

Citation Recognition in Large-Scale Legal Platforms using Transformer Models

Mirio Eggmann
mirio.eggmann@pm.me

Jasmin Fitz
research@fitz-zh.ch

Dario Glasl
dglasl@swisslex.ch

Mitra Purandare
mitra.purandare@ost.ch

*Department of Computer Science
OST - Eastern Switzerland University of Applied Sciences
Rapperswil, Switzerland*

Abstract—Online platforms are popular among legal professionals for daily searches of information needed for drafting opinions, preparing court cases etc. These platforms must handle large volumes of documents ensuring timely availability of information. A key challenge is identifying and linking citations in the documents to their sources within the limited time available for nightly re-indexing.

Traditional methods, such as regular expressions, are fast but lack flexibility and do not support context-dependent citations. The latest state-of-the-art generative language models lack efficiency for the task, processing only 36 documents per hour. We propose transformer-based encoder-only models to recognize citations. We demonstrate that the discriminative BERT-based models can process 3,000 documents in about 32 minutes, exceeding the performance requirements and ensuring scalability and efficiency of large-scale legal platforms.

Index Terms—Named Entity Recognition, Natural Language Processing, Language Models, Legal-Tech, Deep Learning, Transformers

I. INTRODUCTION

Swiss law courts, legislators, and regulators produce large volumes of legal data, e.g., court judgments, on a daily basis. This renders its manual processing and application infeasible. Consequently, legal professionals resort to online platforms as a primary source of latest legal information. An online provider of legal information daily receives a large volume of data like laws, judgements etc. The online platform must offer reliable search and analysis over millions of documents to support the legal professionals in drafting legal opinions, preparing court cases etc. Automated tools are essential for effective management of legal information.

Citations are a crucial component of legal text. To show that a proposition is supported by law, legal documents contain citations to laws, court decisions, or other legal literature. The citations in the legal documents received from legal sources are almost always plain text. A plain text citation prohibits its further use as it hinders efficient search and information retrieval. Recognizing a citation and linking it to the referenced article makes legal information accessible, searchable, and cheaper. This in turn reduces the cost of legal research by making information easily navigable on the web-platform of the online provider.

To link the cited documents, it is essential to identify citations in the text and also, decompose the citations further in sub-parts. While real-time performance is not

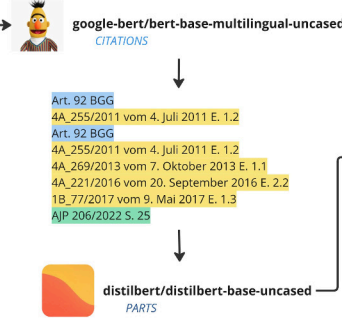
the focus for the online platform, the platform must be capable of processing approximately 3,000 documents during the nightly re-indexing. A typical nightly re-indexing time is about 6 hours between midnight and early morning to ensure updated searchability by the next working day. Common solutions like regular expressions (regex) and rule-based matching for citation identification are fast but are hard to maintain and lack flexibility. They struggle with new citation formats, typographical errors, and most importantly, context-dependent citations. In contrast, transformer-based models [1] are better suited for this task due to their ability to understand context and generalize across citation styles. They not only improve robustness and accuracy, but also significantly enhance maintainability: new citation types or style changes can be accommodated through model retraining or even few shot prompting instead of laborious rule or regex adjustments.

Owing to the breakthrough of attention mechanisms [1] and the consequent rise of large language models (LLMs) [2], [3], there is a growing trend of using generative LLMs for a wide range of tasks, often as a default solution. For the task of citation recognition, the LLMs are too slow. The inference using Meta-Llama-3.1-8B-Instruct-bnb-4bit with vLLM takes 16.33 minutes for 10 documents, showing lack of efficiency and scalability for our task. This demonstrates the unsuitability of generative LLMs for timely processing and structured retrieval of legal information.

We overcome these limitations by leveraging discriminative transformer-based models [4]. We fine-tune a Google BERT multilingual base model (uncased) to identify and classify citations into the three categories CASELAW, LAW and LITERATURE. We then fine-tune DistilBERT (uncased) [5] to identify citation components like COURT, DATE, ARTICLE, etc. We distill knowledge [6] from Llama 3.3 70B [7], [8] to fine-tune DistilBERT, resulting in a three orders of magnitude smaller model. The resulting chain of models depicted in Fig. 1 achieves a two orders of magnitude speedup while maintaining precision and recall on par with Llama.

Our paper is organized as follows: Section II presents background information such as related work and domain knowledge about legal data including its challenging specifics. It is followed by Section III describing our approach. Section IV shows evaluation results and Section V concludes the work.

dass selbständig eröffnete Zwischenentscheide über den Ausstand eines Gutachters nach Art. 92 BGG in der Regel zwar anfechtbar sind und nicht erst zusammen mit dem Endentscheid angefochten werden können, womit verhindert werden soll, dass ein Experte mit einem Gutachten betraut wird, dessen Ergebnisse wegen Befangenheit ohnehin nicht verwertet werden könnten (Urteil 4A_255/2011 vom 4. Juli 2011 E. 1.2); dass Art. 92 BGG indessen nicht zum Tragen kommt, wenn das Gutachten bereits erstellt ist und es mithin lediglich darum geht, ob dieses im Recht liegende Beweismittel verwertbar ist, da es mit Bezug auf die Verwertbarkeit eines Beweismittels nicht darauf ankommen kann, ob die behauptete Unverwertbarkeit sich aus der Befangenheit des Gutachters oder aus anderen Gründen ergibt (Urteile 4A_255/2011 vom 4. Juli 2011 E. 1.2; 4A_269/2013 vom 7. Oktober 2013 E. 1.1; vgl. für den Ausstand einer Richterin bzw. eines Staatsanwalts: Urteile 4A_221/2016 vom 20. September 2016 E. 2.2; 1B_77/2017 vom 9. Mai 2017 E. 1.3); auch publiziert in AJP 2016/2022 S. 25



