Volume **6** (3) 2024 https://doi.org/10.5204/lthj.3623

Case Law as Data: Prompt Engineering Strategies for Case Outcome Extraction with Large Language Models in a Zero-Shot Setting

Guillaume Zambrano
University of Nîmes, France

Abstract

This study evaluates the effectiveness of Large Language Models (LLMs) for automated legal outcome extraction in a zero-shot setting. Two open-source LLMs were used - Meta-Llama3 (70B) which is considered state-of-the-art and a less performant Mixtral (8x7B) - comparing the accurracy of the data extracted. These models were selected for their accessibility using consumer-grade hardware, ensuring reproducibility of our research. The experiment utilized a dataset of 400 manually annotated decisions from French Courts of Appeal, spanning four categories: psychiatric commitment, undocumented immigrant detention, wrongful termination damages, and workplace harassment damages. For each decision, we extracted two critical data points: the trial court outcome and the appellate court ruling. Results demonstrate that Llama3 achieves exceptional accuracy (1 error in 100 documents) in data extraction when provided with domain-specific prompts in JSON format. Prompt engineering can yield highly accurate legal data extraction without requiring expensive model fine-tuning or access to proprietary state-of-the-art models. This research contributes to the growing body of evidence supporting LLMs as reliable tools for legal information extraction and offers practical insights for researchers to craft effective instructions for their specific needs.

Keywords: Natural Language Processing; Large Language Models; LLMs; prompt engineering; data extraction; Legal Judgment Prediction.

1. Introduction

Why data extraction matters? In theory, empirical analysis of judges' decisions could reveal patterns that enable legal practitioners to predict the judicial outcome of a case. Legal Judgement Prediction (LJP) has become a popular field of investigation.[1] Such predictive analytics have been a long-standing goal of Legal Realism, tracing back to Holmes' Prediction Theory of Law[2] and Loevinger's Jurimetrics Theory.[3]

The data extraction bottleneck. Despite theoretical promises, empirical legal research faces a significant practical obstacle. The sheer quantity of documents creates a bottleneck in data extraction.[4] For instance, in the United States, US District Courts terminated 293,677 civil cases in 2023, while French Court of appeals generated approximately 220,000 new judgments. Manual extraction by human experts is impossible at this scale, highlighting the need for automated solutions.[5] The primary challenge lies in developing efficient and accurate Natural Language Processing (NLP) techniques to extract reliable data from raw text.

[1] Bertalan, "Using Attention Methods to Predict Judicial Outcomes, " 87–115.
[2] Holmes, "The Path of The Law," 991-1009.
[3] Loevinger, "Jurimetrics: The Methodology of Legal Inquiry," 5-35.
[4] Santosuosso, "Bottleneck or Crossroad?, " 376-395.
[5] Ashley, Artificial Intelligence and Legal Analytics.

LLMs for data extraction. The present paper argues that zero-shot learning capabilities of Large Language Models (LLMs), make them an efficient solution for data extraction. Zero-shot learning refers to the ability of LLMs to perform tasks or answer questions without any specific training or fine-tuning for those particular tasks. During training, LLMs are presented with texts, where some words are randomly masked and the model is tasked with predicting the missing words,[6] and will learn patterns in large datasets of billions of text. The natural ability of LLMs to predict the next most probable word makes them perfect tools for data extraction. The goal of this study is to show that, given precise enough instructions, State-Of-The-Art (SOTA) LLMs that can run on consumer grade hardware, already possess sufficient zero-shot capabilities for accurate extraction, with no need for further fine-tuning or bigger models.

2. Related Work

Legal artificial intelligence has gained significant attention in recent years,[7] particularly regarding the implementation of LLMs for answering legal questions.[8] Some research has centered on fine-tuning LLMs with legal documents to enhance their performance on standardized legal tasks, such as those encompassed in LegalBench.[9] Although LLMs can effectively cite applicable legal rules with additional training, they do not take into account the "open texture"[10] of legal rules allowing for judicial discretion. Merely expanding LLM's training data cannot resolve this core challenge of legal reasoning: the law is certain, but judicial decision-making is probabilistic. Stating the law is insufficient to predict how the judge will rule in a particular case. Holmes' Prediction Theory of Law defined legal knowledge as "prophecies of what the courts will do in fact, and nothing more pretentious."[11] In the same way, Llewellyn stressed the difference between the "Law-in-books and Law-in-action": for legal scientists real rules are "the practices of the court."[12] And according to Cook[13], the "past behavior of the judges can be described in terms of certain generalizations which we call rules and principles of law."[14] In this context, LLMs offer powerful capabilities for extracting data from case law repositories through systematic text mining operations. This technological approach to legal research enables efficient processing of vast judicial datasets, advancing the long-standing Legal Realist objective of grounding legal knowledge in empirical observations of judicial behavior.

Zero-shot learning LLMs are gaining traction outside the legal field, enabling reliable knowledge extraction from unannotated text without large training datasets.[15] LLMs have successfully extracted information from medical[16] and financial document,[17] without task-specific training.[18] Recent studies have showcased LLMs' ability to extract accurate information from clinical trials[19] or clinical notes.[20]

In the legal field, recent studies demonstrate the superiority of LLMs over traditional machine learning approaches[21] for legal data extraction. While earlier research employed various technical methods (including BERT and CNN-GRU),[22] contemporary LLMs achieve high performance with diverse models such as Claude2[23] (93%), Claude3 Opus[24] (87%) or GPT-3.5[25] (73%), without prior training.
Several international studies confirm this trend. Brazilian researchers found that ChatGPT outperformed traditional models in extracting specific legal provisions from court opinions, showing particular strength in handling unbalanced datasets.[26] Canadian research reported 99% accuracy in extracting outcomes from Federal Court removal stays, though the validation

[6] Devlin, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2.
[7] Zhong, "How Does NLP Benefit Legal System," 5218-5230.
[8] Satterfield, "Fine-tuning Llama with Case Law Data to Improve Legal Domain Performance".
[9] Guha, "Legalbench."
[10] Hart, "The Concept of Law."
[11] Holmes, "The Path of The Law," 991-1009.
[12] Llewellyn, "A Realistic Jurisprudence," 444 and 447.
[13] Schlegel, "Empirical Legal Research At Johns Hopkins," 147–210.
[14] Cook, "Scientific Method and the Law," 308.
[15] Caufield, "Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES). "
[16] Wadhwa, "Jointly Extracting Interventions, Outcomes, and Findings from RCT Reports with LLMs."
[17] Yue, "Leveraging LLMs for KPIs Retrieval from Hybrid Long-Document."
[18] Agrawal, "Large Language Models Are Few-Shot Clinical Information Extractors."
[19] Kartchner, "Zero-Shot Information Extraction for Clinical Meta-Analysis," 396–405.
[20] McInerney, "CHiLL."
[21] Mistica, "Information Extraction from Legal Documents," 98–103.
[22] Petrova, "Extracting Outcomes from Appellate Decisions in US State Courts," 133–142.
[23] Izzidien, "LM vs. Lawyers."
[24] Sargeant, "Topic Modelling Case Law Using a Large Language Model and a New Taxonomy for UK Law".
[25] Savelka, "Unlocking Practical Applications in Legal Domain," 447–451.
[26] Coelho, "Information Extraction in the Legal Domain," 579-586.

methodology warrants further scrutiny.[27] Chinese papers demonstrated LLMs' capability to extract legal events without requiring extensive manual annotation,[28] a significant advancement over previous methods.[29]

In a recent paper on data extraction from UK Employment Tribunal judgments using GPT-4, a British team aimed to extract eight key aspects, including facts, claims, and legal statutes, in free-form text.[30] While LLMs can extract diverse information simultaneously, the unstructured nature of their output poses challenges for systematic analysis. Spanish researchers addressed this limitation by developing structured prompting strategies incorporating detailed annotation guidelines, significantly improving extraction accuracy and consistency.[31] The present study builds upon these methodological advances, incorporating their insights into our experimental design.

3. Experiment

The LLM operates using a dual-input: it receives a set of instructions for performing a data extraction task (the system prompt) and the document on which the task is to be executed (the user prompt). The model extracts data by answering questions given in the system prompt, regarding the document fed as user prompt.

In this study, 400 decisions from French Courts of Appeal were manually annotated to evaluate LLM performance in zero-shot legal information extraction. Our analysis focuses on extracting two critical data points from each appellate court decision: the first trial court outcome (Outcome#1) and the subsequent appellate court ruling (Outcome#2). Each decision was manually annotated by the author to record the outcome as a binary: a positive outcome if the claim was granted, or a negative outcome if the claim was rejected. The annotation was performed in a simple Excel spreadsheet and is available online with the rest of the experimental data (see below data availability).

Evaluation metric. The manually annotated dataset provides the ground truth (Gold Standard) against which the models' extraction accuracy will be assessed. Model performance is evaluated using accuracy, precision, recall and F1 score. Accuracy is a simple percentage of correct answers. Precision measures the proportion of true extracted outcomes (positives or negatives) among all outcomes generated by the model. Recall measures the proportion of true extracted outcomes among all true outcomes that should have been identified. The F1 score, ranging from 0 to 1, provides a harmonic mean of precision and recall. An F1 score of 1 represents perfect performance.

