	**29.52**
Refugee Law	**10.01**	5.93	4.68	4.92	**13.03**

Table 5: Entity types prediction scores averaged on 3 groups: generic (`location, date, norp`), legal entities applicable to most legal domains (Gen Legal: `org, law, claimant_info, procedure, doc_evidence, law_case`), and legal entities specific to refugee law (Refugee Law: `credibility, determination, explanation, legal_ground, law_report`) *Gen* groups the results of RoBERTa and DeBERTa-v3, *CH* refers to CaseHOLD

the task description (5.1) the QA-based experiment is a relatively easier task, making the comparison difficult.

Based on the predicted entity types, it appears that the template suggested for *Experiment Llama2* is not consistently well comprehended, resulting in a lack of clarity regarding the task. In some cases, it returns not just one entity type, but multiple, leading to incorrect predictions. It seems that, instead of relying on manually crafted prompts and templates, which have been acknowledged to be su-optimal as mentioned by Jiang et al. (2020a), there is significant room for improvement in this regard.

6.4 Entity Types Comparison

For this evaluation, we categorize the type of entity into three groups: those that can be encountered in any text with the same meaning, those that are commonly found in legal texts, and those that are highly specific to the domain of the dataset, refugee law. The combined results are summarized in Table 5. Entities related to refugee law tend to yield the lowest performance across all models and settings. Pile of Law outperforms other models even on generic entities. At the same time, RoBERTa and DeBERTa surpass models specifically trained on legal data for generic legal entities, possibly due to larger exposure and a larger vocabulary.

6.5 Failure Cases Analysis

We identify four types of errors across the two experiments:

1. Random Prediction: this refers to cases where the predicted entity type is entirely random and unrelated to the context.

2. Contextually Accurate: this describes situations where the predicted entity type is incorrect, but within the context of the sentences, it is plausible in terms of syntax and meaning.

3. Closely Related: instances where the predicted entity class is incorrect, yet it is closely related to the actual gold entity type. For example, it misclassifies a legal ground (which is very precisely one of the 5 reasons for being granted refugee status, see Appendix A) for an explanation of the decision (which is more generic).

4. False Negative, it predicts an entity type that is not in the list of entity types given as input.

5. Prompt Error, if the answer provided deviates from the prompt instruction, we categorize it as incorrect; we consistently consider that an answer with more than five tokens is incorrect, as it signifies that the response extends beyond providing just the entity type.

***Experiment MLM* errors** To assess the occurrence of error types, we sample 10 errors per entity type per model, i.e. a total of 700 errors for this experiment. Table 4 presents the findings and an example per error type. There is no instance of a False Negative error; the models never predict an entity type that is not in the input predefined list as we constraint the model to a multiple-choice task, from our pre-defined list of entity types. The most common error is simply an incorrect prediction, with the second most frequent error being the prediction of a closely related entity. This may be due to the choice of categories, some of which express subtle legal nuances. Another positive sign is the

presence of more than 10% of incorrect predictions that are nevertheless accurate in the context of the input sentence.

***Experiment Llama2* errors** Similarly, we sample 10 errors per entity type, i.e. a total of 135 errors. Table 4 presents the findings. It's worth noting that the use of QA-style prompts leads to a significant number of prompt errors and false negatives, which we believe could be mitigated to some extent by improving the initial prompt template in future work. Additionally, a common misclassification pattern occurs with `norp` entities, which are always adjectives, but are misclassified as their noun counterparts, as illustrated in the example provided in Table 4. Similarly, acronyms for tribunals (e.g., *RPD* for *Refugee Protection Division*) are classified as *units of length*, an issue that might be rectified by providing more contextual information. Finally, entities like *consistent explanation* are occasionally misclassified as `explanation` when they should be categorized as a `credibility` assessment, possibly due to missing adjectives or entity length-related challenges.

7 Conclusion

Our investigation includes LM selection, input entity length, prompt types, and entity types, in an attempt to understand model strengths and limitations. In summary, our study shows that Llama2 performs best on specific entities and displays potential for improvement with better prompting strategies. However, it also seems that Llama2 repeatedly overlooks syntactic cues. Masked language models mostly appear to be lacking sufficient context within our experimental setup, where they are confronted with a highly challenging task. Law-oriented LMs exhibit varying results, possibly influenced by training corpus differences, and the Pile of Law model shows the best performance on our *AsyLex* dataset. Despite inherent challenges, LMs can accurately identify certain entity types, including multi-token ones, but encounter difficulties with legal sub-domains like Refugee Law. Future research may explore optimized prompts and few-shot learning strategies. Furthermore, assessing the average precision of the entity type ranking predictions generated by the LM, and conducting experiments on additional datasets, would also be necessary.

References

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Claire Barale, Michael Rovatsos, and Nehal Bhuta. 2023. Automated refugee case analysis: A NLP pipeline for supporting legal practitioners. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2992–3005, Toronto, Canada. Association for Computational Linguistics.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting language models for linguistic structure. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663, Toronto, Canada. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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, Online. Association for Computational Linguistics.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535, Toronto, Canada. Association for Computational Linguistics.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 87–96, Melbourne, Australia. Association for Computational Linguistics.