Citation	Confidence	Type
Art. 92 Article BGG Law	99.92%	Law
4A_255/2011 Dossiernummer vom 4. Juli 2011 Date E. 1.2 Consideration	99.65%	Caselaw
Art. 92 Article BGG Law	99.89%	Law
4A_255/2011 Dossiernummer vom 4. Juli 2011 Date E. 1.2 Consideration	99.66%	Caselaw
4A_269/2013 Dossiernummer vom 7. Oktober 2013 Date E. 1.1 Consideration	99.65%	Caselaw
4A_221/2016 Dossiernummer vom 20. September 2016 Date E. 2.2 Consideration	99.68%	Caselaw
1B_77/2017 Dossiernummer vom 9. Mai 2017 Date E. 1.3 Consideration	99.63%	Caselaw
AJP Publisher 206 Volume /2022 Year S. 25 Page	86.57%	Literature

Fig. 1: Proposed citation recognition approach using fine-tuned BERT variants.

II. BACKGROUND

A. Related Work

Named Entity Recognition (NER) is an important task in the field of Natural Language Processing (NLP). NER identifies named entities in a text and classifies them into predefined categories. Most NER systems are designed to recognize common types of entities such as people, locations, and organizations. NER has been applied to the legal domain to extract such information from legal texts [9], [10]. A German research group has developed a German BERT model for legal NER [11]. SwissBERT is a multilingual model fine-tuned on Switzerland-related news articles in the languages German, French, Italian, and Romansh [12].

Some language models known as LegalLMs [13] start from the standard XLM-RoBERTa model [14] and are additionally pre-trained on the MultiLegalPile dataset [15], a multilingual corpus of legal texts covering 24 languages. LEGAL-BERT [16] is another family of BERT models [17] for the legal domain, intended to assist legal NLP.

Legal text processing poses a variety of challenges which are addressed as a part of the COLIEE 2021 automatic legal text processing competition [18], [19]. Albeit different from our work, most research in legal domain is focused on inferring missing and most appropriate new citations in text [20], [21]. Transformer models can also be applied to predict where to add a new reference and which new reference to cite [22].

Our work is inspired by the current research developments in legal text processing for identifying missing citations. However, none of the existing works focus on the problem presented in this paper, namely recognition of existing citations!

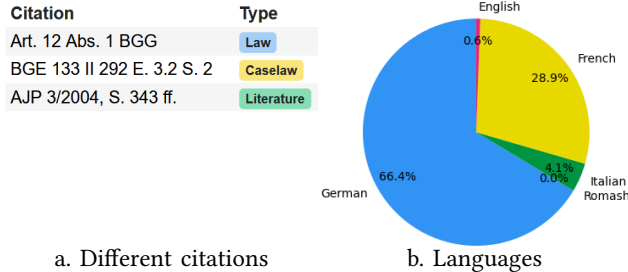


Fig. 2: Legal documents: languages and citations

B. Legal Data

The paper focuses on the recognition of citations. Fig. 2 a. shows the most common types of citations: references to laws (LAW), judicial decisions (CASELAW), and literature citations (LITERATURE). Fig. 2 b. shows the multilingual nature of the documents.

Judgments of the Federal Supreme Court of Switzerland are cited differently depending on where they are published, e.g., official publication in a printed booklet [23], a journal etc. Judgments of the Federal Administrative Court and Federal Criminal Court follow different citation rules. In addition to the federal courts, cantonal and district courts also follow different citation methods. The following citations refer to the same document: ATF 121 V 362 consid. 1b ; 99 V 98 consid. 4, BGE 121 V 362 E. 1b; 99 V 98 117 V 287 cons. 4, 99 V 98 cons. 4, 121 V 362 Erw. 1b und 99 V 98 Erw. 4, ATF 99 V 98, BGE 99 V 98, DTF 99 V 98, 99 V 98.

Some of these variations occur because of language differences (ATF in French, BGE in German, and DTF in Italian), others because of different ways to separate multiple citations. Adding to the difficulty, various courts use abbreviations, e.g., Italian-speaking courts anonymize people with acronyms like CO. CO is also the French and Italian abbreviation for the Swiss Code of Obligations. In cases like these, the context is required to know if a law is cited or merely a person is referred to. Traditional methods fail to distinguish between the two cases.