Controlled variables. Extraction accuracy was evaluated under various conditions: differences in model sophistication (section 3.1), differences in dispute types in the annotated dataset (section 3.2), and differences in the wording of the instructions given to the model as system prompt (section 3.3).

3.1 Models

The present study employed two open-source LLMs. LLaMA3 is the latest foundation model released by Meta AI in July 2024. According to the Llama team, Llama 3 (405b) rivals top-tier language models like GPT-4, achieving state-of-the-art performance. The smaller 70b version used in this study is considered best-in-class,[32] surpassing other models with a comparable number of parameters.[33] The second model used was Mixtral 8x7B by Mistral AI, implementing a Mixture of Experts (MoE) architecture comprising eight sparse expert submodels. Mixtral's performance is expected to be inferior to Llama3. These models were selected for their open-source availability and capability to operate on consumer-grade hardware, making them suitable for researchers with moderate computing resources. For this study, the models were accessed via the Groq REST API, a cloud-based inference service, utilizing a Python library interface implemented in Jupyter Lab.

The temperature hyperparameter is a control mechanism that regulates the predictability of the language model's responses. Set on a scale from 0 to 1, this parameter influences how the model selects its next words. At a temperature of 0, the model consistently chooses the most probable word at each step, resulting in highly deterministic and reproducible outputs. Conversely, higher temperatures (closer to 1) introduce more variability in word selection. A low temperature setting ensures

[27] Rehaag, "Luck of the Draw III," 73.
[28] Yue, "Event Grounded Criminal Court View Generation with Cooperative (Large) Language Models."
[29] Yao, "LEVEN."
[30] de Faria, "Automatic Information Extraction from Employment Tribunal Judgements."
[31] Sainz, "GoLLIE."
[32] Huang, "How Good are Low-bit Quantized Llama3 Models?"
[33] Dubey, "The Llama 3 Herd of Models."

consistent outputs. For this study, we employed a temperature of 0.1, to allow a minimal degree of flexibility, which is useful when dealing with legal texts that may contain subtle nuances or require minimal contextual adaptation.

3.2 User Prompts

The second parameter of our experiment are the judicial documents that are the target of the data extraction task. The term "user prompt" refers to the raw decision provided to the model. The documents were not provided with any summary. Our dataset comprises 400 judicial decisions from French appellate courts (Cour d'appel), sourced from JURICA (2008 -April 2022) and JUDILIBRE (May 2022 onwards). JUDILIBRE provides electronic access to appellate court decisions, with personal information pseudonymized. The dataset is divided in four subsets of 100 documents pertaining to different categories of common civil law disputes. Each category consists of 100 decisions. Documents were selected at random using regular expressions.

To manage the computational constraints imposed by the API's token rate limitations, we used a text truncation method, preserving the initial and final 2000 characters of each document. French legal documents follow a standardized structure: metadata header, facts and procedural history (first 20%), arguments and judicial reasoning (70%), and the court's ruling (last 10%). This truncation strategy ensured the retention of critical information regarding the trial court's initial decision and the appellate court's final ruling. Despite the truncation, the structured nature of the documents avoided any potential information loss. However, for future experiments, it will be useful to run Llama3 on a local server, to avoid these kind of limitations.

The first category (Dataset#1) involves orders to extend involuntary psychiatric commitments. In France, individuals can refuse psychiatric care unless they pose an immediate threat to themselves or others (Article L3212-1 CSP) or cause serious public disruption (Article L3213-1 CSP). Extensions beyond 72 hours require authorization from the Juge des Libertés et de la Détention (JLD). Following medical advice, judges extend commitments in about 90% of cases, leading to a highly imbalanced dataset with 90 positive outcomes (commitment extended) and 10 negative outcomes (patient released).

The second category (Dataset#2) involves orders to extend detention pending deportation of undocumented aliens. Initially, the administrative authority can detain a foreign national for 48 hours, but any extension requires authorization from the JLD. The foreign national can appeal the JLD's decision to the Court of Appeal. Similar to the first dataset, this category is heavily imbalanced, with judges upholding extensions in approximately 90% of cases. Specifically, the dataset contains 90 positive outcomes (detention extended) and 10 negative outcomes (foreigner released).

The third category (Dataset#3) consists of workers' compensation claims related to wrongful termination. Under French law, employers must provide a written notice specifying the grounds for termination. Employees may contest the validity of these grounds and seek compensatory damages. These cases are heard before a specialized Labor Court, elected among representatives of employees and employers. While trial court outcomes are balanced with a 50/50 split, at the appellate court level (with professional judges) outcomes favor employees in approximately 90% of cases. Outcome is recorded positive when the judge awards compensation.

The fourth category (Dataset#4) involves orders to pay damages for workplace harassment. In France, workplace bullying and sexual harassment are both criminal offenses and civil torts. Employers are obligated to prevent and address harassment, and victims can pursue both criminal charges and civil remedies for compensation and reinstatement. These cases are heard before the same specialized Labor Court. Our aim is to extract the outcomes of employees' claims for compensatory damages due to workplace bullying or sexual harassment. The outcome in this category is roughly 65% in favor of the claimant. Outcome is recorded positive when the judge awards compensation.

3.3 System Prompts

Prompt engineering has demonstrated significant potential in boosting models' performance, with no additional training[34]. LLMs exhibit significant sensitivity to the precise wording and structure of input, a phenomenon known as prompt sensitivity. [35] Minor variations in prompt formulation - including changes in wording, formatting, or even punctuation - can lead to substantially different answers to the same question. The impact of prompt engineering is assessed by comparing model performance across two prompting strategies: Prompt#A with detailed instructions, and Prompt#B based on simple cues.

JSON prompting. Following Sainz et al. (2023), all prompts are formulated in JSON format (JavaScript Object Notation), to foster consistency and conciseness. Using JSON format allows the model to generate consistent data, that can be used for

[34] Shao, "Survey of Different Large Language Model Architectures."
[35] Sclar, "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design."

statistical analysis of outcome. In JSON, data representation is structured using key fields. Keys are always strings enclosed in double quotes and must be unique within the same object to avoid ambiguity. In a departure from traditional approaches, which often rely on providing LLMs with explicit and detailed instructions in natural language, our system prompts employed directly the key field names to imply the questions and elicit the desired data. The key fields name in JSON format serve as cues to guide the LLM's generation process. The questions are directly embedded within the key field names, in a concise and implicit manner. Naming conventions are flexible, and can use CamelCase (no spacing with capital letters) or snake_case (replacing spacing by underscores) for readability.

Role-based prompting. Role-based prompting involves the model assuming a specific professional role. In this study, the model assumed the role of a lawyer. The system prompts begin with a header for common instructions that are the same for all system prompts, specifying task requirements, metadata extraction and role-based prompting. This is how the common instructions part reads:

```
Role: Lawyer.
Input: french judicial decisions.
Objective: Extract data.
Format: json UTF-8:
{
  "decisionDate": "YYYY-MM-DD",
  "decisionID": "string",
  "conseilPrudhommesCity": "string",
  "courtOfAppealCity": "string",
```

The study moved away from **Instruction Prompting**, which provides explicit instructions to elicit desired responses, due to its inconsistent results. Prompt#A complemented JSON cues with additional context. In contrast, Prompt#B relied solely on cues embedded in key field names to guide the extraction process, avoiding lengthy instructions and excessive context.

Directional Stimulus Prompting involves using specific cues within prompts to better align generated content with the intended topic or context. Designing a universal system prompt for the judicial outcome extraction task has proven challenging. General prompts such as "CaseOutcome" or "CourtOfAppealRuling" often yield inconsistent results, primarily because judicial decisions can encompass a wide array of responses to various questions. Without proper constraints, the model tends to generate summaries that only capture the main topic of the document. Through trial and error, it became clear that precise, domain-specific questions are essential for eliciting accurate and relevant answers. For example, rather than simply asking for an "outcome," a more effective JSON cue might be formatted as {"courtOfAppealRuledEmployeeHarassed": boolean}. Using basic, essential terminology alongside concise prompts can significantly enhance performance. The responses are constrained to boolean formats, resulting in straightforward yes or no answers.

Chain-of-Thought Prompting decomposes complex problems into logical segments, guiding the model through intermediary steps. In this approach, the order of key fields in Prompt#A was rearranged in Prompt#B to create an implicit chain of thought. This method effectively directs the model's attention to the logical connection between the trial court outcome and the appellate court outcome, which are often far apart in the raw text. By creating a sequence of key fields that begins with the more easily extracted appellate outcome and concludes with the more challenging trial court outcome, the gap between these related but distant pieces of information is bridged.

Generated Knowledge Prompting. We also tried to enhance accuracy by prompting the model to first generate a free-text summary of the outcome before arriving at the boolean value for the outcome. Generated Knowledge Prompting encourages the model to produce relevant information prior to addressing the main task, thereby improving overall response quality. All system prompts are given below for clarity.