Hongliang Dai, Yangqiu Song, and Haixun Wang. 2021. Ultra-fine entity typing with weak supervision from a masked language model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1790–1799, Online. Association for Computational Linguistics.

Elena V. Epure and Romain Hennequin. 2022. Probing pre-trained auto-regressive language models for named entity typing and recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1408–1417, Marseille, France. European Language Resources Association.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus.

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

Adi Haviv, Jonathan Berant, and Amir Globerson. 2021. BERTese: Learning to speak to BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3618–3623, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

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.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020a. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020b. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Bill Yuchen Lin, Dong-Ho Lee, Ming Shen, Ryan Moreno, Xiao Huang, Prashant Shiralkar, and Xiang Ren. 2020. TriggerNER: Learning with entity triggers as explanations for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8503–8511, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sebastian Nagel. 2016.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jaromir Savelka. 2023. Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 447–451, New York, NY, USA. Association for Computing Machinery.

Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. 2023. PromptNER: Prompt locating and typing for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12492–12507, Toronto, Canada. Association for Computational Linguistics.

Simone Tedeschi and Roberto Navigli. 2022. MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

Wenzheng Zhang, Wenyue Hua, and Karl Stratos. 2022. Entqa: Entity linking as question answering. In *International Conference on Learning Representations (ICLR)*.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*. Association for Computing Machinery.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision, ICCV*, pages 19–27.

Appendix

Type	Description	Examples
LOCATION	cities, countries, regions	"toronto, ontario"
DATE	absolute or relative dates or periods	"june, 4th 1996", "two years"
NORP	adjectives of nationalities, religious, political or ethnic groups or communities	"hutu", "nigerian", "christian"
ORG	tribunals, companies, NGOs	"immigration appeal division", "human rights watch"
CREDIBILITY	mentions of credibility	"lack of evidence", "inconsistencies"
DETERMINATION	outcome of the decision (accept/reject)	"appeal is dismissed", "not a convention refugee"
CLAIMANT_INFO	age, gender, citizenship, occupation	"28 year old", "citizen of Iran", "female"
PROCEDURE	steps in the claim and legal procedure events	"removal order", "sponsorship for application"
DOC_EVIDENCE	pieces of evidence, proofs, supporting documents	"passport", "medical record", "marriage certificate"
EXPLANATION	reasons given by the panel for the determination	"fear of persecution", "no protection by the state"
LEGAL_GROUND	referring to the Convention, refugee status is granted for reasons of race, religion, nationality, membership of a particular social group or political opinion	"homosexual", "christian"
LAW	citations: legislation and international conventions	"section 1(a) of the convention"
LAW_CASE	citations: case law and past decided cases	"xxx v. minister of canada, 1994"
LAW_REPORT	country reports written by NGOs or the United Nations	"amnesty international: police and military torture of women in mexico, 2016"

Table 6: Pre-defined list of legal entity types

Type	Extended List
LOCATION	city, country, region, state, province, area, nation, land, republic, district, territory, division, zone
DATE	date, day of the month, appointment, particular date, date stamp, time, timestamp, calendar date, schedule
NORP	nationality, religious community, political group, ethnic groups, community, racial group, party, faction, ideological group, belief community
ORG	tribunal, firm, ngo, company, corporation, business, nonprofit, association, charity, court, judicial body
CREDIBILITY	plausibility, authenticity, integrity, trustworthiness, reliability, credibility, believability, credibility, credibleness
DETERMINATION	verdict", result, resolution, judgment, approval, denial, decline, rejection, approval, determination, finding, conclusion, decision, grant, refusal, positive decision, negative decision
CLAIMANT_INFO	data, employment, resident, national, inhabitant, information, gender, age, citizen, citizenship, sex, job, occupation, profession
PROCEDURE	affidavit, documentary evidence, proof, testimony, exhibit, record, file, paperwork, operation, procedure, legal procedure, legal process, judicial procedure, legal steps, judicial process
DOC_EVIDENCE	proof, evidence, document, written document, written evidence, written proof, written record, written report, written statement, written testimony, written witness statement
EXPLANATION	explanation, clarification, interpretation
LEGAL_GROUND	reason, ground, legal ground, justification, rationale, foundation, legal basis, legal justification
LAW	convention, international convention, law, legislation, legal code, treaty, agreement, protocol, statute
LAW_CASE	citation, jurisprudence, case, law, case law, legal case, lawsuit, legal matter, legal precedent, judicial decisions, legal rulings
LAW_REPORT	country report, report, official report, written report, ngo report, national report, state report, regional report, nonprofit report, non-governmental organization report, charity report

Table 7: Extended pre-defined list of legal entity types (151 types)

Error Type	RoBERTa	DeBERTa	CaseHOLD	PoL	LexLM	# Total	%
Random Prediction	109	81	90	112	103	495	70.71
Contextually Accurate	7	27	25	16	12	87	12.43
Closely Related	24	32	25	12	25	118	16.86
False Negative	-	-	-	-	-	-	0.00
Total	140	140	140	140	140	700	100

Table 8: Error types figures per (number of occurrences) and in percentage for all studied LM