INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Forecasting the European Courts of Human Rights' Decisions: fine-tuning LLM for Predicting Human Rights Violations

Sara Batista Vicente Fernandes

Master in Business Analytics

Supervisor:
Professor Doctor Renata Braga Klevenhusen,
Fluminense Federal University

Doctor Francisco Cruz,
ISCTE Executive Education

September, 2025

BUSINESS
SCHOOL

Forecasting the European Courts of Human Rights' Decisions: fine-tuning LLM for Predicting Human Rights Violations

Sara Batista Vicente Fernandes

Master in Business Analytics

Supervisor:
Professor Doctor Renata Braga Klevenhusen,
Fluminense Federal University

Doctor Francisco Cruz,
ISCTE Executive Education

September, 2025

In loving memory of my grandmother, Rosalina, whose love guides my every step.

Acknowledgment

To my friends and family, for always encouraging me to keep going. A special thank-you to André for unfailing motivation and for believing in me when it mattered most.

To my supervisors, for rigorous guidance and continued trust throughout this journey. Your patience and high standards made a real difference.

To Professor Raul Laureano, for believing in my potential despite my background in Law. That trust not only made this project possible but also helped me get to where I am today.

Resumo

Esta dissertação analisa a previsão de decisões do Tribunal Europeu dos Direitos Humanos a partir do texto das petições apresentadas. O problema é de classificação binária: antecipar se o Tribunal declarará ou não a violação da Convenção Europeia dos Direitos Humanos. Ao contrário de estudos baseados em variáveis estruturadas, este trabalho focou-se apenas no conteúdo textual. Para escalar o uso de *Large Language Models*, introduziu-se uma sumarização com instruções no GPT-4o que reduz o comprimento da entrada e o custo computacional, preservando a informação jurídica essencial sem perda de desempenho.

Foi realizada uma análise comparativa do *fine-tuning* em três modelos GPT e, em paralelo, experiências para (i) comparar o desempenho *zero-shot* em petições sumarizadas versus texto integral, avaliando eventual perda de informação, e (ii) estudar o efeito dos ajustes de hiperparâmetros. Os resultados mostram que se obtém desempenho preditivo relevante mesmo com conjuntos de treino pequenos. A melhor configuração, baseada no GPT-4o, alcançou 73% de *accuracy*, 81% de *recall* e F1-score de 78% na classe "violação". Estes resultados cumpriram os critérios de sucesso pré-definidos e demonstraram o potencial dos LLMs para apoiar a triagem inicial de petições, identificando casos mais propensos a envolver violações de direitos humanos.

As conclusões contribuem para o debate sobre a previsão de decisões judiciais com inteligência artificial, propondo um processo simples e reproduzível para a classificação de textos jurídicos. Embora persistam desafios de interpretabilidade, o estudo indica que os LLMs podem complementar o trabalho judicial, promovendo consistência e eficiência sem substituir o raciocínio humano.

Palavras-chave: Tribunal europeu dos direitos humanos, direitos humanos, previsão de decisões judiciais, classificação de texto jurídico, generative pre-trained, modelos de linguagem

Classificação JEL: C45, K38

Abstract

This master's thesis examines the possibility of predicting decisions of the European Court of Human Rights based on applicants' petitions, using exclusively the petition text. The task was framed as a binary classification problem: predicting whether the Court will or will not find a violation of the European Convention on Human Rights. Unlike previous studies that relied on structured metadata and manually engineered features, this work focused solely on textual content. To enable more scalable use of Large Language Models, a GPT-4o-prompted summarisation step was introduced to reduce input length and computational cost without affecting performance, while preserving essential legal information.

A comparative analysis of fine-tuning across three GPT models was conducted, alongside experiments that (i) compared zero-shot performance on summarised petitions versus full text to assess potential information loss, and (ii) examined the effect of hyperparameter adjustments. The results show that meaningful predictive performance can be achieved even with relatively small training sets. The best configuration, based on GPT-4o, achieved 73% accuracy, 81% recall, and an F1-score of 78% for the "violation" class. These results met the predefined success criteria and demonstrated the potential of LLMs to support the early triage of petitions, identifying cases more likely to involve human-rights violations.

The findings contribute to ongoing debates on judicial decision prediction with artificial intelligence, proposing a simple and reproducible pipeline for legal text classification. Although challenges remain regarding interpretability, this study shows that LLMs can complement judicial work by improving consistency and efficiency without replacing human reasoning.

Keywords: European court of human rights, human rights law, judicial decision prediction, legal text classification, generative pre-trained, large language models
JEL Classification: C45, K38

ACRONYMS & ABBREVIATIONS

AI: Artificial Intelligence

CEPEJ: European Commission for the Efficiency of Justice

CRISP-DM: Cross-Industry Standard Process for Data Mining

DL: Deep Learning

e.g.: exempli gratia (for example)

ECHR: European Convention on Human Rights

ECtHR: European Court of Human Rights

GPT: Generative Pre-trained Transformer

ICAIL: International Conference on Artificial Intelligence and Law

LDA: Latent Dirichlet Allocation

LLM: Large Language Model

ML: Machine Learning

NLP: Natural Language Processing

pp: percentage points

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

SHAP: SHapley Additive exPlanations

SLR: Systematic Literature Review

vs: versus

INDEX

INDEX OF FIGURES

INDEX OF TABLES

1. INTRODUCTION

1.1. Context and motivation

Judicial systems across Europe are increasingly challenged by growing caseloads and limited resources. Among these institutions, the European Court of Human Rights (ECtHR) plays a critical role in safeguarding fundamental rights under the European Convention on Human Rights (Council of Europe, 1950). However, the volume of petitions it receives far exceeds its processing capacity, resulting in significant backlogs and delays (ECtHR, 2024).

At the same time, advances in Artificial Intelligence (AI) – particularly Large Language Models (LLMs) – have opened new possibilities for analysing unstructured legal texts. These models are capable of processing legal arguments and complex narratives in ways that traditional tools cannot. As a result, their potential for supporting legal decision-making and improving procedural efficiency has been increasingly recognised (Lai et al., 2024).

1.2. Research problem and objectives

This study addresses whether outcomes at the ECtHR can be predicted from the text of the initial petition[1] alone, as available at filing and before any subsequent procedural developments. The practical motivation is early triage under resource constraints, where the ethical cost of missing potential violations is higher than flagging non-violations; accordingly, evaluation prioritises the correct identification of "violation" cases. The central research question is: *Why and how can large language models be applied effectively to predict ECtHR decisions?*

In line with this problem framing, the primary objective is to establish the feasibility and utility of prediction from unstructured petition narratives without recourse to structured metadata or later procedural artefacts, and to assess performance on a held-out test set against predefined targets. Subordinate objectives follow from this aim and include identifying which modelling choices within LLMs, with GPT adopted as the representative family, offer the most favourable balance between effectiveness and efficiency, characterising data efficiency by examining how performance evolves with training size at the intake stage, and quantifying cost–performance trade-offs so that any improvement in predictive quality can be weighed against additional expenditure or time. Any resulting system is conceived strictly as advisory support for triage under genuine human oversight.

[1] Examples of petitions can be found in the ECtHR HUDOC database under "Communicated Cases": https://hudoc.echr.coe.int/#{%22documentcollectionid2%22:[%22COMMUNICATEDCASES%22]}

1.3. Research contributions

This study contributes to the growing field of AI-assisted legal prediction by:

- Framing ECtHR outcome prediction as an intake-stage, text-only task (petition at filing; no structured metadata), according to constraints for practical triage.
- Applying LLMs to ECtHR petitions through a simplified, cost-aware evaluation setup suitable for reproducible assessment.
- Benchmarking against representative prior NLP (Natural Language Processing) work to situate results under comparable assumptions.
- Assessing performance across model families/variants and input-text representations, with attention to effectiveness-versus-efficiency trade-offs.
- Demonstrating the potential of LLMs for early triage, with evaluation that prioritises recall for the "violation" class given the asymmetric ethical cost of missed cases.
- Embedding an institutional, legal, and ethical perspective, positioning any predictor as advisory under human oversight.

1.4. Structure of the thesis

The remainder of this thesis is organised as follows. Chapter 2 reviews the relevant literature. Chapter 3 describes the methodology, including data preparation, modelling choices, and the evaluation protocol. Chapter 4 presents and discusses the results. Chapter 5 provides an integrated institutional, legal, and ethical discussion situated in the ECtHR context.

2. SYSTEMATIC LITERATURE REVIEW

This chapter presents a review of the literature on judicial decision prediction, examining how various methodologies have been applied in previous research. It begins by contextualising the evolution of LLMs and their relevance to the analysis of legal text. Next, it clarifies the typology of predictive tasks in this field and highlights the specific research gap addressed by this study. Finally, the protocol adopted for the Systematic Literature Review (SLR) is described, including the criteria used to identify, screen, and evaluate relevant publications.

2.1. Overview of research in judicial decision prediction

The prediction of judicial decisions has gained increasing academic and practical relevance, largely due to advances in AI, particularly in NLP (Aletras et al., 2016). As legal data continues to grow exponentially and case analysis becomes more complex (Ariai & Demartini, 2025), the need for automated tools capable of supporting legal professionals and contributing to a more efficient, accessible, and fair justice system has been widely acknowledged (Aguiar et al., 2024). Among recent technological developments, LLMs have revolutionised the way machines understand and generate text (Hagos et al., 2024). However, their application in predictive legal tasks presents new challenges, since LLMs were not initially designed for classification purposes (Ariai & Demartini, 2025).