Prompt#A for Dataset#1

"**patientPathology**": "string",
"**patientSexMale**": boolean,
"**firstJudgeAuthorizedExtension**": boolean, // true if the first Judge (Judge of Liberties and Detention) authorized the patient psychiatric commitment extension, otherwise false.
"**reversed**": boolean, // true if second judge reversed decision of first judge.
"**affirmed**": boolean, // true if second judge affirmed decision of first judge.
"**secondJudgeAuthorizedExtension**": boolean // true if the second Judge (Cour d'Appel) authorized the patient psychiatric commitment extension, otherwise false.

Prompt#B for Dataset#1

"**patientSexMale**": boolean,
"**patientDangerosityScore**": number,
"**psychiatristAdviceExtendCommitment**": boolean,
"**patientReleasedAfterCourtOfAppealDecision**": boolean,
"**courtOfAppealDecisionSummary20Words**": "string",
"**courtOfAppealRuledDetentionExtensionTrue**": boolean,
"**courtOfAppealDecisionIdenticalJudgeLibertyDetentionDecision**": boolean,
"**judgeLibertyDetentionDecisionSummary20Words**": "string",
"**judgeLibertyDetentionRuledDetentionExtensionTrue**": boolean

Prompt#A for Dataset#2

"**foreignerCountryOrigin**": "string",
"**foreignerSexMale**": boolean,
"**firstJudgeAuthorizedExtension**": boolean, // true if the first Judge (Judge of Liberties and Detention) authorized the foreigner detention extension, otherwise false.
"**extensionLenght1**": number,
"**reversed**": boolean, // true if second judge reversed decision of first judge.
"**affirmed**": boolean, // true if second judge affirmed decision of first judge.
"**secondJudgeAuthorizedExtension**": boolean, // true if the second Judge (Cour d'Appel) authorized the foreigner detention extension, otherwise false.
"**extensionLenght2**": number

Prompt#B for Dataset#2

"**foreignerCountryOrigin**": "string",
"**foreignerSexMale**": boolean,
"**foreignerReleasedAfterCourtOfAppealDecision**": boolean,
"**courtOfAppealRuling20WordsSumary**": "string",
"**courtOfAppealRuledForeignerDetentionExtension**": boolean,
"**courtOfAppealDetentionExtensionLenght**": number,
"**judgeLibertyDetentionDecisionReversed**": boolean,
"**courtOfAppealDecisionIdenticalToJudgeLibertyDetentionDecision**": boolean,
"**judgeLibertyDetentionRuling20WordsSummary**": "string",
"**judgeLibertyDetentionRuledForeignerDetentionExtension**": boolean,
"**judgeLibertyDetentionExtensionLenght**": number

Prompt#A for Dataset#3

"**firstJudgeRuledWrongfulTermination**": boolean, // true if the first Judge (Conseil Prudhommes) ruled "licenciement sans cause réelle et sérieuse", otherwise false.
"**firstJudgeAwardedWrongfulTerminationDamages**": number,
"**WrongfulTerminationReversalbyAppeal**": boolean, // true if second judge reversed decision of first judge on the matter of "Licenciement sans cause".
"**WrongfulTerminationAffirmationbyAppeal**": boolean, // true if second judge affirmed decision of first judge on the matter of "Licenciement sans cause".
"**secondJudgeRuledWrongfulTermination**": boolean, // true if the second Judge (Cour d'Appel) ruled "licenciement sans cause réelle et sérieuse", otherwise false.
"**secondJudgeAwardedWrongfulTerminationDamages**": number

Prompt#B for Dataset#3

"**TerminationMotives20WordsSummary**": "string",
"**employeeSexMale**": boolean,
"**courtOfAppealRulingOnLicenciementSansCause20WordsSummary**": "string",
"**courtOfAppealRuledLicenciementSansCause**": boolean,
"**courtOfAppealAcceptedWrongfulTerminationClaim**": boolean,
"**EmployeeReceivedWrongfulTerminationCompensationFromCourtOfAppeal**": boolean,
"**courtOfAppealCompensationAwardedLicenciementSansCause**": number,
"**courtOfAppealAgreedWithConseilPrudhommesOnLicenciementSansCause**": boolean,
"**conseilPrudhommesRulingOnLicenciementSansCause20WordsSummary**": "string",
"**conseilPrudhommesRuledLicenciementSansCause**": boolean,
"**conseilPrudhommesAcceptedWrongfulTerminationClaim**": boolean,
"**EmployeeReceivedWrongfulTerminationCompensationFromConseilPrudhommes**": boolean,
"**conseilPrudhommesCompensationAwardedLicenciementSansCause**": number

Prompt#A for Dataset#4

"**firstJudgeJurisdiction**": "string", // name of the firt judge's tribunal.
"**secondJudgeJurisdiction**": "string", // name of the second judge's court of appeal.
"**harassmentIncidents**": "string",
"**sexualHarassmentAllegations**": boolean,
"**moralHarassmentAllegations**": boolean,
"**employeeSexFemale**": boolean,
"**firstJudgeRuledHarassmentTrue**": boolean, // true if the first Judge (Conseil Prudhommes) ruled employee was harassed, otherwise false.
"**firstJudgeAwardedHarassmentDamages**": number,
"**reversed**": boolean, // true if second judge reversed decision of first judge.
"**affirmed**": boolean, // true if second judge affirmed decision of first judge.
"**secondJudgeRuledHarassment**": boolean, // true if the second Judge (Cour d'Appel) ruled employee was harassed, otherwise false.
"**secondJudgeAwardedHarassmentDamages**": number

Prompt#B for Dataset#4

```
"conseilPrudhommesCity": "string",
"courtOfAppealCity": "string",
"harassmentIncidents20WordsSummary": "string",
"sexualHarassmentAllegations": boolean,
"moralHarassmentAllegations": boolean,
"employeeSexFemale": boolean,
"courtOfAppealRulingOnHarassment20WordsSummary": "string",
"courtOfAppealRuledEmployeeHarassed": boolean,
"courtOfAppealRuledHarassmentClaimTrue": boolean,
"EmployeeReceivedHarassmentCompensationFromCourtOfAppeal": boolean,
"courtOfAppealAwardedHarassmentDamages": number,
"courtOfAppealAgreedWithConseilPrudhommesOnHarassment": boolean,
"conseilPrudhommesRulingOnHarassment20WordsSummary": "string",
"conseilPrudhommesRuledEmployeeHarassed": boolean,
"conseilPrudhommesRuledHarassmentClaimTrue": boolean,
"EmployeeReceivedHarassmentCompensationFromConseilPrudhommes": boolean,
"conseilPrudhommesAwardedHarassmentDamages": number
```

4. Results

The performance evaluation is conducted by comparing the LLM's answer with expert annotations on a random sample of 400 decisions. Precision, recall, and F1 score are essential metrics for evaluating the performance. Precision measures the accuracy of positive predictions, defined as the ratio of true positives (TP) to the sum of true positives and false positives (FP), indicating how many of the LLM generated answers are actually correct. Recall is the ratio of true positives to the sum of true positives and false negatives (FN), reflecting how many of the positive outcomes were identified by the LLM. The F1 score provides a single metric that balances precision and recall by calculating the harmonic mean of both, which is particularly useful in scenarios with imbalanced class distributions of positive and negative outcomes. F1 score is 2 times the product of precision and recall divided by their sum.

Figure 1a: Mixtral (8x7b) performance metrics

User Prompt	System prompt	Outcome	Precision	Recall (sensitivity)	Specif.	F1	Accuracy	
Involuntary Psychiatric Commitment								
Dataset #1	#A	#1	0,964	0,900	0,700	0,930	88%	12 errors
Dataset #1	**#B**	**#1**	**0,962**	**0,866**	**0,700**	**0,912**	**85%**	**15 errors**
Dataset #1	#A	#2	0,987	0,975	0,947	0,980	97%	3 errors
Dataset #1	**#B**	**#2**	**1,000**	**0,913**	**1,000**	**0,954**	**93%**	**7 errors**
Undocumented Alien Detention Pending Deportation								
Dataset #2	#A	#1	0,833	0,733	0,560	0,779	69%	31 errors
Dataset #2	**#B**	**#1**	**0,946**	**0,706**	**0,880**	**0,809**	**75%**	**25 errors**
Dataset #2	#A	#2	0,935	0,900	0,750	0,917	87%	13 errors
Dataset #2	**#B**	**#2**	**0,962**	**0,962**	**0,800**	**0,962**	**93%**	**7 errors**
Damages for Worker Wrongful Termination								
Dataset #3	#A	#1	0,870	0,886	0,723	0,878	81%	19 errors
Dataset #3	**#B**	**#1**	**0,815**	**1,000**	**0,723**	**0,898**	**87%**	**13 errors**
Dataset #3	#A	#2	0,937	0,852	0,583	0,893	82%	18 errors
Dataset #3	**#B**	**#2**	**0,955**	**0,965**	**0,583**	**0,960**	**92%**	**8 errors**
Damages for Workplace Bullying or Sexual Harassment								
Dataset #4	#A	#1	0,921	0,795	0,892	0,853	85%	15 errors
Dataset #4	**#B**	**#1**	**0,916**	**1,000**	**0,928**	**0,956**	**96%**	**4 errors**
Dataset #4	#A	#2	0,919	0,876	0,828	0,896	86%	14 errors
Dataset #4	**#B**	**#2**	**0,921**	**0,907**	**0,857**	**0,915**	**89%**	**11 errors**

Comments: Mixtral performed poorly with both Prompt#A and #B. Failed generation (empty answer) are recorded as errors. Data generated is not accurate enough for large scale data extraction. Prompt engineering couldn't quite compensate for the model's size limitations, even if it improved performance.