Fig. 3 shows that the legal documents may contain long citations (> 700 characters). Based on the citation rules of

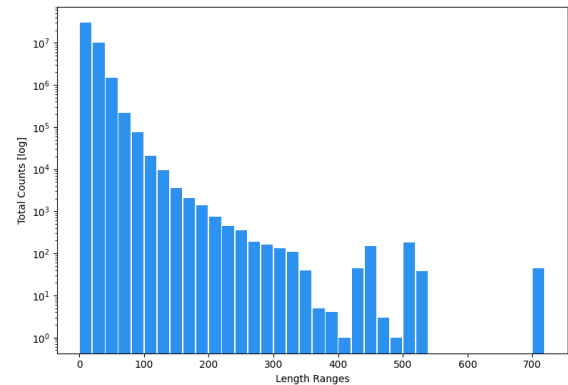


Fig. 3: Number of characters in citations

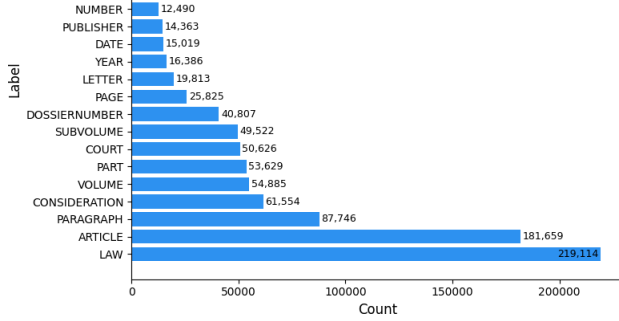


Fig. 4: 15 of the 21 different part types used in the citations of the Swiss Federal Court

the Swiss Federal Court, Fig. 4 shows that there are about 21 different part types, e.g., VOLUME, COURT, PUBLISHER, DATE, SENTENCE, CONSIDERATION, REFERENCE etc. The numerous types demand a fine-grained classification scheme to capture the different parts of a citation.

III. METHODS

The task of citation recognition is split into two sub-tasks:

- Task 1: recognize citations in a document and classify the citations into the three categories CASELAW, LAW and LITERATURE.
- Task 2: decompose the citation into sub-parts.

Both tasks can be accomplished by discriminative models based on encoder-only transformer models as well as generative-style decoder-only transformer models.

a) *Using Generative Models*: This approach involves inference using few-shot prompting on a generative model. Models from closed-source vendors like OpenAI, Anthropic etc. are ruled out due to security and data privacy risks. Additionally, external hosting is not an option because it leads to vendor lock-in, making provider switching difficult and costly. If the performance of off-the-shelf inference is not acceptable, models can also be fine-tuned for the specific tasks. We choose open-source quantized Unsloth Llama 3.2 1B Instruct [24], Llama 3.2 3B Instruct [25] and Llama 3.1 8B Instruct [26] and fine-tune these for citation recognition.

b) *Using Discriminative Models*: Fig. 1 illustrates the model chaining approach using BERT models. NER is employed for classification of tokens into REGULAR and CITATION tags. A second discriminative BERT model, e.g., DistillBERT, further identifies its sub-parts. Dividing the task enables us to develop optimized models for each sub-task. The output can be analyzed at each step, increasing explainability as well as ease of debugging and error analysis. Moreover, the second model receives only the relevant text as input, benefitting its performance.

The tags should themselves represent the distinct types of citations and their parts. The citation tokens are differentiated into the most common types of citations, namely, LAW, CASELAW and LITERATURE. Fig. 5 shows that the tags follow the “inside-outside-beginning” (IOB2) format. A B-CASELAW tag indicates the beginning of a CASELAW and

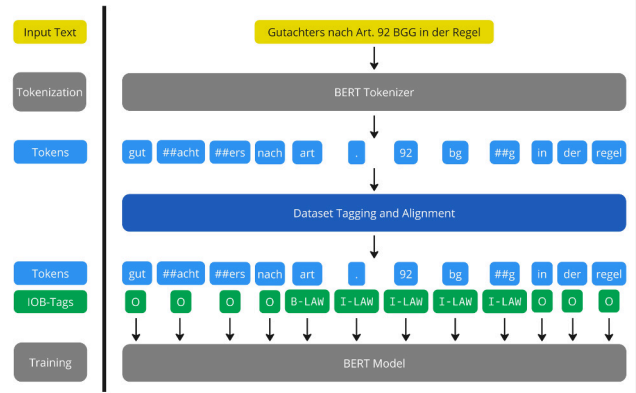


Fig. 5: Input tokenization and NER tags of a BERT-style discriminative model

an I-CASELAW tag indicates continuation of a CASELAW. An O-tag indicates that the token is not a citation token.

Furthermore, various citation styles require a granular classification scheme to capture parts in a citation. There is a NER tag corresponding to each of the 21 part types used in the citation rules of the Swiss Federal Court, shown in Fig. 4. These tags are used to precisely distinguish the different parts.

A. Data Processing

A large amount of already recognized citations is available. These comprise legal documents and labeled citations therein identified using rule-based and regex methods. Fig. 6 depicts our data preprocessing step involving deduplication and splitting into the usual train, validate and test sets.

Using Prodigy [27], we manually annotate a few hand-picked documents covering a variety of document types and citation types. To assess the models thoroughly, we separate the manually annotated set into special cases set and golden test set. The golden test set is a randomly chosen sample of the available data. The special test set contains special corner cases like multiple citations, context-dependent citations, citations with spelling mistakes etc., that are not correctly identified by the traditional methods.