To contextualise this study, Medvedeva et al. (2022) identified three main types of tasks in judicial decision prediction: (1) Outcome Identification, which involves extracting the verdict from the judgement text; (2) Outcome-Based Judgement Classification, in which a decision is inferred based on the case content without relying on explicit verdict mentions; and (3) Outcome Forecasting, which predicts outcomes using only information available prior to the final decision, such as petitions. This third category is the most complex, as it requires deeper semantic understanding and probabilistic reasoning.

The ability to anticipate court decisions holds significant promise. For legal professionals, predictive models may support case strategy, client advice, and risk assessment. For institutions, such models offer tools to improve transparency, detect bias, and guide policy, ultimately contributing to a more efficient and fair justice system (Lai et al., 2023).

The rise of LLMs marks a turning point in AI research, particularly in tasks involving natural language understanding and generation (Badr, 2024). Powered by transformer architectures, models such as GPT can capture complex linguistic patterns and semantic relationships across long textual contexts (Raza et al., 2025).

LLMs belong to the broader field of deep learning (DL) (Blank, 2023), differing significantly from traditional machine learning (ML) techniques, which often relies on pre-processing and manual feature engineering for unstructured data (Yang, 2024). Although LLMs appear to learn directly from raw text, the input is first tokenised and then converted into numerical vectors before being processed. This transformation is handled automatically by model providers and is generally invisible to end users (Capgemini Applied Innovation Exchange, 2024).

By enabling a more flexible and holistic analysis of legal texts, LLMs allow researchers to move beyond traditional approaches that rely on manually engineered features such as citations patterns or topic models. This shift reflects a broader trend in AI towards increased model autonomy, scalability, and reduced dependence on manual data preparation (Shu et al., 2024).

2.2. Protocol for the Systematic Literature Review

A Systematic Literature Review is intended to identify, evaluate, and synthesise scientific studies relevant to a specific research question, using transparent and structured methods (Kitchenham & Brereton, 2013). In the present study, the SLR aims to organise and assess the current state of knowledge on the use of AI techniques for predicting judicial decisions, with a particular focus on the ECtHR. Given the growing interest in this field, the review provides a rigorous and comprehensive overview of methodologies, applications, and limitations found in the literature.

The review follows the PRISMA 2020 methodology (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), which offers an updated and comprehensive framework for conducting and reporting systematic reviews with transparency and rigour (Page et al., 2021). The protocol presented below details each stage of the review process, from the formulation of research questions to the synthesis and evaluation of the selected studies.

In line with the research objective, this literature review seeks to answer the general question: “How have court decisions been predicted?”

More specifically, it addresses the following three sub-questions:

- RQ1: “What data has been used to predict court decisions?”
- RQ2: “What methodology has been used?”
- RQ3: “What are the contributions, limitations, and suggestions for future studies?”

To maximise the effectiveness of the review, the articles included in this SLR were selected from the Scopus database, given its recognised advantages. This platform is notable not only for its extensive coverage of multidisciplinary publications, but also for the depth and flexibility

of its search capabilities. It enables the use of detailed queries across various research domains and supports Boolean operators such as “AND” and “NOT”, allowing for highly specific keyword combinations. This results in a more precise and relevant selection of articles (Rosili et al., 2021). Such capacity to refine and focus on queries is essential for identifying literature that specifically addresses the intersection between AI and judicial decision prediction.

Although the Web of Science database was initially considered, a preliminary analysis revealed substantial overlap with the results obtained from Scopus. Moreover, Scopus offered a broader range of relevant articles for the scope of this research, particularly after refinement of the search queries. Therefore, Scopus was deemed sufficient and was selected as the sole source for literature retrieval.

The search was conducted by querying the title, abstract, and keywords fields in Scopus. To ensure broad and diverse retrieval, English language was used. The selection of terms such as “prediction”, “court decision”, and “artificial intelligence”, along with their synonyms, was informed by an analysis of frequently used expressions in previous publications, as well as by the researcher’s own legal expertise, ensuring alignment with current terminology at the intersection of jurisprudence and technology.

Using the following search string: ("predict*" OR "forecast") AND ("court decision" OR "legal decision" OR "law decision" OR "case outcome" OR "judicial case" OR "judicial decision") AND ("machine learning" OR "artificial intelligence" OR "AI" OR "neural networks"), a total of 174 documents were retrieved from Scopus. These were subjected to the inclusion and exclusion criteria (Table 2.1), and after a preliminary screening of their abstracts, 22 articles were selected for detailed analysis.

Table 2.1 - Inclusion and exclusion criteria applied in the SLR

Inclusion criteria	Articles or conference papers Articles published from 2020 onwards Articles in English or Portuguese
Exclusion criteria	Articles that do not have an impact factor Duplicate articles Articles without a practical component, which do not address the prediction of court decisions

The search and selection process were conducted between January and February 2024. Although the dissertation was finalised in 2025, no new studies were added after the initial search period to preserve methodological consistency.

To structure the analysis of the selected studies and respond to the research questions, three systematic tables were developed – one for each specific research question. Each table was guided by a set of quality criteria designed to reflect the most relevant aspects to be assessed within each dimension of the review.

The detailed criteria used to assess each research question are presented in the appendices. For RQ1, the focus was on the clarity and detail with which datasets were described, including their sources, legal domain specificity, outcome types analysed, and whether the emphasis was on prediction or classification tasks. For RQ2, attention was given to modelling approaches, including data preparation techniques, types of AI or NLP models, comparative analyses across models, performance evaluation metrics (e.g., accuracy, precision, recall), and the extent to which issues such as explainability and the treatment of false positives and false negatives were addressed. For RQ3, the assessment covered the main contributions of each study, such as ethical considerations, acknowledged methodological limitations, and suggested directions for future research.

Each article included in the review was analysed using these three tables. The criteria were assessed quantitatively, on the basis of their presence in each paper. This structured approach ensures a transparent, replicable, and comprehensive analysis of the state of the art in judicial decision prediction using AI.

The analysis of each article was primarily conducted through a column-based comparison across the constructed tables, allowing for the identification of points of convergence and divergence between studies. Following a comprehensive review, it became necessary to determine which articles exhibited the highest quality and the strongest alignment with the scope of this study. To support this assessment, a set of evaluation criteria was created based on the three research questions. For each criterion, a score was assigned according to whether the article clearly met it (1), partially addressed it (0.5), or did not meet it (0), in line with quality assessment recommendations in the literature (Kitchenham & Charters, 2007). This scoring system enabled a transparent and structured classification of the reviewed studies.

Table 2.2 - Quality assessment criteria used for evaluating the selected studies in the SLR

Question	Quality criteria
What data has been used to predict court decisions?	1. Does it describe the type of data used and, when applicable, the size of training, test and validation sets? 2. Does it perform outcome forecasting? 3. Does it specify whether the data used is of a general nature or focused on a specific legal field? 4. Does it indicate the types of outcomes of the court cases under analysis?
What methodology has been used?	1. Does the article clearly identify its best-performing model and briefly describe how it works? 2. Does it describe how the data was prepared? 3. Does it compare different models? 4. Does it indicate the metrics used to evaluate the results, including but not limited to accuracy? 5. Does it reflect on the impact of false positives and false negatives in the context of judicial prediction? 6. Does it address the explainability of the model or discuss the importance of model interpretability in the legal context?
What are the contributions, limitations and suggestions for future studies?	1. Does it include and discuss ethical considerations, especially regarding the use of data and the application of artificial intelligence models in judicial decisions? 2. What are the main contributions to predicting judicial decisions using AI? 3. Does it recognize and discuss limitations, be they methodological, data or scope? 4. Does it offer any suggestions for future research?

The flowchart below outlines the methodology followed in this research to identify, select, and analyse scientific articles focused on the prediction of court decisions.