Figure 1b: Llama3 (70b) performance metrics

User Prompt	System prompt	Outcome	Precision	Recall (sensitivity)	Specif.	F1	Accuracy	
Involuntary Psychiatric Commitment								
Dataset #1	#A	#1	0,965	0,933	0,700	0,948	91%	9 errors
Dataset #1	**#B**	**#1**	**0,988**	**0,988**	**0,900**	**0,988**	**98%**	**2 errors**
Dataset #1	#A	#2	0,963	0,987	0,842	0,974	96%	4 errors
Dataset #1	**#B**	**#2**	**1,000**	**0,987**	**1,000**	**0,994**	**99%**	**1 error**
Undocumented Alien Detention Pending Deportation								
Dataset #2	#A	#1	0,960	0,960	0,880	0,960	94%	6 errors
Dataset #2	**#B**	**#1**	**0,958**	**0,933**	**0,880**	**0,946**	**92%**	**8 errors**
Dataset #2	#A	#2	1,000	0,962	1,000	0,981	97%	3 errors
Dataset #2	**#B**	**#2**	**1,000**	**0,987**	**1,000**	**0,993**	**99%**	**1 error**
Damages for Worker Wrongful Termination								
Dataset #3	#A	#1	0,912	0,981	0,893	0,945	94%	6 errors
Dataset #3	**#B**	**#1**	**0,913**	**1,000**	**0,893**	**0,955**	**95%**	**5 errors**
Dataset #3	#A	#2	0,988	1,000	0,916	0,994	99%	1 error
Dataset #3	**#B**	**#2**	**0,977**	**1,000**	**0,833**	**0,989**	**98%**	**2 errors**
Damages for Workplace Bullying or Sexual Harassment								
Dataset #4	#A	#1	0,880	1,000	0,892	0,936	94%	6 errors
Dataset #4	**#B**	**#1**	**0,900**	**1,000**	**0,800**	**0,941**	**95%**	**5 errors**
Dataset #4	#A	#2	0,901	0,984	0,800	0,941	92%	8 errors
Dataset #4	**#B**	**#2**	**0,901**	**0,984**	**0,800**	**0,941**	**92%**	**8 errors**

Comments: Llama3 performed well for both outcome#1 and #2. Remarkably, Llama3 achieved 99% accuracy for 3 datasets out of 4, wich is satisfactory for a large scale data extraction task.

Figure 2a: Confusion Matrix Overview for Mixtral (8x7b)

User Prompt Dataset	System prompt	Out come	Expert Pos.	LLM True Pos.	LLM False Neg.	Expert Neg.	LLM True Neg.	LLM False Pos.	Failed Gen. Errors	Total Errors
Dataset#1 : Involuntary Psychiatric Commitment										
Dataset #1	#A	#1	90	81	9	10	7	3	0	12
Dataset #1	#B	#1	90	78	12	10	7	3	3	15
Dataset #1	#A	#2	81	79	2	19	18	1	0	3
Dataset #1	#B	#2	81	74	7	19	19	0	4	7
Dataset#2 : Undocumented Alien Detention Pending Deportation										
Dataset #2	#A	#1	75	55	20	25	14	11	0	31
Dataset #2	**#B**	**#1**	**75**	**53**	**22**	**25**	**22**	**3**	**1**	**25**
Dataset #2	#A	#2	80	72	8	**20**	15	5	1	13
Dataset #2	**#B**	**#2**	**80**	**77**	**3**	**20**	**16**	**4**	**2**	**7**
Dataset#3 : Damages for Worker Wrongful Termination										
Dataset #3	#A	#1	53	47	6	47	34	13	11	19
Dataset #3	**#B**	**#1**	**53**	**53**	**0**	**47**	**34**	**13**	**1**	**13**
Dataset #3	#A	#2	88	75	13	12	7	5	11	18
Dataset #3	**#B**	**#2**	**88**	**85**	**3**	**12**	**7**	**5**	**3**	**8**
Dataset#4 : Damages for Workplace Bullying or Sexual Harassment										
Dataset #4	#A	#1	44	35	9	56	50	6	7	15
Dataset #4	**#B**	**#1**	**44**	**44**	**0**	**56**	**52**	**4**	**0**	**4**
Dataset #4	#A	#2	65	57	8	35	29	6	7	14
Dataset #4	**#B**	**#2**	**65**	**59**	**6**	**35**	**30**	**5**	**0**	**11**

Comments: Prompt engineering didn't improve performance for Datasets#1 and #2. Minimalist prompt slightly improved performance for Datasets#3 and #4. Remarkably, minimalist prompting achieved perfect accuracy for positive outcome#1 of datasets#3 and #4, but missed negative outcomes.

Figure 2b: Confusion Matrix Overview for Llama3 (70b)

User Prompt Dataset	System prompt	Out come	Expert Pos.	LLM True Pos.	LLM False Neg.	Expert Neg	LLM True Neg.	LLM False Pos.	Failed Gen. Errors	Total Errors
Dataset#1 : Involuntary Psychiatric Commitment										
Dataset #1	#A	#1	90	84	6	10	7	3	0	9
Dataset #1	#B	#1	90	89	1	10	9	1	0	2
Dataset #1	#A	#2	81	80	1	19	16	3	0	4
Dataset #1	#B	#2	81	80	1	19	19	0	0	1
Dataset#2 : Undocumented Alien Detention Pending Deportation										
Dataset #2	#A	#1	75	72	3	25	22	3	0	6
Dataset #2	**#B**	**#1**	**75**	**70**	**5**	**25**	**22**	**3**	**0**	**8**
Dataset #2	#A	#2	80	77	3	**20**	20	0	0	3
Dataset #2	**#B**	**#2**	**80**	**79**	**1**	**20**	**20**	**0**	**0**	**1**
Dataset#3 : Damages for Worker Wrongful Termination										
Dataset #3	#A	#1	53	52	1	47	42	5	0	6
Dataset #3	**#B**	**#1**	**53**	**53**	**0**	**47**	**42**	**5**	**0**	**5**
Dataset #3	#A	#2	88	88	0	12	11	1	0	1
Dataset #3	**#B**	**#2**	**88**	**88**	**0**	**12**	**10**	**2**	**0**	**2**
Dataset#4 : Damages for Workplace Bullying or Sexual Harassment										
Dataset #4	#A	#1	44	44	0	56	50	6	0	6
Dataset #4	**#B**	**#1**	**44**	**44**	**0**	**56**	**51**	**5**	**0**	**5**
Dataset #4	#A	#2	65	64	1	35	28	7	0	8
Dataset #4	**#B**	**#2**	**65**	**64**	**1**	**35**	**28**	**7**	**0**	**8**

Comments: Remarkably, the total number of errors for outcome#1 and #2 stays constant for dataset#2 and #3. Prompt engineering managed to shave only 1 error for dataset#4. Best performance improvement fot dataset#1, with 10 errors less. Llama3 already achieved perfect or near-perfect accuracy with baseline prompt for 3 datasets out of 4, and further improvement remained out of reach.

Figure 3a: Mixtral (8x7b) Prompt Engineering Performance Improvement: Delta #B - #A

User Prompt Dataset	LLM Precis. Delta #B-#A	LLM Recall Delta #B-#A	LLM F1 Delta #B-#A	LLM Pos. Accur. Delta #B-#A	LLM Neg. Accur. Delta #B-#A	LLM Global Accur. Delta #B-#A
Dataset#1 : Involuntary Psychiatric Commitment						
Outcome #1	0,962 - 0,964 (-0,2%)	0,866 - 0,900 (-3,77%)	0,912 – 0,930 (-1,9%)	(78 – 81) / 90 (-3,33%)	(7 – 7) / 10 (0%)	(85 – 88) / 100 (-3%)
Outcome #2	1,000 – 0,987 (+1,3%)	0,913 – 0,975 (-6,3%)	0,954 – 0,980 (-2,6%)	(74 – 79) / 81 (-6,17%)	(19 – 18) / 19 (+5,26%)	(93 – 97) / 100 (-4%)
Dataset#2 : Undocumented Alien Detention Pending Deportation						
Outcome #1	0,946 – 0,833 (+13,5%)	0,706 – 0,733 (-3,6%)	0,809 – 0,779 (+3,8%)	(53 – 55) / 75 (-2,6%)	(22 – 14) / 25 (+32%)	(75 – 69) / 100 **(+6%)**
Outcome #2	0,962 – 0,935 (+2,8%)	0,962 – 0,900 (+6,8%)	0,962 – 0,917 (+4,9%)	(77 – 72) / 80 (+6,25%)	(16 – 15) / 20 (+5%)	(93 – 87) / 100 **(+6%)**
Dataset#3 : Damages for Worker Wrongful Termination						
Outcome #1	0,815 – 0,870 (-6%)	1,000 – 0,886 (+12,8%)	0,898 – 0,878 (+2,27%)	(53 – 47) / 53 (+11,32%)	(34 – 34) / 47 (0%)	(87 – 81) / 100 (+6%)
Outcome #2	0,955 – 0,937 (+1,9%)	0,965 – 0,852 (+13,2%)	0,960 – 0,893 (+7,5%)	(85 – 75) / 88 (+11,36%)	(7 – 7) / 12 (0%)	(92 – 82) / 100 (+10%)
Dataset#4 : Damages for Workplace Bullying or Sexual Harassment						
Outcome #1	0,916 – 0,921 (-0,5%)	1,000 – 0,795 (+25,7%)	0,956 – 0,893 (+12%)	(44 – 35) / 44 (+20,45%)	(52 – 50) / 56 (+3,57%)	(96 – 85) / 100 (+11%)
Outcome #2	0,921 – 0,919 (+0,2%)	0,907 – 0,876 (+3,5%)	0,915 – 0,896 (+2,1%)	(59 – 57) / 65 (+3,07%)	(30 – 29) / 35 (+2,07%)	(89 – 86) / 100 (+3%)