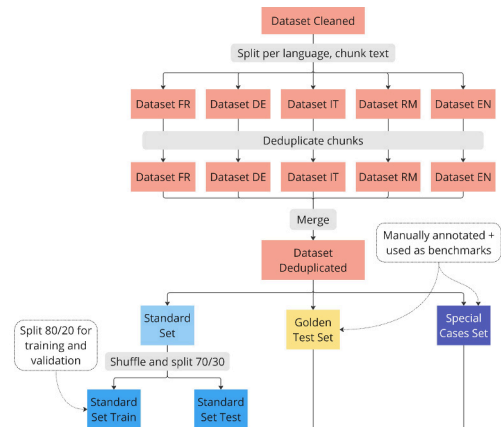


Fig. 6: Data processing via chunking, deduplication, and train-test split

a) *Chunking*: Due to the context size limitations of the models, documents are divided into smaller chunks for fine-tuning. Based on Fig. 3, we divide the documents into chunks of max. 1,000 characters for all models including Llama. Llama models support even larger contexts. However, larger context comes with a speed penalty hindering scalability [28]. A proper chunking size makes sure that sufficient context is provided to the model, allowing the use of surrounding information [29]. Care is taken not to split citations during chunking.

b) *Deduplication of Chunks*: The dataset contains many duplicate chunks. We use the MinHash algorithm to filter out similar chunks. The MinHash algorithm is a probabilistic method for estimating the similarity between two sets by hashing the elements of the sets [30]. The deduplication process is implemented using the DataTrove library from Hugging Face [31]. Deduplication removes around 12% of the chunks.

B. Task 1: Recognizing Citations

Table I lists the selected models for the task. Most of the models are based on BERT. Note that GLiNER [32] is also a NER model that uses a bidirectional transformer encoder (BERT-like) for parallel entity extraction. Two Switzerland-specific models are included, namely, ZurichNLP/swissbert [33] and ZurichNLP/swissbert-NER. These models support the Swiss languages in the dataset, namely, German, French, Italian, and Romansh. These models come already pre-trained on Swiss specific text, which may help with the legal text.

Quantization is a technique that reduces the precision of a model’s numerical computations in order to decrease memory and computational requirements. It is widely used to make the models faster and cost-effective for large-scale applications. In addition to the BERT variants, we also include quantized models meta-llama/Llama-3.1-8B-Instruct and unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit in our list.

a) *Fine-Tuning*: We employ LoRA [34] with its Hugging Face implementation [35] of the PEFT library [36] to fine-tune the Llama and BERT models. To find the best hyperparameters for the models, the Ray Tune library [37] is used.

C. Task 2: Recognizing Parts in a Citation

Unlike task 1, this task suffers from a cold start due to the lack of training data. No existing solution is present to generate the parts information. One possibility is to reverse engineer the parts from the matching regular expression from the named groups. It is a tedious task as the group names can be misnomer, not informative, and not fine-grained. Moreover, one needs to re-execute the regular expression to get the parts. This is not always possible. The matching regular expression may not be available as it can be generated dynamically, stitching together expressions during production.

In such cases, in-context learning with off-the-shelf generative models is a natural choice. The Llama models (unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit and unsloth/Llama-3.3-70B-Instruct-bnb-4bit) deliver strong performance

TABLE I: SPECIFICATIONS OF THE EVALUATED MODELS

Model	Params	Lang.	Cased
urchade/gliner_large-v2.1	459.0M	Eng	Y
urchade/gliner_medium-v2.1	209.0M	Eng	Y
urchade/gliner_small-v2.1	166.0M	Eng	Y
urchade/gliner_multi-v2.1	209.0M	Multi	Y
google-bert/bert-base-cased	109.0M	Eng	Y
google-bert/bert-base-uncased	110.0M	Eng	N
google-bert/bert-base-multil-uncased	168.0M	Multi	N
google-bert/bert-large-uncased	336.0M	Eng	N
dslim/bert-base-NER-uncased	109.0M	Eng	N
dslim/bert-large-NER	334.0M	Eng	Y
distilbert/distilbert-base-uncased	67.0M	Eng	N
distilbert/distilbert-base-cased	65.8M	Eng	Y
FacebookAI/xlm-roberta-large	561.0M	Multi	Y
microsoft/deberta-v3-base	86.0M	Eng	Y
microsoft/deberta-v3-large	304.0M	Eng	Y
microsoft/mdeberta-v3-base	86.0M	Multi	Y
ZurichNLP/swissbert	160.0M	Multi	Y
ZurichNLP/swissbert-ner	152.0M	Multi	Y
meta-llama/Llama-3.2-1B-Instruct	1.2B	Multi	Y
meta-llama/Llama-3.2-3B-Instruct	3.2B	Multi	Y
unsloth/Meta-Llama-3.1-8B-Inst-bnb-4bit	8.0B	Multi	Y

when prompted with examples. However, even the smaller Meta-Llama-3.1-8B-Instruct-bnb-4bit can process only 36 documents per hour — making it unsuitable for deployment on our constrained hardware (one DGX H100).

Discriminative BERT-based models are significantly smaller, faster and hence, suitable for recognizing citations in production. Since the generative Llama model achieves higher accuracy, it can serve as a teacher to train a smaller, more efficient student model — a technique known as knowledge distillation [38]. Fig. 7 illustrates our pipeline for training compact discriminative models guided by the Llama model. We prompt the teacher model to identify parts of citations. The citations with their part labels are then used to fine-tune the student model (hard distillation).