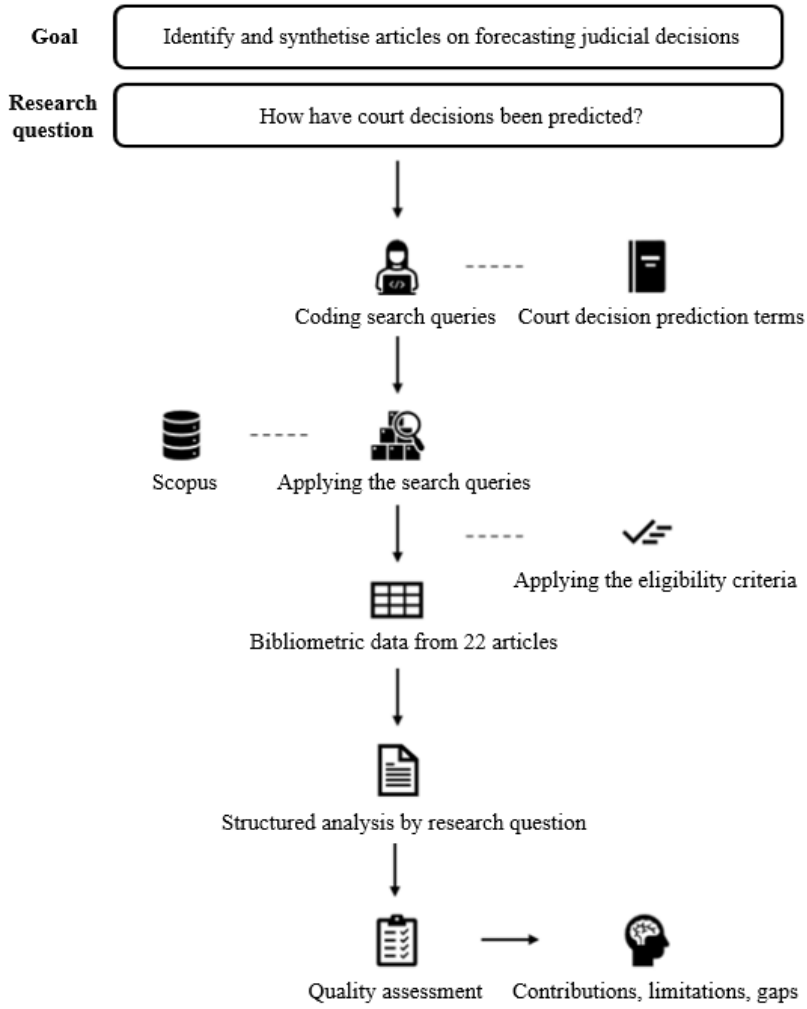

Figure 2.1 - Methodological flowchart of the SLR

2.3. Descriptive overview of selected articles

Following the application of inclusion and exclusion criteria, and after the removal of duplicates, a total of 22 articles were selected for analysis. Table 2.3 presents a bibliometric overview of these studies, including year of publication, authorship, and source.

Table 2.3 - Overview of the 22 studies selected for the SLR

ID	Year	Title	Source	Authors
1	2020	Deep Learning Algorithm for Judicial Judgements Prediction Based on BERT	2020 5th International Conference on Computing, Communication and Security (ICCCS)	Wang, Y.; Gao, J.; Chen, J.
2	2020	Predicting Outcomes of Legal Cases based on Legal Factors using Classifiers	Procedia Computer Science	Shaikh, R. A.; Sahu, T. P.; Anand, V.
3	2020	Predicting the outcome of judicial cases using semantic analysis	2020 IEEE Symposium Series on Computational Intelligence (SSCI)	Pande, R.; Alam, S.

ID	Year	Title	Source	Authors
4	2020	Using machine learning to predict decisions of the European Court of Human Rights	Artificial Intelligence and Law	Medvedeva, M.; Vols, M.; Wieling, M.
5	2020	Verdict Prediction for Indian Courts Using Bag of Words and Convolutional Neural Network	2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)	Pillai, V. G.; Chandran, L. R.
6	2021	A model for predicting court decisions on child custody	PLOS ONE	Muñoz Soro, J. F.; Serrano-Cinca, C.
7	2021	AI Model for Predicting Legal Judgements to Improve Accuracy and Explainability of Online Privacy Invasion Cases	Applied Sciences	Park, M.; Chai, S.
8	2021	Automatic Judgement Forecasting for Pending Applications of the European Court of Human Rights	Proceedings of the Fifth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2021)	Medvedeva, M.; Üstun, A.; Xu, X.; Vols, M.; Wieling, M.
9	2021	Case-level prediction of motion outcomes in civil litigation	Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law (ICAIL)	McConnell, D. J.; Zhu, J.; Pandya, S.; Aguiar, D.
10	2021	Predicting the litigation outcome of ppp project disputes between public authority and private partner using an ensemble model	Journal of Business Economics and Management	Zheng, X.; Liu, Y.; iang, J.; Thomas, L. M.; Su, N.
11	2021	Predicting the supreme court decision on appeal cases using hierarchical convolutional neural network	International Journal of Speech Technology	Sivaranjani, N.; Jayabharathy, J.; Teja, P. C.
12	2022	A hybrid CNN + BILSTM deep learning-based DSS for efficient prediction of judicial case decisions	Expert Systems with Applications	Ahmad, S.; Asghar, M.Z.; Alotaibi, F. M.; Al-Otaibi, Y. D.
13	2022	Artificial Intelligence-Assisted Decision-Making Method for Legal Judgements Based on Deep Neural Network	Mobile Information Systems	Ma, W.
14	2022	Correlation of Language Processing and Learning Techniques for Legal Support System	2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)	Budhiraja, A.; Sharma, K.
15	2022	Legal Judgements Prediction for Canadian Appeal Cases	2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)	Almuslim, I.; Inkpen, D.
16	2022	Predicting Brazilian Court Decisions	PeerJ Computer Science	Lage-Freitas, A.; Allende-Cid, H.; Santana, O.; Oliveira-Lage, L.
17	2022	Predicting the Category and the Length of Punishment in Indonesian Courts Based on Previous Court Decision Documents	Computers	Nuranti, E. Q.; Yulianti, E.; Husin, H. S.
18	2022	Using Artificial Intelligence to Predict Decisions of the Turkish Constitutional Court	Social Science Computer Review	Sert, M. F.; Yıldırım, E.; Haşlak, İ.

ID	Year	Title	Source	Authors
19	2022	Using attention methods to predict judicial outcomes	Artificial Intelligence and Law	Bertalan, V. G. F.; Ruiz, E. E. S.
20	2023	Building accurate legal case outcome prediction models	2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)	Umamaheswari, S.; Aartisha, S.; Kanimozhi, J.; Subashini, R.
21	2023	Joining metadata and textual features to advise administrative courts decisions: A cascading classifier approach	Artificial Intelligence and Law	Mentzingen, H.; Antonio, N.; Lobo, V.
22	2023	Prediction of court decision from Arabic documents using deep learning	Expert Systems	Zahir, J.

As shown in Table 2.3, several of the selected articles were published in the journal Artificial Intelligence and Law, underscoring its relevance as a central venue for research in this domain. In addition, the presence of studies published across a wide range of international conferences demonstrates the growing and widespread interest in the topic of judicial decision prediction. Most publications involve multiple authors, which suggests a collaborative effort in addressing the complexity associated with the application of AI techniques in legal contexts.

2.4. Structured analysis by research question

To address the research question "How have court decisions been predicted?", a structured analysis was conducted based on the three specific sub-questions defined earlier (see 2.2).

2.4.1. "What data has been used to predict court decisions?"

In response to the second research question, a detailed description of the types of data used in the reviewed studies – including legal domains covered, nature of the prediction task, data formats, dataset sizes, and geographic sources – is provided in Table 2.4.

Table 2.4 - Data used to predict court decisions in the selected studies

ID	Jurisdiction	Legal Field	Dataset Description	Outcome Forecasting?	Court Case Outcomes	Study Period
1	China	Criminal Law	Unstructured data 158,117 training, 17,569 validation, 43,922 test	No: Outcome-Based Judgement Classification	MCC + Reg	NA
2	India	Criminal Law	Structured data 86 cases	No: Outcome-Based	Binary classification	2017-2018

ID	Jurisdiction	Legal Field	Dataset Description	Outcome Forecasting?	Court Case Outcomes	Study Period
				Judgement Classification		
3	New Zealand	Employment Law	Unstructured data 218 cases	No: Outcome-Based Judgement Classification	Binary classification	2017
4	International	Human Rights Law	Unstructured data 3,132 training, 8,400 test	No: Outcome-Based Judgement Classification	Binary classification	2014-2017
5	India	Criminal and Civil Law	Unstructured data 400 cases	No: Outcome Identification	Binary classification + Regression	NA
6	Spain	Family Law	Structured data 1,884 cases	Yes	Binary classification	2016-2020
7	USA	Online Privacy Law	Unstructured data 1,098 cases	No: Outcome Identification	Binary classification	2000-2018
8	International	Human Rights Law	Unstructured data 5,456 training, 1,568 test	Yes	Binary classification	2000-2020
9	USA	Civil law	Structured and unstructured data 7,904 cases	Yes	Binary classification	2004-2019
10	China	PPP	Unstructured data 198 cases	No: Outcome-Based Judgement Classification	Binary classification	2013-2018
11	India	Criminal and Civil Law	Unstructured data 20 years of cases	Yes	Binary classification	2000-2019
12	USA	General	Structured data 96,404 training, 24,102 test	No: Outcome-Based Judgement Classification	Multi-class classification	1946-2018
13	China	Criminal Law	Unstructured data	No: Outcome-Based Judgement Classification	Multi-class classification + Regression	NA
14	India	Criminal and Civil Law	Unstructured data	No: Outcome-Based Judgement Classification	Multi-class classification	NA
15	Canada	Criminal and Civil Law	Unstructured data 3,670 cases	No: Outcome-Based Judgement Classification	Binary classification	1982-2021
16	Brazil	General	Unstructured data 3,234 training, 809 test	No: Outcome-Based Judgement Classification	Binary classification	2018-2019

ID	Jurisdiction	Legal Field	Dataset Description	Outcome Forecasting?	Court Case Outcomes	Study Period
17	Indonesia	Criminal Law	Unstructured data 20,367 training, 2.263 test	No: Outcome-Based Judgement Classification	Multi-class classification + Regression	NA
18	Turkey	Constitutional Law	Unstructured data 344 training, 86 test	No: Outcome-Based Judgement Classification	Binary classification	2012-2020
19	Brazil	Criminal Law	Unstructured data 2,467 cases	No: Outcome-Based Judgement Classification	Binary classification	NA
20	India	Criminal Law	Unstructured data 1,017 cases	No: Outcome-Based Judgement Classification	Binary classification + Regression	NA
21	Brazil	Administrative Law	Unstructured data 1,108 cases	Yes	Binary classification	2016-2020
22	Morocco	General	Unstructured data 617 training, 155 test	No: Outcome-Based Judgement Classification	Multi-class classification	NA

The analysis reveals a clear trend: criminal law predominates across the reviewed studies. This is likely due to the binary nature of many criminal decisions (such as guilty vs. not guilty), which lend themselves well to classification tasks (Picinali, 2017). Additionally, criminal cases often receive greater public attention and are more accessible, encouraging data-driven research in this domain. At the same time, the review shows considerable representation across other legal areas, including family law, human rights, constitutional law, employment law, online privacy, and administrative procedures. This diversity indicates that while criminal law remains the primary focus, researchers are increasingly exploring other areas as legal data becomes more widely accessible globally.