Comments: Despite the model's modest size, prompt engineering managed to improve Mixtral performance across the board. Global accuracy for outcome#1 is solid: 93% for dataset#1, 93# for dataset#2, 92% for dataset#3 and 89% for dataset#4. Mean improvement was between 6% and 10% for 3 datasets out of 4.

Figure 3b: Llama3 (70b) Prompt Engineering Performance Improvement: Delta #B - #A

User Prompt Dataset	LLM Precis. Delta #B-#A	LLM Recall Delta #B-#A	LLM F1 Delta #B-#A	LLM Pos. Accur. Delta #B-#A	LLM Neg. Accur. Delta #B-#A	LLM Global Accur. Delta #B-#A
Dataset#1 : Involuntary Psychiatric Commitment						
Outcome #1	0,988 - 0,965 (+2,3%)	0,988 - 0,933 (+5,8%)	0,988 – 0,948 (+4,2%)	(89 – 84) / 90 (+5,55%)	(9 – 7) / 10 (+20%)	(98 – 91) / 100 (+7%)
Outcome #2	1,000 – 0,963 (+3,8%)	0,987 – 0,987 (0%)	0,994 – 0,974 (+2%)	(80 – 80) / 81 (0%)	(19 – 16) / 19 (+15,7%)	(99 – 96) / 100 (+3%)
Dataset#2 : Undocumented Alien Detention Pending Deportation						
Outcome #1	0,958 – 0,960 (-0,2%)	0,933 – 0,960 (-2,8%)	0,946 – 0,960 (-1,4%)	(70 – 72) / 75 (-2,6%)	(22 – 22) / 25 (0%)	(92 – 94) / 100 **-2%**
Outcome #2	1,000 – 1,000 (0%)	0,987 – 0,962 (+2,5%)	0,993 – 0,981 (+1,2%)	(79 – 77) / 80 (+2,5%)	(20 – 20) / 20 (0%)	(99 – 97) / 100 **+2%**
Dataset#3 : Damages for Worker Wrongful Termination						
Outcome #1	0,913 – 0,912 (+0,1%)	1,000 – 0,981 (+1,9%)	0,955 – 0,945 (+1%)	(53 – 52) / 53 (+1,8%)	(42– 42) / 47 (0%)	(95 – 94) / 100 (+1%)
Outcome #2	0,977 – 0,988 (-1,1%)	1,000 – 1,000 (0%)	0,989 – 0,994 (-0,5%)	(88 – 88) / 88 (0%)	(10 – 11) / 12 (-8,3%)	(98 – 99) / 100 (-1%)
Dataset#4 : Damages for Workplace Bullying or Sexual Harassment						
Outcome #1	0,900 – 0,880 (+2,2%)	1,000 – 1,000 (0%)	0,950 – 0,936 (+1,4%)	(44 – 44) / 44 (0%)	(51 – 50) / 56 (+1,7%)	(95 – 94) / 100 (+1%)
Outcome #2	0,901 – 0,901 (0%)	0,984 – 0,984 (0%)	0,941 – 0,940 (+0,1%)	(64 – 64) / 65 (0%)	(28 – 28) / 35 (0%)	(92 – 92) / 100 (0%)

Comments: The different runs with Llama3 show that the model size and pre-training is the dominant factor for performance. Prompt engineering of the system prompt had a negligible impact on performance. Llama3 seems to have reached a glass ceiling of near-perfect accuracy, that can't be broken by clever prompt engineering. The resistant errors most likely are the consequence of the poor quality of judicial documents that were fed to the model.

4.1 Dataset #1

Figure 4: Confusion Matrix for Mixtral (8x7b): Prompt#A for Outcome 1

Precision 0,964 – Recall 0,900 – Specificity. 0,700 – F1 0,930 – Accuracy 88%

Outcome 1 1st judge	**Expert answer: Neg. (RELEASE)**	**Expert answer: Pos. (EXTENSION)**	**Total (LLM)**
Mixtral answer: Neg. (RELEASE)	TN: 7	FN: 9	16
Mixtral answer: Pos. (EXTENSION)	FP: 3	TP: 81	84
Total (expert)	10	90	100

Figure 5: Confusion Matrix for Mixtral (8x7b): Prompt#A for Outcome 2

Precision 0,987 – Recall 0,975 – Specif. 0,947 – F1 0,980 – Accuracy 97%

Outcome 2 2nd judge	**Expert answer: Neg. (RELEASE)**	**Expert answer: Pos. (EXTENSION)**	**Total (LLM)**
Mixtral answer: Neg. (RELEASE)	TN: 18	FN: 2	20
Mixtral answer: Pos. (EXTENSION)	FP: 1	TP: 79	80
Total (expert)	19	81	100

Figure 6: Confusion Matrix for Llama3 (70b): Prompt#A for Outcome 1

Precision 0,965 – Recall 0,933 – Specif. 0,700 – F1 0,948 – Accuracy 91%

Outcome 1 1st judge	Expert answer: Neg. (RELEASE)	Expert answer: Pos. (EXTENSION)	Total (LLM)
LlaMa answer: Neg. (RELEASE)	TN: 7	FN: 6	13
LlaMa answer: Pos. (EXTENSION)	FP: 3	TP: 84	87
Total (expert)	10	90	100

Figure 7: Confusion Matrix for Llama3 (70b) : Prompt#A for Outcome 2

Precision 0,963 – Recall 0,987 – Specif. 0,842 – F1 0,974 – Accuracy 96%

Outcome 2 2nd judge	Expert answer: Neg. (RELEASE)	Expert answer: Pos. (EXTENSION)	Total (LLM)
LlaMa answer: Neg. (RELEASE)	TN: 16	FN: 1	17
LlaMa answer: Pos. (EXTENSION)	FP: 3	TP: 80	83
Total (expert)	19	81	100

Figure 8: Confusion Matrix for Mixtral (8x7b) : Prompt#B for Outcome 1

Precision 0,962 – Recall 0,866 – Specif. 0,700 – F1 0,912 – Accuracy 85%

Outcome 1 1st judge	Expert answer: Neg. (RELEASE)	Expert answer: Pos. (EXTENSION)	Total (LLM)
Mixtral answer: Neg. (RELEASE)	TN: 7	FN: 9	16
Mixtral answer: Pos. (EXTENSION)	FP: 3	TP: 78	81
Empty: failed gen.		3	3
Total (expert)	10	90	100

Figure 9: Confusion Matrix for Mixtral (8x7b) : Prompt#B for Outcome 2

Precision 1,000 – Recall 0,913 – Specif. 1,000 – F1 0,954 – Accuracy 93%

Outcome 2 2nd judge	Expert answer: Neg. (RELEASE)	Expert answer: Pos. (EXTENSION)	Total (LLM)
Mixtral answer: Neg. (RELEASE)	TN: 19	FN: 3	22
Mixtral answer: Pos. (EXTENSION)	FP: 0	TP: 74	74
Empty: failed gen.		4	4
Total (expert)	19	81	100

Figure 10: Confusion Matrix for Llama3 (70b): Prompt#B for Outcome 1

Precision 0,988 – Recall 0,988 – Specif. 0,900 – F1 0,988 – Accuracy 98%

Outcome 1 1st judge	Expert answer: Neg. (RELEASE)	Expert answer: Pos. (EXTENSION)	Total (LLM)
LlaMa answer: Neg. (RELEASE)	TN: 9	FN: 1	10
LlaMa answer: Pos. (EXTENSION)	FP: 1	TP: 89	90
Total (expert)	10	90	100

Figure 11: Confusion Matrix for Llama3 (70b): Prompt#B for Outcome 2

Precision 1,000 – Recall 0,987 – Specif. 1,000 – F1 0,994 – Accuracy 99%

Outcome 2 2nd judge	Expert answer: Neg. (RELEASE)	Expert answer: Pos. (EXTENSION)	Total (LLM)
LlaMa answer: Neg. (RELEASE)	TN: 19	FN: 1	20
LlaMa answer: Pos. (EXTENSION)	FP: 0	TP: 80	80
Total (expert)	19	81	100