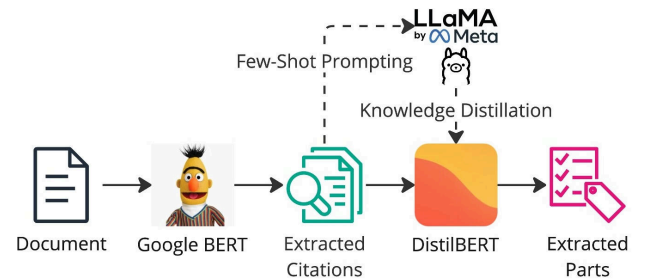


Fig. 7: Knowledge distillation with Llama (teacher) and DistilBERT (student)

For the citations generated after data processing shown in Fig. 6, we use few-shot prompting using the teacher to identify parts in the citations. The citation grouping into training, validation, test, golden and special case sets remains as described before. We analyze the two Llama variants to serve as the teacher. We use vLLM, an optimized inference library to run inference. The final teacher model is selected by evaluating the models on the hand-crafted golden test set of 100 citations. Table II justifies the use of the larger 70B model as a better teacher.

TABLE II: PRECISION (P), RECALL (R), AND F1 SCORES OF TEACHER LLMs ON THE GOLDEN TEST SET

Model	P	R	F1
unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit	0.92	0.93	0.92
unsloth/Llama-3.3-70B-Instruct-bnb-4bit	0.99	0.98	0.99

IV. RESULTS

For the training and inference results in this section, we use a single DGX H100 with 80GB of GPU memory.

A. Task 1: Recognizing Citations

We employ phased model selection. Training is split into multiple rounds with increasing amount of training data. Only the winning models with the highest recall test scores as well as acceptable inference and training times are selected for the next training round.

Importance of High Recall: High recall is a key priority for citation recognition as it influences downstream document linking. False positives (recognized citations that do not actually exist) simply do not get linked later in the process, causing no significant impact. However, false negatives (citations that exist but are not recognized) result in valid citations being missed. Ensuring high recall is a must to ensure quality of the legal platform for the legal professionals.

a) *Phase 1:* We train all models listed in Table I on 100,000 training chunks. The test scores in Fig. 8 show that ‘uncased’ models perform consistently better than their ‘cased’ counterparts. Capitalization does not alter the meaning in legal citations - cased models introduce unnecessary complexity and fail to generalize [4].

Language of the Models: The language of the model has only a small impact on the performance. Legal citations across the most common languages in the training set (German, French, English, Italian) often follow similar numeric and symbolic conventions (e.g. Art. 2 or § 15). The evaluation demonstrates that these patterns are effectively captured by the multilingual models. However, since these patterns rely less on language-specific contextual nuances, the English models perform comparably well.

Size of the Models: The best performing model (DistilBERT base uncased) has only 67M parameters. The largest model with over 8B parameters (Llama 3.1 8B) is at the 5th place. Google BERT-large has 24 layers. Its base version has only 12 layers, and DistilBERT reduces it further to 6 [4], [5]. More layers allow capturing complex patterns, semantic

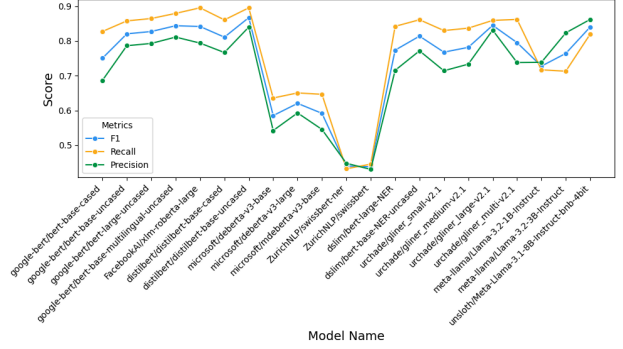


Fig. 8: Phase 1: F1, precision, recall test scores after training models on 100K samples

relationships, and long-range dependencies in the data. Legal citations are usually short, and their parts are close together. Citations often follow structured patterns that can be effectively recognized by the lower layers of the model. This explains why smaller models outperform larger ones.

A new version of BERT called ModernBERT [39] extends the standard transformer architecture [1] to be smarter, better, and faster, supporting longer context (8,000 tokens). However, the fine-tuned ModernBERT on 100k samples exhibited an F1 score of only 45.5%. Since we do not need such a long context, as seen in Fig. 3, it is not pursued further.

b) *Phase 2:* The best performing models in phase 1 are then trained on 1 million training samples. Table III lists the performance metrics and the training duration.

Generative vs. Discriminative Models: Contrary to the trend of using LLMs via in-context learning (aka, few-shot prompting), Table III shows that generative models do not satisfy our performance constraints. Discriminative models are equally performant in terms of precision and recall. Moreover, they are smaller and consequently, faster to train (avg. 1.4 hours for BERT vs. 1.4 days for Llama). Faster training allows us to support new evolving citations as the training can also be done overnight. We eliminate models with prohibitively large training times for the next phase.