Most studies focus on outcome-based judgment classification, which leverages textual features from legal documents to categorise decisions rather than directly predicting verdicts. By contrast, outcome forecasting, which aims to predict case results based solely on pre-judgement materials, remains less common due to limited access to pre-decisional data and the added technical and ethical complexity of predicting outcomes before judgments are rendered (Medvedeva et al., 2022).

In terms of data types, unstructured text dominates, reflecting the reality that legal information is usually embedded in full-length case documents, judicial opinions, and filings. This prevalence necessitates the application of NLP techniques to extract meaningful features

(Frankenreiter & Nyarko, 2023). The predominance of binary classification aligns with the dichotomous nature of many legal outcomes. Although dataset sizes vary, most studies rely on corpora exceeding 1,000 cases, reflecting a preference for datasets large enough to support meaningful model training and evaluation.

Geographically, India, China, and the United States of America are the most represented countries. The prominence of China and the USA likely reflects both the availability of legal data and the maturity of research infrastructures in these regions (Schmallenbach et al., 2024), while India's strong presence may be attributed to its emerging role as a centre for AI research (Chahal et al., 2021). Some studies extend their scope to less commonly analysed jurisdictions, such as Morocco and Turkey, indicating a growing interest in global perspectives.

2.4.2. "What methodology has been used?"

To answer the second question, the models, techniques, evaluation metrics, and additional methodological elements are detailed in Table 2.5.

Table 2.5 - Modelling approaches and evaluation criteria in the selected studies

ID	Best Performing Model	Data Preparation	Comparison of Models	Evaluation Metrics	FN and FP Impact	Model Explainability
1	BERT + FC	Word segmentation; POS-tagging; stop-word removal; TF-IDF vectorization	Yes	Precision: 89,71% Recall: 89,79% F1-Score: 89,64%	Only define	No
2	CART	Manual extraction of 19 legal-factor features; min-max normalization of the features	Yes	Accuracy: 91,86% Precision: 92,86% Recall: 90,70% F1-Score: 91,76	Only to compute metrics	Reports per-feature statistics but offers no deeper interpretability discussion
3	CapsNet	Paragraph extraction; manual labelling; keyword-based feature selection; cosine similarity; LDA topic-clusters; tokenization; word-embedding vectorization	Yes	Accuracy: 75,19%	No	No
4	SVM	Stopword and capitalization removal; TF-IDF	No	Accuracy: 75%	No	Coefficient visualization of top n-grams

ID	Best Performing Model	Data Preparation	Comparison of Models	Evaluation Metrics	FN and FP Impact	Model Explainability
		n-grams; tokenization				
5	CNN	Tokenization; stop-word removal; BoW keyword extraction; word-embedding vectorization	No	Accuracy: 85%	No	Uses keyword inspection of true/false cases
6	MLP neural network	Manual labelling; numeric counts of factual elements and legal principles	Yes	Accuracy: 86,4%	No	Decision-tree rules extracted to reveal winning/losing patterns, emphasising model interpretability for lawyer
7	CART	Keyword extraction; NTA to map privacy-invasion factors in legal context	Yes	Accuracy: 83,16%	No	Network-text analysis visualises how privacy-invasion factors drive judgements, offered to keep the model explainable
8	H-BERT	Tokenization; stop-word and capitalization removal; TF-IDF vectorization	Yes	Precision: 67% Recall: 67% F1-Score: 67%	Only to compute metrics	Inspects SVM n-gram coefficients and notes need for explainable models
9	Adaboost	TF-IDF vectorization; attorney case-entropy feature	Yes	Accuracy: 64,4%	No	Feature-importance analysis explains which n-grams and attorney features drive predictions; authors emphasise need for transparent, explainable models in legal AI
10	Ensemble Model (GBDT + kNN + MLP)	Manually extract of legal factors and encode each as binary features	Yes	Accuracy: 96,42% Precision: 96,66% Recall: 96,38% F1-Score: 96,03%	Only to compute metrics	Explains outcomes through the 17 legal factors
11	H-CNN	Word-embedding vectorization	Yes	Accuracy: 81,13%	No	Lists most influential parameters
12	CNN + BiLSTM	Chi-square test for feature selection	Yes	Accuracy: 91,52% Precision: 91.74% Recall: 89.04% F1-Score: 90.44%	Only to compute metrics	No

ID	Best Performing Model	Data Preparation	Comparison of Models	Evaluation Metrics	FN and FP Impact	Model Explainability
13	BERT12multi	Word-embedding vectorization	Yes	F1-Score: 82,2%	No	No
14	M-AttBLSTM-CNN (TextCNN + Att-BLSTM)	Stop-word and suffix removal; text conversion into features; TF-IDF vectorization	Yes	F1-Score: 94,62%	No	Proposes future inclusion of rationale
15	RCNN	Stop-word, capitalization, citations, number and punctuation removal; lemmatization; word-embedding vectorization	Yes	Accuracy: 93,46% Precision: 91,4% Recall: 93,5% F1-Score: 92,4%	Only to compute metrics	Notes “black-box” issues and plans future attention-based explanations
16	XGBoost	Stemming; stop-word, capitalization and punctuation removal; TF-IDF vectorization	Yes	F1-Score: 80,22% Accuracy: 81,35% Precision: 81,08% Recall: 79,74%	Only to compute metrics	No
17	CNN+Attention	Token normalisation; section annotation into 10 features; word-embedding vectorization	Yes	F1-Score: 76,81% Precision: 77,08% Recall: 77,36% Accuracy: 77,32%	Only to compute metrics	Ablation shows which sections drive accuracy, providing limited insight
18	MLP neural network	TF-IDF vectorization; stop-word, capitalization, special characters and punctuation removal; tokenization	No	Accuracy: 90%	Only to compute metrics	No
19	HAN	Stop-word removal; tokenization; word-embedding vectorization	Yes	Accuracy: 99,78% F1-Score: 98,55% Recall: 99,32% Precision: 98,58%	Only to compute metrics	Hierarchical-attention network weights plotted; paper lists top words driving convictions vs acquittals to keep the model transparent
20	Naïve Bayes	ND	Yes	Accuracy: 89%	No	No
21	Cascade RF+XGBoost	Ubiquitous and non-discriminative words and stop-	Yes	F1-Score: 90% Precision: 92,9%	Only to compute metrics	Uses SHAP to visualise topic and metadata contributions for individual

ID	Best Performing Model	Data Preparation	Comparison of Models	Evaluation Metrics	FN and FP Impact	Model Explainability
		word removal; POS tagging; tokenization; stemming; text summarization; stop words removal; vectorization		Recall: 87,3%		predictions, explicitly aiming to "preserve explainability" in legal use
22	CNN	Stop-word, punctuation, digits, extra white spaces and tabulations removal; word-embedding vectorization	Yes	Accuracy: 80,51%	No	Notes deep models are "black boxes"

There is a clear predominance of deep learning (DL) approaches across the reviewed studies. Traditional machine learning (ML) algorithms – such as CART, AdaBoost, and XGBoost – remain in use, primarily in research that relies on structured or manually engineered features. This trend reflects DL architectures' superior capacity to process and understand complex, unstructured legal texts (Johnson et al., 2025).

Preprocessing of legal documents is critical due to both the length and complexity of texts in this domain and the presence of specialised legal vocabulary. Common preprocessing steps include vectorisation, stop-word removal, lemmatisation, tokenisation, and the transformation of legal concepts into structured features. Several studies also employ additional strategies – such as paragraph segmentation or rule extraction – to enhance input quality and support model interpretability (Keeling et al., 2020).

Most articles conduct comparative evaluations of multiple models, emphasising benchmarking as standard practice (Guha et al., 2024). Accuracy, precision, recall, and F1-score are the most prevalent evaluation metrics. High performance is generally reported for outcome-based classification tasks, whereas studies focused on outcome forecasting tend to show lower results, reflecting the increased uncertainty and complexity of predicting legal outcomes before judgments are rendered (Medvedeva & McBride, 2023).

While performance evaluation and benchmarking receive considerable attention, analyses of model interpretability and the consequences of errors are less developed. Where interpretability is addressed, common techniques include feature importance analysis, SHAP (SHapley Additive exPlanations) values, and, in some cases, visualisations of influential parameters or rule extraction methods. These approaches help clarify how models arrive at

decisions and promote greater transparency. Nonetheless, it remains rare for studies to systematically evaluate the practical or ethical significance of false positives and false negatives. This highlights a pressing need for more robust interpretability frameworks and stronger integration of ethical reflection into methodological practice.