4.2 Dataset #2

Figure 12: Confusion Matrix for Mixtral (8x7b) : Prompt#A for Outcome#1

Precision 0,833 – Recall 0,733 – Specif. 0,560 – F1 0,779 – Accuracy 69%

Outcome 1 1st judge	Expert answer: Neg. (RELEASE)	Expert answer: Pos. (EXTENSION)	Total (LLM)
Mixtral answer: Neg. (RELEASE)	TN: 14	FN: 20	34
Mixtral answer: Pos. (EXTENSION)	FP: 11	TP: 55	66
Total (expert)	25	75	100

Figure 13: Confusion Matrix for Mixtral (8x7b) : Prompt#A for Outcome#2

Precision 0,935 – Recall 0,900 – Specif. 0,750 – F1 0,917 – Accuracy 87%

Outcome 2 2nd judge	Expert answer: Neg. (RELEASE)	Expert answer: Pos. (EXTENSION)	Total (LLM)
Mixtral answer: Neg. (RELEASE)	TN: 15	FN: 8	23
Mixtral answer: Pos. (EXTENSION)	FP: 5	TP: 72	77
Total (expert)	20	80	100

Figure 14: Confusion Matrix for LlaMa3 (70b): Prompt#A for Outcome#1

Precision 0,960 – Recall 0,960 – Specif. 0,880 – F1 0,960 – Accuracy 94%

Outcome 1 1st judge	Expert answer: Neg. (RELEASE)	Expert answer: Pos. (EXTENSION)	Total (LLM)
LlaMa answer: Neg. (RELEASE)	TN: 22	FN: 3	25
LlaMa answer: Pos. (EXTENSION)	FP: 3	TP: 72	75
Total (expert)	25	75	100

Figure 15: Confusion Matrix for LlaMa3 (70b): Prompt#A for Outcome#2

Precision 1,000 – Recall 0,962 – Specif. 1,000 – F1 0,981 – Accuracy 97%

Outcome 2 2nd judge	Expert answer: Neg. (RELEASE)	Expert answer: Pos. (EXTENSION)	Total (LLM)
LlaMa answer: Neg. (RELEASE)	TN: 20	FN: 3	23
LlaMa answer: Pos. (EXTENSION)	FP: 0	TP: 77	77
Total (expert)	20	80	100

Figure 16: Confusion Matrix for Mixtral (8x7b): Prompt#B for Outcome#1

Precision 0,946 – Recall 0,706 – Specif. 0,880 – F1 0,809 – Accuracy 75%

Outcome 1 1st judge	Expert answer: Neg. (RELEASE)	Expert answer: Pos. (EXTENSION)	Total (LLM)
Mixtral answer: Neg. (RELEASE)	TN: 22	FN: 21	43
Mixtral answer: Pos. (EXTENSION)	FP: 3	TP: 53	56
Empty		1	1
Total (expert)	25	75	100

Figure 17: Confusion Matrix for Mixtral (8x7b): Prompt#B for Outcome#2

Precision 0,962 – Recall 0,962 – Specif. 0,800 – F1 0,962 – Accuracy 93%

Outcome 2 2nd judge	Expert answer: Neg. (RELEASE)	Expert answer: Pos. (EXTENSION)	Total (LLM)
Mixtral answer: Neg. (RELEASE)	TN: 16	FN: 2	18
Mixtral answer: Pos. (EXTENSION)	FP: 3	TP: 77	80
Empty	1	1	2
Total (expert)	20	80	100

Figure 18: Confusion Matrix for LlaMa3 (70b): Prompt#B for Outcome#1

Precision 0,958 – Recall 0,933 – Specif. 0,880 – F1 0,946 – Accuracy 92%

Outcome 1 1st judge	Expert answer: Neg. (RELEASE)	Expert answer: Pos. (EXTENSION)	Total (LLM)
LlaMa answer: Neg. (RELEASE)	TN: 22	FN: 5	27
LlaMA answer: Pos. (EXTENSION)	FP: 3	TP: 70	73
Total (expert)	25	75	100

Figure 19: Confusion Matrix for LlaMa3 (70b): Prompt#B for Outcome#2

Precision 1,000 – Recall 0,987 – Specif. 1,000 – F1 0,993 – Accuracy 99%

Outcome 2 2nd judge	Expert answer: Neg. (RELEASE)	Expert answer: Pos. (EXTENSION)	Total (LLM)
LlaMa answer: Neg. (RELEASE)	TN: 20	FN: 1	21
LlaMA answer: Pos. (EXTENSION)	FP: 0	TP: 79	79
Total (expert)	20	80	100

4.3 Dataset #3

Figure 20: Confusion Matrix for Mixtral (8x7b): Prompt#A for Outcome#1

Precision 0,870 – Recall 0,886 – Specif. 0,723 – F1 0,878 – Accuracy 81%

Outcome 1 (1st judge)	Expert answer: Neg. (REJECTED)	Expert answer: Pos. (ACCEPTED)	Total (LLM)
Mixtral answer: Neg. (REJECTED)	TN: 34	FN: 1	35
Mixtral answer: Pos. (ACCEPTED)	FP: 7	TP: 47	54
Empty (failed generation)	6	5	11
Total (expert)	47	53	100

Figure 21: Confusion Matrix for Mixtral (8x7b): Prompt#A for Outcome#2

Precision 0,937 – Recall 0,852 – Specif. 0,583 – F1 0,893 – Accuracy 82%

Outcome 2 (2nd judge)	Expert answer: Neg. (REJECTED	Expert answer: Pos. (ACCEPTED)	Total (LLM)
Mixtral answer: Neg. (REJECTED)	TN: 7	FN: 2	9
Mixtral answer: Pos. (ACCEPTED)	FP: 5	TP: 75	80
Empty (failed generation)	0	11	11
Total (expert)	12	88	100

Figure 22: Confusion Matrix LlaMa3 (70b): Prompt#A for Outcome#1

Precision 0,912 – Recall 0,981 – Specif. 0,893 – F1 0,945 – Accuracy 94%

Outcome 1 (1st judge)	Expert answer: Neg. (REJECTED	Expert answer: Pos. (ACCEPTED)	Total (LLM)
LlaMA answer: Neg. (REJECTED)	TN: 42	FN: 1	43
LlaMA answer: Pos. (ACCEPTED)	FP: 5	TP: 52	57
Total (expert)	47	53	100

Figure 23: Confusion Matrix for LlaMa3 (70b): Prompt#A for Outcome#2

Precision 0,988 – Recall 1,000 – Specif. 0,916 – F1 0,994 – Accuracy 99%

Outcome 2 (2nd judge)	Expert answer: Neg. (REJECTED)	Expert answer: Pos. (ACCEPTED)	Total (LLM)
LlaMa answer: Neg. (REJECTED)	TN: 11	FN: 0	11
LlaMa answer: Pos. (ACCEPTED)	FP: 1	TP: 88	89
Total (expert)	12	88	100

Figure 24: Confusion Matrix for Mixtral (8x7b): Prompt#B for Outcome#1

Precision 0,815 – Recall 1,000 – Specif. 0,723 – F1 0,898 – Accuracy 87%

Outcome 1 (1st judge)	Expert answer: Neg. (REJECTED)	Expert answer: Pos. (ACCEPTED)	Total (LLM)
Mixtral answer: Neg. (REJECTED)	TN: 34	FN: 0	34
Mixtral answer: Pos. (ACCEPTED)	FP: 12	TP: 53	65
Empty: gen. failed	1		1
Total (expert)	47	53	100

Figure 25: Confusion Matrix for Mixtral (8x7b): Prompt#B for Outcome#2

Precision 0,955 – Recall 0,965 – Specif. 0,583 – F1 0,960 – Accuracy 92%

Outcome 2 (2nd judge)	Expert answer: Neg. (REJECTED)	Expert answer: Pos. (ACCEPTED)	Total (LLM)
Mixtral answer: Neg. (REJECTED)	TN: 7	FN: 1	8
Mixtral answer: Pos. (ACCEPTED)	FP: 4	TP: 85	89
Empty: gen. failed	1	2	3
Total (expert)	12	88	100

Figure 26: Confusion Matrix for LlaMa3 (70b): Prompt#B for Outcome#1

Precision 0,913 – Recall 1,000 – Specif. 0,893 – F1 0,955 – Accuracy 95%

Outcome 1 (1st judge)	Expert answer: Neg. (REJECTED)	Expert answer: Pos. (ACCEPTED)	Total (LLM)
LlaMa answer: Neg. (REJECTED)	TN: 42	FN: 0	42
LlaMa answer: Pos. (ACCEPTED)	FP: 5	TP: 53	58
Total (expert)	47	53	100

Figure 27: Confusion Matrix for LlaMa3 (70b): Prompt#B for Outcome#2

Precision 0,977 – Recall 1,000 – Specif. 0,833 – F1 0,989 – Accuracy 98%

Outcome 2 (2nd judge)	Expert answer: Neg. (REJECTED)	Expert answer: Pos. (ACCEPTED)	Total (LLM)
LlaMa answer: Neg. (REJECTED)	TN: 10	FN: 0	10
LlaMa answer: Pos. (ACCEPTED)	FP: 2	TP: 88	90
Total (expert)	12	88	100