TABLE III: PHASE 2: PRECISION (P), RECALL (R), AND F1 TEST SCORES AFTER TRAINING ON 1 MILLION SAMPLES

Model	P	R	F1	T
google-bert/bert-large-uncased	0.93	0.95	0.94	3.3h
google-bert/bert-base-multilingual-uncased	0.92	0.95	0.94	1.7h
FacebookAI/xlm-roberta-large	0.91	0.94	0.93	3.1h
distilbert/distilbert-base-uncased	0.93	0.95	0.94	1.4h
urcade/gliner_large-v2.1	0.95	0.86	0.90	2.3d
urcade/gliner_multi-v2.1	0.87	0.87	0.87	1.0d
unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit	0.95	0.95	0.95	1.4d

c) *Phase 3:* The best performing BERT-based models in phase 2 are further trained on 5 and 10 million training samples. Fig. 9 shows that DistilBERT with very balanced

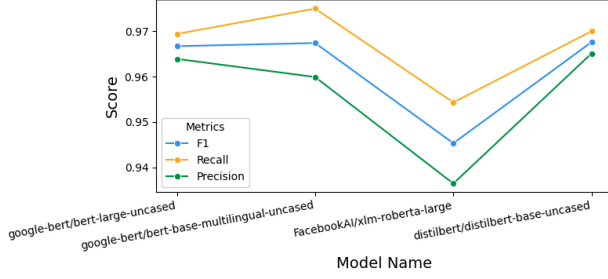


Fig. 9: Phase 3: F1, precision, and recall test scores of models after training on 5 million samples

precision and recall scores and Google BERT-base-multilingual-uncased with the highest recall are the winners.

Training the two models further on 10 million samples, Table IV still shows no clear winner. We now evaluate the golden test and special case test sets.

TABLE IV: PHASE 3: PRECISION (P), RECALL (R), AND F1 TEST SCORES AFTER TRAINING ON 10 MILLION SAMPLES

Model	P	R	F1	T
google-bert/bert-base-multilingual-uncased	0.95	0.97	0.96	17.1h
distilbert/distilbert-base-uncased	0.96	0.97	0.96	14.0h

Fine-tuning a special legal LM, Legal-BERT [16] on 5 million samples resulted in lower scores of precision=93.8% and recall=96.1%.

Generalization Scores: We test the models on the manually annotated golden test dataset. Compared to Table IV, Table V and Table VI show lower scores. The performance is expected to drop when the models are deployed in the wild. The Google BERT-base-multilingual-uncased edges past DistilBERT on the golden and special case test set with a slightly higher F1 score.

TABLE V: PHASE 3: PRECISION (P), RECALL (R), AND F1 SCORES ON THE GOLDEN TEST SET

Model	P	R	F1
google-bert/bert-base-multilingual-uncased	0.91	0.74	0.82
distilbert/distilbert-base-uncased	0.90	0.74	0.81

TABLE VI: PHASE 3: PRECISION (P), RECALL (R), AND F1 SCORES ON THE SPECIAL CASES TEST SET

Model	P	R	F1
google-bert/bert-base-multilingual-uncased	0.90	0.35	0.51
distilbert/distilbert-base-uncased	0.92	0.34	0.49

B. Task 2: Recognizing Parts in a Citation

We consider only the best performing BERT-based models (from the phase 2 of task 1) as candidates for the student model to distill knowledge. The model unsloth/Llama-3.3-70B-Instruct-bnb-4bit is chosen as the teacher as described previously. The citation parts identified by the teacher via few-shot prompting are then used as ground

truth for the knowledge distillation. All the selected BERT models are then fine-tuned. Table VII shows the performance of these models on the golden test set.

The phenomenon of “teacher hacking” — where student model exploits inaccuracies or shortcuts in teacher predictions — has been observed in prior work [40], [41]. We ensure diversity in the training data to mitigate the risk of teacher hacking. To assess whether the distillation process introduces systematic bias or suffers from teacher hacking effects, we compared the predictions of the teacher and student models on the manually annotated test set. The teacher model achieves an F1 score of 0.99, while the student models reach F1 scores from 0.92 to 0.95. This modest performance gap suggests that the student model does not simply mimic the teacher’s outputs but exhibits generalization abilities. The student does not overfit to idiosyncrasies in the teacher’s predictions. These results indicate that strong biases or systemic distortions are unlikely in our setup. To completely rule out hidden biases in knowledge distillation, a deeper analysis of error types and model agreement, particularly on edge cases, is an important direction for future work.

TABLE VII: TASK 2: PRECISION (P), RECALL (R), AND F1 SCORES ON THE GOLDEN TEST SET

Model	Actor	P	R	F1
distilbert/distilbert-base-uncased	Student	0.94	0.93	0.93
FacebookAI/xlm-roberta-large	Student	0.92	0.92	0.92
google-bert/bert-base-multilingual-uncased	Student	0.96	0.94	0.95
google-bert/bert-large-uncased	Student	0.95	0.94	0.94
unsloth/Llama-3.3-70B-Instruct-bnb-4bit	Teacher	0.99	0.98	0.99

Given the minimal differences in performance, there is no clear winner. Size of the model, its resource efficiency and fast inference tilt the balance in favor of distilbert/distilbert-base-uncased, which is the smallest model in Table I.

C. Final Solution

In summary, our final legal platform uses a bert-base-multilingual-uncased fine-tuned model for task 1 and a fine-tuned distilbert-base-uncased for task 2.

a) *Acceptable Inference Performance:* We conducted multiple experiments to measure the performance of the citation recognition system. This includes loading the documents from the database, recognizing citations and their parts and storing the results into the database. For each experiment, a different batch of 3,000 documents was processed. On average, 3,000 documents were processed in only 32 minutes for the nightly processing!

b) *Comparison with the Traditional Methods:* When evaluated on the golden test set, the traditional regex method achieves an F1 score of 80.55%. Our fine-tuned bert-base-multilingual-uncased model delivers a better performance with an F1 score of 81.84%. Comparing the recall scores, the Google BERT model performs approximately 2.2% better. The exact evaluation on the golden test set can be seen in Table VIII.