2.4.3. "What are the contributions, limitations, and suggestions for future studies?"

Finally, to address the third question, the main findings and recommendations are presented in Table 2.6.

Table 2.6 - Contributions, limitations and suggestions identified in the selected studies

ID	Ethical Considerations	Contributions	Limitations	Suggestions
1	-	Decision support for legal professionals.	Dataset limitations.	Model improvement and architecture exploration.
2	-	Decision support for legal professionals.	Dataset limitations; Manual or noisy data processing; Performance issues (on multi-accused cases).	Feature engineering and data representation; Generalization and domain expansion.
3	-	Decision support for legal professionals.	Dataset limitations; Manual or noisy data processing; Performance issues.	Model improvement and architecture exploration; Dataset enhancement and scaling; Real-world deployment and usability.
4	-	Contextual and extralegal awareness.	Performance issues (when predicting truly future cases); Legal or semantic oversimplification; Generalization constraints.	Feature engineering and data representation; Generalization and domain expansion; Model improvement and architecture exploration.
5	-	Fairness, impartiality and transparency.	Dataset limitations; Manual or noisy data processing; Performance issues (in multi-charge scenarios); Legal or semantic oversimplification.	Dataset enhancement and scaling; Feature engineering and data representation; Model improvement and architecture exploration.
6	-	Decision support for legal professionals; Efficiency and cost reduction.	Manual or noisy data processing; Generalization constraints.	-
7	Trade-off between performance and transparency.	Legal risk management and compliance support; Fairness, impartiality and transparency.	Generalization constraints.	Generalization and domain expansion.

ID	Ethical Considerations	Contributions	Limitations	Suggestions
8	-	Fairness, impartiality and transparency.	Performance issues (across models and years); Generalization constraints; Lack of explainability or interpretability.	Feature engineering and data representation; Generalization and domain expansion; Explainability and interpretability; Model improvement and architecture exploration.
9	Need for transparency and explainability to ensure accountability; Ethical, fair, and non-discriminatory AI in legal contexts.	Decision support for legal professionals; Fairness, impartiality and transparency.	Generalization constraints.	Generalization and domain expansion; Feature engineering and data representation; Explainability and interpretability.
10	-	Decision support for legal professionals; Efficiency and cost reduction.	Dataset limitations; Manual or noisy data processing; Performance issues (for cases with multiple accused and mixed outcomes).	Feature engineering and data representation; Generalization and domain expansion.
11	-	-	Generalization constraints.	Generalization and domain expansion.
12	-	Decision support for legal professionals; Efficiency and cost reduction.	Dataset limitations.	Model improvement and architecture exploration; Generalization and domain expansion; Feature engineering and data representation.
13	-	Efficiency and cost reduction.	-	Explainability and interpretability; Real-world deployment and usability.
14	-	Decision support for legal professionals.	Dataset limitations; Manual or noisy data processing.	Feature engineering and data representation; Generalization and domain expansion.
15	-	Research and jurisprudential consistency.	Manual or noisy data processing; Lack of explainability or interpretability.	Model improvement and architecture exploration; Explainability and interpretability.
16	-	Decision support for legal professionals; Efficiency and cost reduction; Fairness, impartiality and transparency.	Manual or noisy data processing.	Dataset enhancement and scaling; Real-world deployment and usability; Feature engineering and data representation; Model improvement and architecture exploration.
17	-	-	Performance issues (for the 'heavy' category); Manual or noisy data processing.	Feature engineering and data representation.

ID	Ethical Considerations	Contributions	Limitations	Suggestions
18	-	Efficiency and cost reduction.	Dataset limitations; Legal or semantic oversimplification; Performance issues (because the model relies on post-decision content).	Model improvement and architecture exploration; Feature engineering and data representation; Dataset enhancement and scaling.
19	-	Decision support for legal professionals; Contextual and extralegal awareness.	Generalization constraints.	Model improvement and architecture exploration; Dataset enhancement and scaling.
20	-	Fairness, impartiality and transparency.	-	Model improvement and architecture exploration; Real-world deployment and usability; Explainability and interpretability.
21	Limitations of global explainability and human interpretability.	Decision support for legal professionals.	Dataset limitations; Lack of explainability or interpretability.	Model improvement and architecture exploration; Feature engineering and data representation; Explainability and interpretability; Real-world deployment and usability.
22	Need for transparency and explainability to ensure accountability.	Efficiency and cost reduction.	Dataset limitations; Lack of explainability or interpretability.	Dataset enhancement and scaling; Real-world deployment and usability; Explainability and interpretability.

The review shows that the main contributions centre on providing decision-support tools for legal professionals, enhancing efficiency, and reducing costs by automating repetitive or preliminary judicial tasks (McConnell et al., 2021). Several studies also advance fairness, impartiality, and transparency by reducing human subjectivity and bias and promoting more consistent jurisprudential outcomes (Soro & Serrano-Cinca, 2021). In some cases, contributions extend to contextual or extralegal awareness and improved consistency in legal reasoning, highlighting the growing sophistication of computational models.

Despite these advances, important limitations persist. The most recurrent issues relate to dataset quality and coverage, including imbalanced samples, manual or noisy data processing, and annotation inconsistencies, all of which compromise model generalisability and reliability (Shaikh et al., 2020). Other frequent limitations include performance drops in multi-accused or multi-charge cases, challenges in predicting truly future cases, and legal or semantic oversimplification. Moreover, a lack of explainability and interpretability remains a significant

concern, particularly when deep learning models operate as “black boxes” and the ethical implications of erroneous outputs remain insufficiently addressed.

Suggestions for future research consistently highlight the need for model improvement and architectural exploration, more robust feature engineering, and expansion and scaling of datasets. Many studies recommend automating labelling and annotation processes to reduce noise and improve reproducibility, while others call for domain expansion to test models across jurisdictions or temporal periods. Another recurring recommendation is to strengthen explainability and interpretability mechanisms – such as using feature-importance analyses, SHAP values, or transparent model designs – to support accountability, ethical oversight, and real-world usability (Park & Chai, 2021; Mentzingen et al., 2023). Finally, several authors emphasise the need for deployment-oriented research, integrating usability assessments to bridge the gap between academic prototypes and practical legal applications.

2.5. Quality assessment and critical insights

In conclusion of the review process, a quality assessment was conducted using a set of predefined criteria aligned with each research question. This step not only enabled the identification of the most complete and informative studies but also allowed recurring weaknesses and research gaps across the literature to be recognised.

The results, summarised in Table 2.7, show scores ranging from 8 to 13, with an average of 10.5. This indicates generally solid methodological practices, though variability in comprehensiveness was observed. A few works – particularly Medvedeva et al. (2021), McConnell et al. (2021), and Mentzingen et al. (2023) – achieved the highest scores and can be considered methodological benchmarks.

Table 2.7 - Quality assessment assigned to each study

ID	RQ1				RQ2						RQ3				Total
	1	2	3	4	1	2	3	4	5	6	1	2	3	4	
1	1	0.5	1	1	1	1	1	1	0.5	0	0	1	1	1	**11**
2	0.5	0.5	1	1	1	1	1	1	0.5	0.5	0	1	1	1	**11**
3	0.5	0.5	1	1	1	1	1	0.5	0	0	0	1	1	1	**9.5**
4	1	0.5	1	1	1	0	1	0.5	0	1	0	1	1	1	**10**
5	0.5	0	1	1	1	1	0	0.5	0	0.5	0	1	1	1	**8.5**
6	0.5	1	1	1	1	1	1	0.5	0	1	0	1	1	0	**10**
7	0.5	0	1	1	1	1	1	0.5	0	1	1	1	1	1	**11**
8	1	1	1	1	1	1	1	1	0.5	0.5	0	1	1	1	**12**
9	0.5	1	1	1	1	1	1	0.5	0	1	1	1	1	1	**12**

ID	RQ1				RQ2						RQ3				Total
	1	2	3	4	1	2	3	4	5	6	1	2	3	4	
10	0.5	0.5	1	1	1	1	1	1	0.5	0.5	0	1	1	1	**11**
11	0.5	1	1	1	1	1	1	0.5	0	0.5	0	0	1	1	**9.5**
12	1	0.5	1	1	1	1	1	1	0.5	0	0	1	1	1	**11**
13	0	0.5	1	1	1	1	1	1	0	0	0	1	0	1	**8.5**
14	0	0.5	1	1	1	1	1	1	0	0	0	1	1	1	**9.5**
15	0.5	0.5	1	1	1	1	1	1	0.5	0	0	1	1	1	**10.5**
16	1	0.5	1	1	1	1	1	1	0	0	0	1	1	1	**10.5**
17	1	0.5	1	1	1	1	1	1	0.5	0.5	0	0	1	1	**10.5**
18	1	0.5	1	1	1	1	0	0.5	0.5	0	0	1	1	1	**9.5**
19	0.5	0.5	1	1	1	1	1	1	0.5	1	0	1	1	1	**11.5**
20	0.5	0.5	1	1	1	0	1	0.5	0.5	0	0	1	0	1	**8**
21	0.5	1	1	1	1	1	1	1	0.5	1	1	1	1	1	**13**
22	1	0.5	1	1	1	1	1	0.5	0	0	1	1	1	1	**11**

For RQ1 (data), the strongest studies provided transparent dataset descriptions, including comprehensive details on training, validation, and test splits, as well as clear definitions of outcome categories (Zhong et al., 2020). Such methodological transparency is crucial for reproducibility and allows for meaningful comparisons across studies. However, many works lack this level of detail; in particular, information regarding data preprocessing, selection criteria, or sample representativeness is often limited, which can undermine the credibility of findings.