4.4 Dataset #4

Figure 28: Confusion Matrix for Mixtral (8x7b): Prompt#A for Outcome#1

Precision 0,921 – Recall 0,795 – Specif. 0,892 – F1 0,853 – Accuracy 85%

Outcome 1 (1st judge)	Expert answer: Neg. (REJECTED)	Expert answer: Pos. (ACCEPTED)	Total (LLM)
Mixtral answer: Neg. (REJECTED)	TN: 50	FN: 5	55
Mixtral answer: Pos. (ACCEPTED)	FP: 3	TP: 35	38
Empty (failed generation)	3	4	7
Total (expert)	56	44	100

Figure 29: Confusion Matrix for Mixtral (8x7b): Prompt#A for Outcome#2

Precision 0,919 – Recall 0,876 – Specif. 0,828 – F1 0,896 – Accuracy 86%

Outcome 2 (2nd judge)	Expert answer: Neg. (REJECTED)	Expert answer: Pos. (ACCEPTED)	Total (LLM)
Mixtral answer: Neg. (REJECTED)	TN: 29	FN: 2	31
Mixtral answer: Pos. (ACCEPTED)	FP: 5	TP: 57	62
Empty (failed generation)	1	6	7
Total (expert)	35	65	100

Figure 30: Confusion Matrix for LlaMa3 (70b): Prompt#A for Outcome#1

Precision 0,880 – Recall 1,000 – Specif. 0,892 – F1 0,936 – Accuracy 94%

Outcome 1 (1st judge)	Expert answer: Neg. (REJECTED)	Expert answer: Pos. (ACCEPTED)	Total (LLM)
LlaMa answer: Neg. (REJECTED)	TN: 50	FN: 0	50
LlaMa answer: Pos. (ACCEPTED)	FP: 6	TP: 44	50
Total (expert)	56	44	100

Figure 31: Confusion Matrix for LlaMa3 (70b): Prompt#A for Outcome#2

Precision 0,901 – Recall 0,984 – Specif. 0,800 – F1 0,940 – Accuracy 92%

Outcome 2 (2nd judge)	Expert answer: Neg. (REJECTED)	Expert answer: Pos. (ACCEPTED)	Total (LLM)
Mixtral answer: Neg. (REJECTED)	TN: 28	FN: 1	29
Mixtral answer: Pos. (ACCEPTED)	FP: 7	TP: 64	71
Total (expert)	35	65	100

Figure 32: Confusion Matrix for Mixtral (8x7b): Prompt#B for Outcome#1

Precision 0,916 – Recall 1,000 – Specif. 0,928 – F1 0,956 – Accuracy 96%

Outcome 1 1st judge	Expert answer: Neg. (REJECTED)	Expert answer: Pos. (ACCEPTED)	Total (LLM)
Mixtral answer: Neg. (REJECTED)	TN: 52	FN: 0	52
Mixtral answer: Pos. (ACCEPTED)	FP: 4	TP: 44	48
Total (expert)	56	44	100

Figure 33: Confusion Matrix for Mixtral (8x7b): Prompt#B for Outcome#2

Precision 0,921 – Recall 0,907 – Specif. 0,857 – F1 0,915 – Accuracy 89%

Outcome 2 2nd judge	Expert answer: Neg. (REJECTED)	Expert answer: Pos. (ACCEPTED)	Total (LLM)
Mixtral answer: Neg. (REJECTED)	TN: 30	FN: 6	36
Mixtral answer: Pos. (ACCEPTED)	FP: 5	TP: 59	64
Total (expert)	35	65	100

Figure 34: Confusion Matrix for LlaMa3 (70b): Prompt#B for Outcome#1

Precision 0,900 – Recall 1,000 – Specif. 0,910 – F1 0,950 - Accuracy 95%

Outcome 1 1st judge	Expert answer: Neg. (REJECTED)	Expert answer: Pos. (ACCEPTED)	Total (LLM)
LlaMa answer: Neg. (REJECTED)	TN: 51	FN: 0	51
LlaMa answer: Pos. (ACCEPTED)	FP: 5	TP: 44	49
Total (expert)	56	44	100

Figure 35: Confusion Matrix for LlaMa3 (70b): Prompt#B for Outcome#2

Precision 0,901 – Recall 0,984 – Specif. 0,800 – F1 0,941 – Accuracy 92%

Outcome 2 2nd judge	Expert answer: Neg. (REJECTED)	Expert answer: Pos. (ACCEPTED)	Total (LLM)
LlaMa answer: Neg. (REJECTED)	TN: 28	FN: 1	29
LlaMA answer: Pos. (ACCEPTED)	FP: 7	TP: 64	71
Total (expert)	35	65	100

5. Discussion

Overall, both LLMs achieved higher performance in data extraction, compared to machine learning techniques. For comparison, Petrova et al.[36] reported a F1 score of 82,32%. Vacek et al.[37] reported an average F1 score of 91%. On a very similar task with pre-trained models, Vaudaux et al. reached 92% with CamemBERT, 94% with FlauBERT and 96% with JuriBERT.[38] Llama3 clearly outperformed these results.

The Mixtral (8x7b) model exhibited variable performance when applied to datasets related to involuntary psychiatric commitment, undocumented alien detention, and workplace-related damages. This variability indicates challenges in consistently identifying true positives and negatives, particularly in datasets with a high number of negative outcomes. The model's performance metrics highlight its limitations in handling imbalanced class distributions.

On the other hand, the Llama3 (70b) model demonstrated superior performance across all datasets, consistently achieving high precision and recall metrics. This suggests that Llama3's larger size and extensive pre-training enable it to extract effectively judicial outcomes from complex documents.

While prompt engineering offered some potential improvements for Mixtral, it was inadequate in overcoming the model's inherent limitations, likely due to its smaller size and less comprehensive training. Mixtral's performance improvements were marginal and often inconsistent across different datasets, indicating that prompt adjustments can refine model responses but cannot fundamentally enhance a model's capabilities beyond its architectural constraints. This underscores the necessity for large-scale, well-trained models to tackle complex data extraction tasks effectively.

Llama3's performance reached a near-perfect accuracy plateau, that couldn't be improved by prompt engineering, suggesting its limitations were caused by the low-quality of input data. Despite high precision and recall rates, minor errors persistance highlights the critical importance of input data quality. While prompt engineering is essential to harness Llama3's full potential, it is insufficient to address errors from poor-quality documents, highlighting the need for robust data preprocessing.

Regarding prompt engineering, results from Prompt #A and Prompt #B align with Sainz et al prompting strategy,[39] with instructions in JSON format. Directional Stimulus prompting in JSON format is a reliable method for extracting judicial

[36] Petrova, "Extracting Outcomes from Appellate Decisions in US State Courts, " 133–142.
[37] Vacek, "Litigation Analytics, " 45–54.
[38] Vaudaux, "Pretrained Language Models v. Court Ruling Predictions, " 38-43.
[39] Sainz. "GoLLIE."

outcomes from unstructured text. As shown in Figure 3, advanced techniques like chain-of-thought prompting or generated knowledge prompting could enhance a weaker model like Mixtral, but they have limited impact on a more advanced model like Llama3 when it already reached is plateau. Our results suggests that Llama3 can perform data extraction reliably without additional context, comments, definitions or detailed instructions. Llama3 is already well-suited for data extraction tasks, and is able to rely on simple cues in JSON format without lengthy and detailed instructions.

In conclusion, LLMs in zero-shot setting can perform accurate legal data extraction from unstructured text without fine-tuning. To further validate these results and explore the full potential of LLMs in legal data extraction, future research should expand the number of manually annotated cases to confirm the reliability of the results. It will also be useful to evaluate the extraction of numerical values, such as compensation amounts. Creating a large multilingual benchmark of annotated cases accessible to the research community will enable the testing of prompts and models, driving further innovation in this field. Finally, a qualitative investigation of errors should be implemented, to better understand their causes. If LLMs are confirmed as good data extractors, a large scale automated extraction could provide useful data for Legal Empirical research.

Data availability statement

All the data associated with the research is available online at: ZAMBRANO, Guillaume. 2024. "LILA: Litigation Data Extraction with LLMs." OSF. July 13. doi:10.17605/OSF.IO/M4FKT.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article .

Bibliography

Agrawal, Monica, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. "Large Language Models Are Few-Shot Clinical Information Extractors." *ArXiv.org*. (2022). https://doi.org/10.48550/arXiv.2205.12689.

Ashley, Kevin. *Artificial Intelligence and Legal Analytics : New Tools for Law Practice in the Digital Age*. Cambridge University Press. 2017.

Bertalan, Vithor Gomes and Evandro Eduardo Ruiz. "Using Attention Methods to Predict Judicial Outcomes." *Artificial Intelligence and Law* 32, no 1 (2022): 87–115. https://doi.org/10.1007/s10506-022-09342-7.