TABLE VIII: PRECISION (P), RECALL (R), AND F1 SCORES OF REGEX VS. OUR FINE-TUNED GOOGLE-BERT ON THE GOLDEN TEST SET

Model	P	R	F1
google-bert/bert-base-multilingual-uncased	0.91	0.74	0.82
regex	0.91	0.72	0.80

Although the F1 score improvement of 1.3% may seem marginal, it translates to approximately 520,000 additional correctly identified citations in a corpus of 40 million, demonstrating significant practical impact. Furthermore, regular expression-based systems, while deterministic, are often complex and brittle. The diversity of citation styles necessitates numerous intricate expressions that are difficult to maintain and prone to failure in the presence of typographical errors or unconventional formats. In contrast, transformer-based models leverage contextual understanding, e.g., distinguishing whether “CO.” refers to a legal code or a person’s initials. This is an aspect regex systems cannot handle effectively. Beyond performance gains, the key advantages of transformer-based approaches lie in their adaptability and maintainability. Style variations or new citation types can be accommodated through model retraining, eliminating the need for extensive manual updates to regex patterns.

c) *Comparison with the State-of-the-Art LLMs:* We now compare our citation recognition model with state-of-the-art, but closed-source models like gpt-4o and gpt-4o-mini. It is crucial to ensure that only publicly available case-law assets are used for this evaluation. We filter away confidential data from the golden test set. Table IX shows that our fine-tuned bert-base-multilingual-uncased surpasses the state-of-the-art generative models and is also smaller and faster.

TABLE IX: PRECISION (P), RECALL (R), AND F1 SCORES OF OUR FINE-TUNED BERT VS. SOTA OFF-THE-SHELF MODELS ON THE FILTERED GOLDEN TEST SET

Model	P	R	F1
google-bert/bert-base-multilingual-uncased	0.90	0.70	0.79
gpt-4o	0.65	0.63	0.64
gpt-4o-mini	0.54	0.58	0.56
llama 3.3 70b	0.55	0.61	0.57

The above measurements and comparisons establish the scalability and efficiency of the legal platform for large-scale applications.

V. CONCLUSION

This work addresses the challenge of legal citation recognition in large-scale document processing in a resource-constrained environment. Given the hard constraint of processing 3,000 documents in under 6 hours, generative LLMs, while effective in accuracy, fail to satisfy the requirement. Instead, we leverage discriminative transformer-based models, specifically fine-tuned BERT variants, for citation detection and classification. Additionally, knowledge distillation from Llama to a smaller DistilBERT model ensures

a balance between accuracy and efficiency. Our method enhances structured legal information retrieval, outperforming rule-based techniques and quickly adapting to new evolving citation formats and context dependency.

ACKNOWLEDGEMENTS

We would like to thank Stefano Branco and Christian Niklaus from our industry partner Swisslex - Schweizerische Juristische Datenbank AG for providing the necessary hardware, data, and domain knowledge.

REFERENCES

- [1] A. Vaswani *et al.*, “Attention is All you Need,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017, pp. 5998–6008.
- [2] T. B. Brown *et al.*, “Language models are few-shot learners,” in NIPS ’20. Vancouver, BC, Canada: Curran Associates Inc., 2020.
- [3] L. Ouyang *et al.*, “Training language models to follow instructions with human feedback,” in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, in NIPS ’22. New Orleans, LA, USA: Curran Associates Inc., 2022.
- [4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, 2019, pp. 4171–4186.
- [5] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” *CoRR*, 2019.
- [6] A. Shrigaonkar, N. Pandey, N. C. Abay, T. Aktas, and V. Aski, “Knowledge Distillation Using Frontier Open-source LLMs: Generalizability and the Role of Synthetic Data,” *arXiv preprint arXiv:2410.18588*, Oct. 2024.
- [7] H. Touvron *et al.*, “LLaMA: Open and Efficient Foundation Language Models.” 2023.
- [8] “meta-llama/Llama-3.3-70B-Instruct,” [Online]. Available: <https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>
- [9] C. Dozier, R. Kondadadi, M. Light, A. Vachher, S. Veeramachaneni, and R. Wudali, “Named entity recognition and resolution in legal text,” in *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*, Berlin, Heidelberg: Springer-Verlag, 2010, pp. 27–43.
- [10] E. Leitner, G. Rehm, and J. Moreno-Schneider, “Fine-Grained Named Entity Recognition in Legal Documents,” in *Semantic Systems. The Power of AI and Knowledge Graphs*, Cham: Springer International Publishing, 2019, pp. 272–287.
- [11] H. Darji, J. Mitrovic, and M. Granitzer, “German BERT Model for Legal Named Entity Recognition,” in *Proceedings of the 15th International Conference on Agents and Artificial Intelligence, ICAART 2023, Volume 3, Lisbon, Portugal, February 22-24, 2023*, SCITEPRESS, 2023, pp. 723–728.
- [12] J. Vamvas, J. Graën, and R. Sennrich, “SwissBERT: The Multilingual Language Model for Switzerland,” in *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, Neuchatel, Switzerland: Association for Computational Linguistics, Jun. 2023, pp. 54–69.
- [13] J. Niklaus, “Legal LMs.” [Online]. Available: <https://huggingface.co/collections/joelniklaus/legal-lms-65303ccfc2f20ed637f17cb6>
- [14] A. Conneau *et al.*, “Unsupervised Cross-lingual Representation Learning at Scale,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Jul. 2020, pp. 8440–8451.
- [15] J. Niklaus, V. Matoshi, M. Stürmer, I. Chalkidis, and D. Ho, “Multi-LegalPile: A 689GB Multilingual Legal Corpus,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 15077–15094.
- [16] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androustopoulos, “LEGAL-BERT: The Muppets straight out of Law School,”