Outcome forecasting remains underexplored, as most articles focus on retrospective outcome-based classification rather than predictive tasks based on pre-judgement materials (Medvedeva et al., 2022). Pre-judgement data are often incomplete, and models face difficulties in making reliable predictions from them. Additionally, attempting to forecast outcomes raises extra technical and ethical challenges, such as potential bias and the difficulty of ensuring fairness. These issues explain why research in this area is still relatively rare, leaving substantial opportunities for future investigation.

Regarding RQ2 (methodology), studies that provided detailed documentation of preprocessing steps and engaged in comparative benchmarking received higher methodological evaluations (Mentzingen et al., 2023). A minority of studies also implemented interpretability techniques – such as feature importance measures or SHAP values (Keeling et al., 2020) – to reveal which data attributes most influenced predictions and to provide case-specific

explanations. The growing use of SHAP in legal text analysis reflects a broader move toward transparent AI, exposing how models weigh different legal arguments or facts.

Despite these interpretability advances, most studies offered only limited assessments of the ethical and practical implications of false positives and false negatives. Few works systematically examined how classification errors could translate into unfair outcomes or biased recommendations. The integration of fairness metrics or comprehensive error analysis remains the exception rather than the rule. This persistent gap underscores the need for future research to more thoroughly address the consequences of predictive errors, ensuring that advances in legal AI are accompanied by strong interpretability frameworks and ethical oversight – consistent with the growing recognition of these issues in recent literature (Dressel & Farid, 2021).

For RQ3 (contributions, limitations, and suggestions), most studies articulated clear contributions in terms of efficiency, fairness, and reduced judicial burden (Soro & Serrano-Cinca, 2021; Budhiraja & Sharma, 2022). However, these claims were often reported descriptively rather than supported through systematic evaluation, which affects the robustness of the contributions reported. Ethical considerations were also addressed unevenly across studies (Park & Chai, 2021; Zahir, 2022), which contributed to lower quality scores in several cases.

Suggestions for future work tended to be high-level and conceptual, with few studies outlining concrete methodological steps or providing evidence-based justifications for their recommendations. Across the literature, there is a consistent call for more rigorous ethical assessment, clearer methodological rationale, and stronger integration of interpretability into predictive modelling. These gaps highlight significant opportunities for research that unifies methodological transparency, fairness evaluation, and explainability within a coherent predictive framework.

Overall, the assessment shows that, although judicial decision prediction displays methodological maturity in places, critical gaps persist – especially in outcome forecasting (RQ1), model interpretability and error analysis (RQ2), and ethical considerations (RQ3).

3. METHODOLOGY

To structure the development of this project, the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology was adopted (Schröer et al., 2021). This well-established framework guides ML workflows through six interconnected phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. In the present study, the methodology was adapted, with the inclusion of an additional intermediate stage. Positioned between Data Preparation and Modelling, this stage involved systematic experimentation with prompts, including the design of summarisation and prediction prompts and the comparison between summaries and full texts as model inputs. Such an extension was required to accommodate the specific characteristics of LLMs, which rely heavily on prompt design for effective performance.

Figure 3.1 presents the general methodological workflow followed in this project. Starting from the initial dataset, the cases were first examined and processed through filtering and outlier removal to prepare the data (see 3.3**Error! Reference source not found.**). In the subsequent s tage, prompt experiments were conducted (see 3.4). Building on these results, LLMs were fine-tuned with cumulative datasets (see 3.5), and the performance of the resulting models was evaluated (see 4.3).

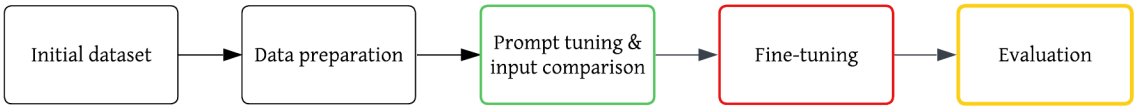

Figure 3.1 - Methodological workflow adopted in the project

3.1. Business Understanding

This subchapter defines the problem addressed by this study and outlines its legal and institutional context. In accordance with the CRISP-DM framework, the Business Understanding phase aims to clarify the project's goals from a domain perspective and translate them into a data science problem. The objective is to assess whether LLMs can be employed to forecast the outcome of petitions submitted to the ECtHR, relying solely on the text of the initial complaint. The following subchapters provide an overview of the Court's role and procedures, the rationale for employing AI in this context, and the specific success criteria guiding the project.

3.1.1. Institutional and procedural context of the ECtHR

The ECtHR is an international court established by the European Convention on Human Rights (ECHR). It is responsible for ensuring that the member states of the Council of Europe respect the rights and freedoms guaranteed under the Convention. Individuals who believe that their rights have been violated by a member state may lodge a complaint (referred to as a petition) before the ECtHR, after exhausting all domestic remedies (ECtHR, 2021).

The Court plays a central role in safeguarding human rights across Europe and has accumulated a vast body of jurisprudence since its establishment in 1959. With 46 member states and thousands of petitions submitted annually, the ECtHR faces a substantial and growing workload, which places pressure on its capacity to deliver timely and consistent decisions. In 2024, approximately 28,800 applications were allocated to a judicial formation, while over 60,000 applications remained pending before the Court by the end of the year (ECtHR, 2025).

Each case submitted to the ECtHR undergoes a multi-stage review process. Initially, a case is screened for admissibility, ensuring that procedural requirements are met (e.g., exhaustion of national remedies, timely submission). Admissible cases are then examined on the merits, and a decision is issued regarding whether there has been a violation of the Convention (ECtHR, 2023).

Cases are typically examined by a chamber of seven judges, and decisions are primarily based on written submissions rather than oral hearings (Council of Europe, 2024). The final judgements include a legal reasoning section that interprets the relevant articles of the Convention and applies them to the facts of the case. The process is inherently complex, and judicial discretion plays a significant role in interpreting the Convention's provisions, particularly in borderline cases (Globalex, 2024).

3.1.2. Problem statement and value of predictive solution

The increasing volume of petitions received by the ECtHR has led to backlogs and significant delays in judgments (European Law Institute, 2024), raising concerns among scholars and practitioners about consistency and the risk of divergent outcomes across similar cases (Wildhaber, 2019). Against this backdrop, the domain objective is to support early triage by flagging petitions for potential rights violations, thereby helping allocate judicial attention and reduce delays – while fully preserving judicial autonomy and human oversight.

Operationally, the task was approached as a binary text-classification problem, where the initial petition serves as the sole input and the model predicts whether the outcome will involve a violation or not. Given the asymmetry of errors in this domain, particular emphasis was placed

on recall for the violation class, to reduce the likelihood of overlooking genuine human-rights breaches. The models were evaluated on more recent cases, approximating a realistic predictive setting.

The system is intended to support early triage by flagging petitions that may involve rights violations, so that resources can be allocated more efficiently in legal aid and advocacy contexts. Additional value is provided to researchers, as decision patterns and consistency can be analysed systematically. The tool is conceived strictly as decision support, with judicial autonomy and human oversight preserved. Ethical risks – such as bias and limits in capturing judicial discretion – are acknowledged and discussed in Chapter 5.

3.1.3. Project success criteria

The success of this project is evaluated in comparison with previous work on ECtHR outcome forecasting. Medvedeva et al. (2021) achieved 67% precision, 67% recall, and 67% F1-score using a training set of over 2,000 cases and a traditional NLP pipeline.

The objective of this study is to achieve comparable or superior performance using simplified architecture based on LLMs, without relying on handcrafted features or structured inputs. Beyond numerical metrics, ethical considerations are prioritised by focusing on sensitivity to the violation class, where failing to detect a true violation could have significant human rights implications.

Success is defined as:

- Achieving an F1-score and accuracy equal to or above the 67% baseline reported by Medvedeva et al. (2021).
- Ensuring high recall – defined here as above 70% – to improve upon the 67% recall reported in the same study for the "violation" class, thereby increasing the likelihood of correctly identifying critical cases and supporting triage.
- Providing some degree of explainability of the model's predictions through post-hoc analysis.

3.2. Data Understanding

The dataset used in this project was obtained from the public repository released together with the paper by Medvedeva et al. (2021), presented at the International Conference on Artificial Intelligence and Law (ICAIL) 2021. It comprises 2,598 cases from the ECtHR, decided between 2006 and 2019, each including structured metadata and the full unstructured text of the

applicant's petition. The final year (2019) likely reflects the availability of decisions at the time the dataset was compiled for that study.

As the petition is submitted at the initial stage of the process, before any judgement is made, the text does not contain information about the Court's final decision. This ensures there is no risk of data leakage from the target variable into the model input. Only the petition text is used to fine-tune the models. The target variable, indicating whether the Court found a violation of the ECtHR (1) or not (0), is provided as a separate structured label in the dataset.

3.2.1. Case structure and target variable

Each case record combines two components: (i) structured metadata, such as country, application number, and articles invoked, and (ii) the unstructured text of the applicant's petition. In this project, only the unstructured petition text is used as input to the predictive model. Structured variables are excluded because this information is present in the text itself. Including both could introduce redundancy without adding meaningful value to the model.