Caufield, J. Harry, Harshad Hegde, Vincent Emonet, Nomi L Harris, Marcin P Joachimiak, Nicolas Matentzoglu, HyeongSik Kim, Sierra Moxon, Justin T Reese, Melissa A Haendel, Peter N Robinson, Christopher J Mungall. "Structured Prompt Interrogation and Recursive Extraction Of Semantics (SPIRES): A Method For Populating Knowledge Bases Using Zero-Shot Learning." *Bioinformatics* 40, no 3 (2024). https://doi.org/10.1093/bioinformatics/btae104.

Coelho, Gustavo, Alimed Celecia, Jefferson de Sousa, Melissa Lemos, Maria Lima, Ana Mangeth, Isabella Frajhof and Marco Casanova. "Information Extraction in the Legal Domain: Traditional Supervised Learning vs. ChatGPT." In *Proceedings of the 26th International Conference on Enterprise Information Systems*. Vol. 1, SciTePress, 579-586. 2024. https://www.scitepress.org/Documents/2024/124998/.

de Faria, Joana Ribeiro, Huiyuan Xie and Felix Steffek. "Automatic Information Extraction From Employment Tribunal Judgements Using Large Language Models." *University of Cambridge Faculty of Law Research Paper*, no 13 2024. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4776160.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina 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*. Volume 1 (Long and Short Papers), 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Dubey, Abhimanyu, Jauhri, Abhinav, Pandey, Abhinav, *et al*. "The Llama 3 Herd of Models." *arXiv preprint* 2024. https://doi.org/10.48550/arXiv.2407.21783.

Guha, Neel, Nyarko, Julian, HO, Daniel, *et al*. "Legalbench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models." *arXiv preprint* 2023. https://doi.org/10.48550/arXiv.2308.11462.

Hart, H. L .A. *The Concept of Law*. Oxford University Press, 1961.

Huang, Wei, Xudong Ma, Haotong Qin, Xingyu Zheng, Chengtao Lv, Hong Chen, Jie Luo, Xiaojuan Qi, Xianglong Liu, Michele Magno. "How Good Are Low-Bit Quantized Llama3 Models? An Empirical Study." *arXiv preprint* 2024. https://doi.org/10.48550/arXiv.2404.14047.

Izzidien, Ahmed, Holli Sargeant and Felix Steffek. "LLM vs. Lawyers: Identifying a Subset of Summary Judgments in a Large UK Case Law Dataset." *arXiv preprint* 2024. https://doi.org/10.48550/arXiv.2403.04791.

Kartchner, David, Selvi Ramalingam, Irfan Al-Hussaini, Olivia Kronick and Cassie Mitchell. "Zero-Shot Information Extraction for Clinical Meta-Analysis using Large Language Models." In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Association for Computational Linguistics, 396–405. 2023. https://doi.org/10.18653/v1/2023.bionlp-1.37.

Loevinger, Lee. "Jurimetrics: The Methodology of Legal Inquiry." *Law and Contemporary Problems* (1963): 5-35.

McInerney, Denis. "CHiLL : Zero-shot Custom Interpretable Feature Extraction from Clinical Notes with Large Language Models." In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore*. Association for Computational Linguistics. 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.568.

Mistica, Meladel, Geordie Zhang, Hui Chia, Kabir Manandhar Shrestha, Rohit Gupta, Saket Khandelwal, Jeannie Paterson, Tim Baldwin and Daniel Beck. "Information Extraction from Legal Documents: A Study in the Context of Common Law Court Judgements." In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*, 98–103, ACLWeb. 2020. https://aclanthology.org/2020.alta-1.12/.

Petrova, Alina, John Armour and Thomas Lukasiewicz. "Extracting Outcomes from Appellate Decisions in US State Courts." In *33rd International Conference on Legal Knowledge and Information Systems (JURIX 2020)*. Volume 334: Legal Knowledge and Information Systems. Frontiers in Artificial Intelligence and Applications. 133–42. https://ebooks.iospress.nl/volumearticle/56170.

Rehaag, Sean. "Luck of the Draw III: Using AI to Using AI to Extract Data About Decision-Making in Federal Court Stays of Removal." *Queen's LJ*. (2024): 49:2 73. https://ssrn.com/abstract=4322881.

Sainz, Oscar, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau and Eneko Agirre. "GoLLIE: Annotation Guidelines Improve Zero-Shot Information-Extraction." *ArXiv.org*. 2024. https://doi.org/10.48550/arXiv.2310.03668.

Santosuosso, Amedeo and Giulia Pinotti. "Bottleneck or Crossroad? Problems of Legal Sources Annotation and Some Theoretical Thoughts." *Stats* 3, no 3 (2020): 376-395. https://doi.org/10.3390/stats3030024.

Sargeant, Holli, Ahmed Izzidien and Feliz Steffek. "Topic Modelling Case Law Using a Large Language Model and a New Taxonomy for UK Law: AI Insights into Summary Judgment". *University of Cambridge Faculty of Law Research Paper* no 21 (2024). https://ssrn.com/abstract=4836558.

Satterfield, Nolan, Parker Holbrook and Thomas Wilcox. "Fine-tuning Llama with Case Law Data to Improve Legal Domain Performance." *OSF Preprints*. 2024. https://osf.io/preprints/osf/e6mjs.

Savelka Jaromir. "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)*. Association for Computing Machinery, 447–451. 2023. https://doi.org/10.1145/3594536.3595161.

Schlegel, John Henry. "Empirical Legal Research At Johns Hopkins : Walter Wheeler Cook And His Friends." In *American Legal Realism and Empirical Social Science*, 147–210. University of North Carolina Press, 1995. www.jstor.org/stable/10.5149/9780807864364_schlegel.9.

Sclar, Melanie, Yejin Choi, Yulia Tsvetkov and Alane Suhr. "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting." *arXiv preprint*. 2023. https://doi.org/10.48550/arXiv.2310.11324.

Shao, Minghao, Abdul Basit, Ramesh Karri and Muhammad Shafique "Survey of Different Large Language Model Architectures: Trends, Benchmarks, and Challenges." *IEEE Access*, 2024. http://dx.doi.org/10.1109/ACCESS.2024.3482107.

Shen, Shirong, Guilin Qi, Zhen Li, Sheng Bi and Lusheng Wang "Hierarchical Chinese Legal Event Extraction via Pedal Attention Mechanism." In *Proceedings of the 28th International Conference on Computational Linguistics*. 2020. https://doi.org/10.18653/v1/2020.coling-main.9.

Sun, Qi Kun Huang, Xiaocui Yang, Rong Tong and Kun Zhang and Soujanya Poria. "Consistency Guided Knowledge Retrieval and Denoising in LLMs for Zero-Shot Document-Level Relation Triplet Extraction." In *Proceedings of the ACM on Web Conference*, 4407-4416. 2024. https://doi.org/10.1145/3589334.3645678.

Vacek Thomas, Ronald Teo, Dezhao Song, Timothy Nugent, Conner Cowling and Frank Schilder. "Litigation Analytics: Case Outcomes Extracted from US Federal Court Dockets." In *Proceedings of the Natural Legal Language Processing Workshop 2019*, 45–54. ACLWeb. https://doi.org/10.18653/v1/W19-2206.

Vaudaux Olivia, Caroline Bazzoli, Maximin Coavoux, Géraldine Vial and Étienne Vergès. "Pretrained Language Models v. Court Ruling Predictions: A Case Study on a Small Dataset of French Court of Appeal Rulings." In *Proceedings of the Natural Legal Language Processing Workshop 2023*, ACLWeb. Singapore: Association for Computational Linguistics. 2023. https://doi.org/10.18653/v1/2023.nllp-1.5.

Wadhwa Somin, Jay DeYoung, Benjamin Nye, Silvio Amir and Byron C. Wallace. "Jointly Extracting Interventions, Outcomes, and Findings from RCT Reports with LLMs." *ArXiv.org*. 2023. https://doi.org/10.48550/arXiv.2305.03642.

Xiang, Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang and Wenjuan Han. "Zero-Shot Information Extraction via Chatting with ChatGPT". *ArXiv preprint*. 2023. https://doi.org/10.48550/arxiv.2302.10205.

Yao Feng, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen and Maosong Sun. "LEVEN: A Large-Scale Chinese Legal Event Detection Dataset." *Findings of the Association for Computational Linguistics : ACL 2022*, January. https://doi.org/10.18653/v1/2022.findings-acl.17.

Yue Chongjian, Xinrun Xu, Xiaojun Ma, Lun Du, Hengyu Liu, Zhiming Ding, Yanbing Jiang, Shi Han and Dongmei Zhang. "Leveraging LLMs for KPIs Retrieval from Hybrid Long-Document: A Comprehensive Framework and Dataset." *ArXiv preprint*. 2023. https://doi.org/10.48550/arxiv.2305.16344.

Yue, Linan, Qi Liu, Lili Zhao, Li Wang, Weibo Gao and Yanqing An. "Event Grounded Criminal Court View Generation with Cooperative (Large) Language Models." *ArXiv preprint*. 2024. https://doi.org/10.48550/arXiv.2404.07001.

Zhong Haoxi, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. "How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence." *ArXiv preprint*. 2020. https://doi.org/10.48550/arXiv.2004.12158.