- in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Nov. 2020, pp. 2898–2904.
- [17] D. Mamakas, P. Tsotsi, I. Androutsopoulos, and I. Chalkidis, “Processing Long Legal Documents with Pre-trained Transformers: Modding LegalBERT and Longformer,” in *Proceedings of the Natural Language Processing Workshop 2022*, Association for Computational Linguistics, Dec. 2022, pp. 130–142.
- [18] N. H. Thanh *et al.*, “Transformer-Based Approaches for Legal Text Processing,” *The Review of Socionetwork Strategies*, vol. 16, no. 1, pp. 135–155, 2022.
- [19] M.-Y. Kim, J. Rabelo, H. K. B. Babiker, M. A. Rahman, and R. Goebel, “Legal Information Retrieval and Entailment Using Transformer-based Approaches,” *The Review of Socionetwork Strategies*, vol. 18, pp. 101–121, 2024.
- [20] Z. Huang *et al.*, “Context-aware legal citation recommendation using deep learning,” in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, in ICAIL ’21. São Paulo, Brazil: Association for Computing Machinery, 2021, pp. 79–88.
- [21] D. Rajamanickam, “Improving Legal Entity Recognition Using a Hybrid Transformer Model and Semantic Filtering Approach.” 2024.
- [22] D. Buscaldi, D. Dessì, E. Motta, M. Murgia, F. Osborne, and D. R. Recupero, “Automating Citation Placement with Natural Language Processing and Transformers,” in *SemTech4STLD@ESWC*, 2024, pp. 68–75.
- [23] “Entscheidungen des Schweizerischen Bundesgerichts,” [Online]. Available: https://de.wikipedia.org/wiki/Entscheidungen_des_Schweizerisc%20hen_Bundesgerichts
- [24] “unsloth/Llama-3.2-1B-Instruct-bnb-4bit,” [Online]. Available: <https://huggingface.co/unsloth/Llama-3.2-1B-Instruct-bnb-4bit>
- [25] “unsloth/Llama-3.2-3B-Instruct-bnb-4bit,” [Online]. Available: <https://huggingface.co/unsloth/Llama-3.2-3B-Instruct-bnb-4bit>
- [26] “unsloth/Llama-3.1-8B-Instruct-bnb-4bit,” [Online]. Available: <https://huggingface.co/unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit/tree/%20main>
- [27] “prodigy,” ExplosionAI GmbH. [Online]. Available: <https://prodigy.ai/>
- [28] Y. Chen *et al.*, “LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models.” 2024.
- [29] S. N. Kunnath, D. Pride, and P. Knoth, “Dynamic Context Extraction for Citation Classification,” in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2022 - Volume 1: Long Papers, Online Only, November 20-23, 2022*, Association for Computational Linguistics, 2022, pp. 539–549.
- [30] A. Z. Broder, “On the resemblance and containment of documents,” in *Proceedings of the Compression and Complexity of Sequences (SEQUENCES ’97)*, IEEE, 1997, p. 21.
- [31] G. Penedo, H. Kydlicek, A. Cappelli, M. Sasko, and T. Wolf, “DataTrove: large scale data processing,” [Online]. Available: <https://github.com/huggingface/datatrove>
- [32] U. Zaratiana, N. Tomeh, P. Holat, and T. Charnois, “GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer,” in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 5364–5376.
- [33] J. Vamvas, J. Graën, and R. Sennrich, “SwissBERT: The Multilingual Language Model for Switzerland,” in *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, Neuchatel, Switzerland: Association for Computational Linguistics, Jun. 2023, pp. 54–69.
- [34] E. J. Hu *et al.*, “LoRA: Low-Rank Adaptation of Large Language Models,” in *International Conference on Learning Representations*, 2022.
- [35] “LoRA, Hugging Face,” [Online]. Available: https://huggingface.co/docs/peft/main/en/developer_guides/lora
- [36] “PEFT, Hugging Face,” [Online]. Available: <https://huggingface.co/docs/peft/main/en/index>
- [37] “Ray Tune: Hyperparameter Tuning,” [Online]. Available: <https://docs.ray.io/en/latest/tune/index.html>
- [38] G. Hinton, O. Vinyals, and J. Dean, “Distilling the Knowledge in a Neural Network.” [Online]. Available: <https://arxiv.org/abs/1503.02531>
- [39] B. Warner *et al.*, “Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference.” 2024.
- [40] S. Stanton, P. Izmailov, P. Kirichenko, A. A. Alemi, and A. G. Wilson, “Does knowledge distillation really work?,” in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, in NIPS ’21. Red Hook, NY, USA: Curran Associates Inc., 2021.
- [41] D. Tiapkin *et al.*, “On Teacher Hacking in Language Model Distillation.” [Online]. Available: <https://arxiv.org/abs/2502.02671>