The target variable is provided in a separate structured field and encodes the binary outcome of the case: violation (1) or no violation (0). The dataset is naturally balanced, with 1,299 cases in each class.

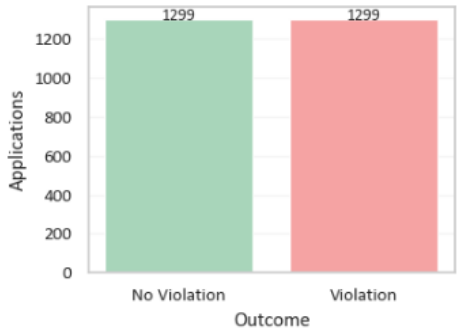

Figure 3.2 - Distribution of cases outcome

A detailed description of the dataset's structured variables is provided in Appendix A.

3.2.2. Geographic distribution of outcomes

To assess geographic patterns, the distribution of judgements by respondent country was analysed. Figure 3.3 shows the number of judgements per country, with Russia, Ukraine, Croatia, and Turkey being the most frequent respondents. However, this distribution is

influenced by factors such as population size, political context, and duration of membership in the Council of Europe.

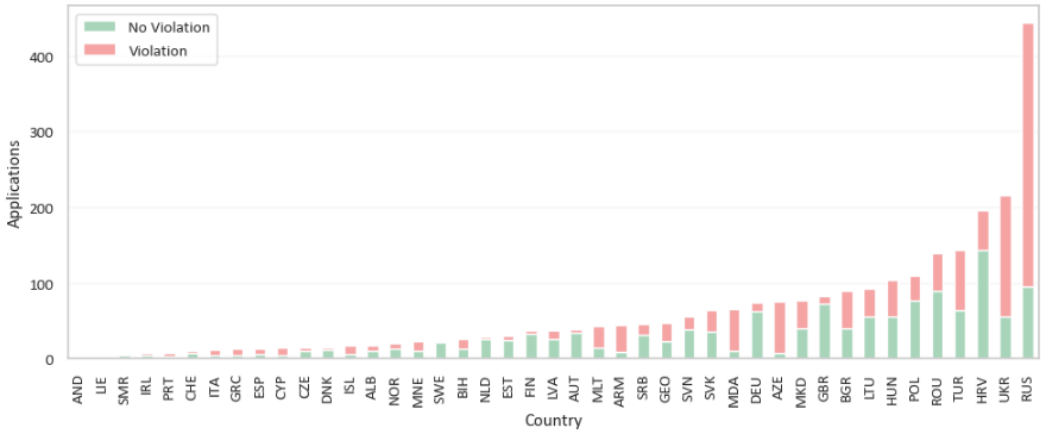

Figure 3.3 - Absolute distribution of violation and no violation outcomes across countries

To better understand the geographic distribution of outcomes, Figure 3.4 presents the violation rate per country. A choropleth map visually summarises disparities across Europe, highlighting countries with higher or lower proportions of violations. These results underscore the potential for implicit geographic bias in the model, even when the country is not explicitly used as an input feature.

Figure 3.4 - Choropleth map showing the violation rate (%) by country

3.2.3. Temporal analysis of applications and judgements

Figure 3.5 shows the evolution in the volume of ECtHR applications over time. The number of applications rises after 2012, peaks around 2017, and then gradually decreases in subsequent years.

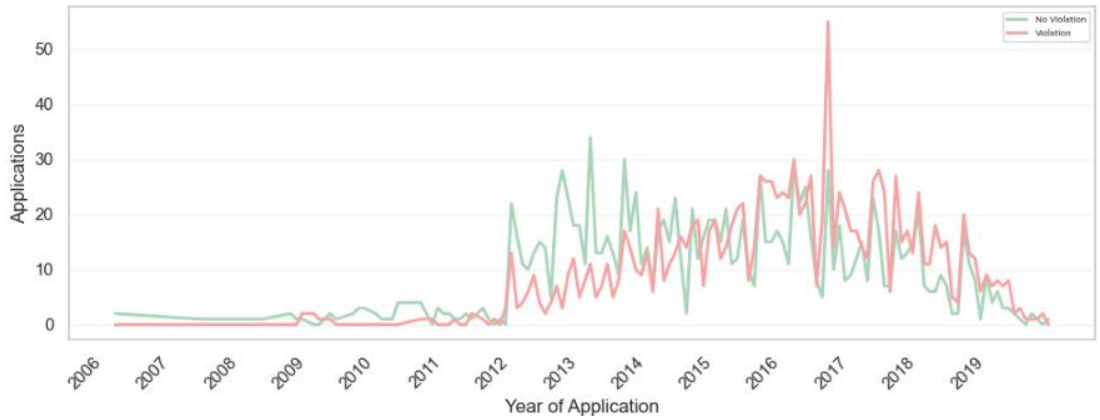

Figure 3.5 - Temporal distribution of applications by outcome

To examine how case duration has evolved, the number of days between the application date and the judgement date was calculated for each case. Figure 3.6 shows the average duration of proceedings per year, separated by outcome.

Overall, the duration of ECtHR proceedings has increased over time for both categories. Notably, cases resulting in a violation tend to take longer than those resulting in non-violation. This pattern may reflect greater complexity or sensitivity in violation cases, further justifying the need for early triage tools that can help prioritise such cases efficiently.

It should be noted that, while Figure 3.5 presents the number of applications by outcome since 2006 (based on the year of submission), Figure 3.6 focuses on the average case duration measured by judgement year. In the dataset, judgement dates for violation cases are only available from 2016 onwards, which explains why violation cases only appear in the duration plot from that year, even though they exist in earlier application years.

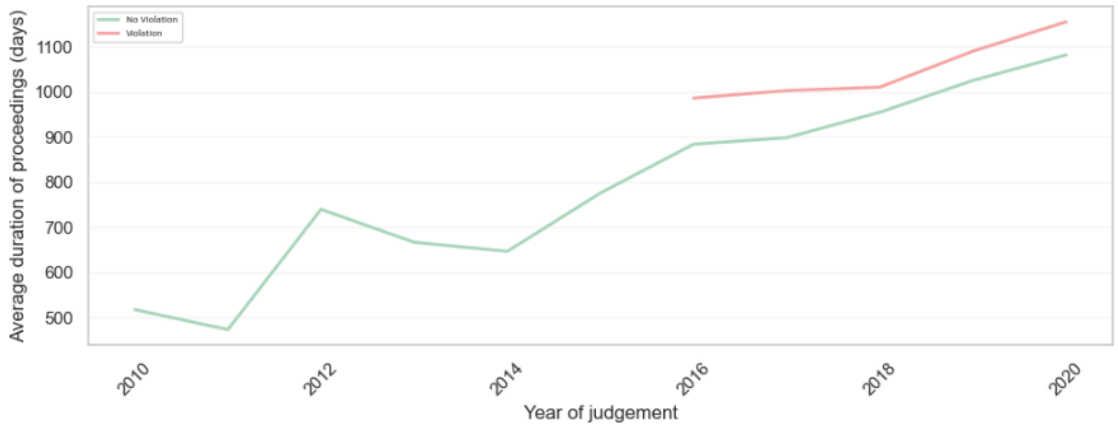

Figure 3.6 - Average duration (days) of proceedings by outcome and judgement year

3.2.4. Petition length and textual patterns

Figure 3.7 shows the distribution of word counts by outcome, with some petitions exceeding 10,000 words. Based on the distribution, outliers were identified at 4,452 words, which corresponds approximately to the upper whisker in both classes. This threshold was later used as a reference in the data preparation phase to filter extremely long cases.

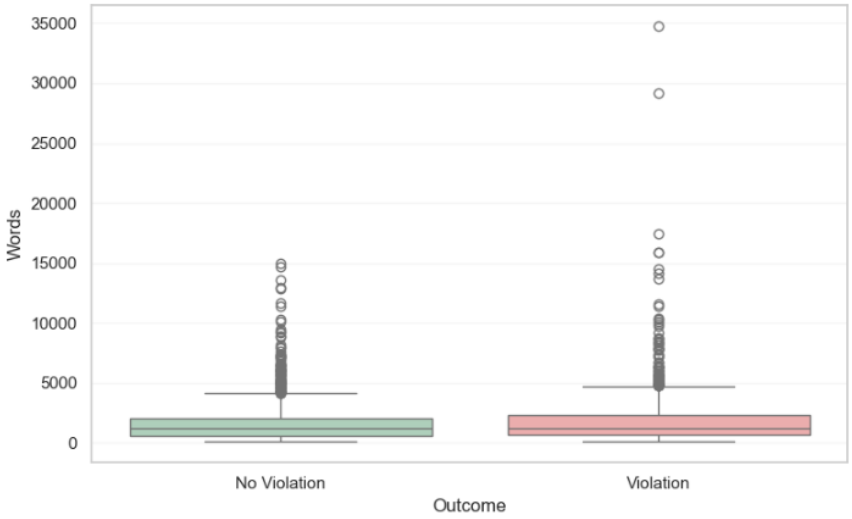

Figure 3.7 - Distribution of petition words count by outcome

To investigate linguistic differences between outcomes, a TF-IDF analysis of bigrams in petition texts was conducted. Figure 3.8 and Figure 3.9 display the most characteristic bigrams for each outcome class.

Petitions that resulted in non-violation often contain formal and procedural expressions, such as *appeal points*, *public prosecutor*, and *relevant domestic*. In contrast, violation cases tend to include language linked to factual circumstances or potential rights violations, such as *police officers*, *ill treatment*, and *conditions detention*. These patterns suggest that LLMs might capture meaningful distinctions between legal reasoning and the nature of allegations, even without explicit labels.

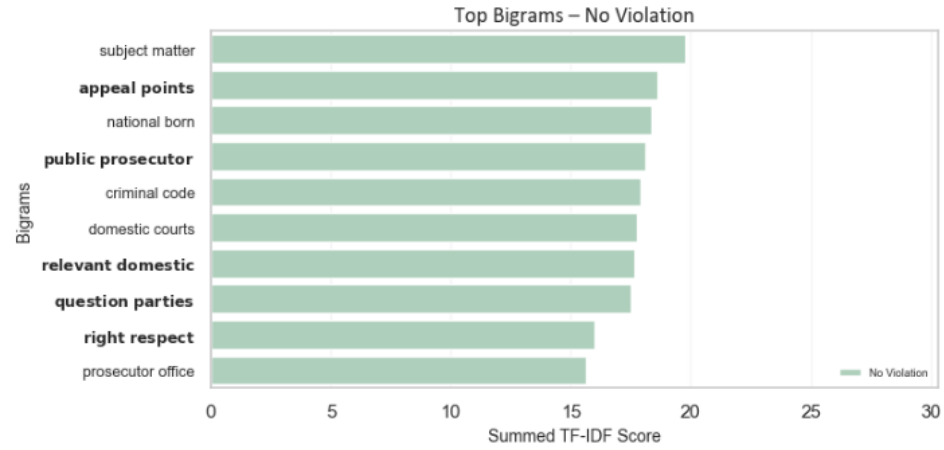

Figure 3.8 - Top bigrams in non-violation cases based on summed TF-IDF scores

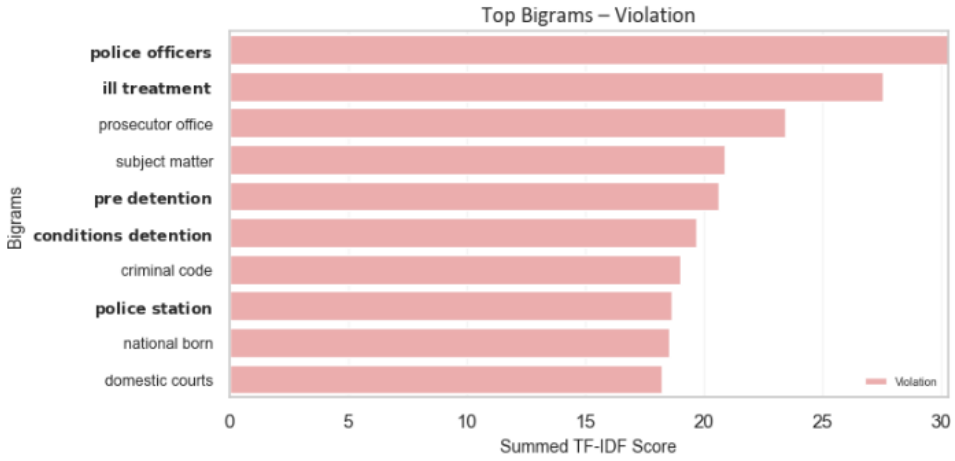

Figure 3.9 - Top bigrams in violation cases based on summed TF-IDF scores

Finally, a topic model based on Latent Dirichlet Allocation (LDA) (Table 3.) was applied to uncover latent themes in the petition texts, separately for violation and non-violation cases. In violation cases, topics linked to detention or police violence are prominent, often involving terms such as *detention*, *treatment*, *police*, and *prison*. Notably, one topic includes *Russia* alongside *investigation* and *abduction*, suggesting that geographic context may be embedded in the language, reinforcing earlier concerns about potential bias. In contrast, non-violation topics include more procedural and family-related themes, with frequent terms like *children*, *mother*, *care*, and *family*, reflecting a different legal context and tone. These distinctions further indicate that the linguistic patterns in petitions reflect underlying legal and contextual differences between outcomes.

Table 3.1 - LDA-generated topics with top keywords in violation and non-violation cases

Topic	Violation	No Violation
1	detention, appeal, 12, 11, 13, 10, december, october, domestic, 15	child, family, children, rights, right, father, order, social, mother, care
2	appeal, judgement, child, public, administrative, right, rights, hearing, criminal, supreme	criminal, appeal, prosecutor, judgement, supreme, investigation, time, hearing, domestic, constitutional
3	police, criminal, detention, treatment, investigation, prosecutor, medical, officers, ill, prison	company, property, act, state, land, judgement, rights, claim, 10, appeal
4	property, land, act, company, 12, state, compensation, public, authorities, eur	detention, prison, treatment, medical, criminal, hospital, police, order, person, appeal
5	investigation, police, criminal, 11, ms, russia, office, abduction, born, administrative	evidence, police, criminal appeal, information, public, act, person, right, order

3.3. Data preparation

The data preparation stage was limited to a single filtering step, aimed at removing extreme outliers in petition length. This ensured that the dataset was consistent and computationally manageable before proceeding with the subsequent prompt experiments (see 3.4).

During the exploratory analysis (see 3.2), it was observed that some cases exceeded 5,000 words, with some surpassing 10,000. Such extremely long texts were not only rare but also introduced significant computational costs during summarisation and fine-tuning.

As the summarisation process was also performed using GPT and incurred a cost proportional to input size, it was decided that a reasonable upper bound should be defined. Based on the distribution of word counts and a visual inspection of outliers via boxplot, petitions containing more than 4,452 words – the threshold before outliers appeared – were excluded. Consequently, 162 cases (6.2% of the original 2,598) were removed, leaving a final dataset of 2,436 petitions for the subsequent prompt experiments and modelling.

3.4. Prompt experiments

Figure 3.10 illustrates the stage of prompt experiments, situated between data preparation and modelling. This stage was designed to select the prompts for summarisation (see 3.4.1) and prediction (see 3.4.2) that would later guide fine-tuning. Only after these choices had been made was the use of summaries re-examined (see 3.4.3), to confirm that no performance loss was introduced when compared to full-text inputs.

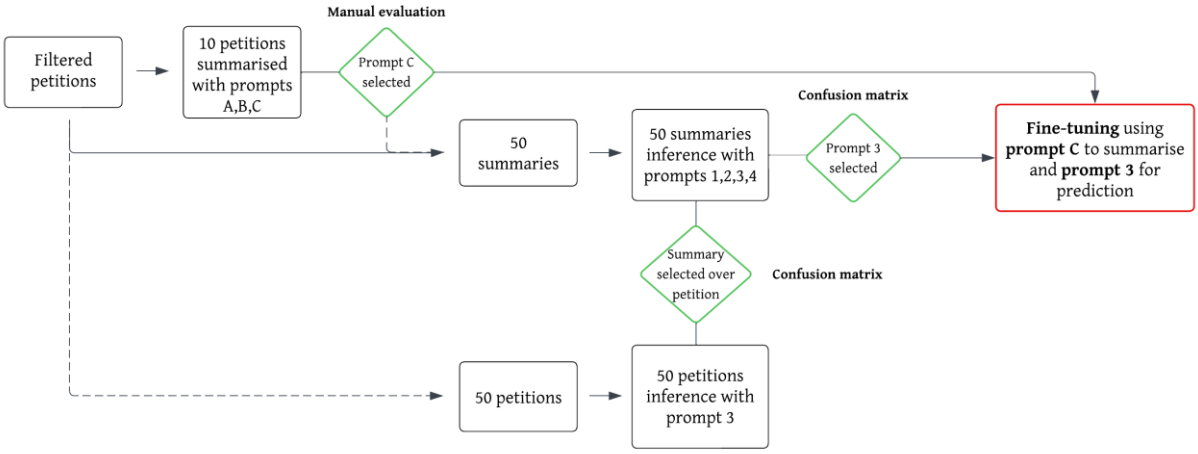

Figure 3.10 - Workflow of prompt experiments

3.4.1. Petition summarisation

To reduce the computational and financial costs of using LLMs such as GPT, full petitions were not employed as model inputs. Instead, compressed summaries retaining essential legal and factual content were generated. This choice was motivated by two considerations:

a) Cost-efficiency: full petitions would significantly increase token usage, thereby raising both fine-tuning and inference costs.

b) Scalability: shorter inputs enable faster processing and make real-world deployment more feasible.

Summarisation was therefore adopted as a strategic step, ensuring that the modelling phase could proceed without sacrificing legal relevance. Summaries were generated with the GPT-4o model via the OpenAI API, selected at this stage for its strong general-purpose performance (Nwanne, W., 2025).

Three candidate summarisation prompts (A, B, and C) were designed, differing in structure and level of guidance. These were applied to a random sample of 10 petitions and qualitatively assessed according to criteria such as factual coverage, legal detail, structural clarity, and references to national decisions. Although the sample size was small, it was considered sufficient for a qualitative comparison, enabling the identification of strengths, weaknesses, and error patterns that might be obscured by quantitative metrics alone. Such limited domain-expert evaluation is consistent with practices in prompt engineering research, where manual review often guides early optimisation before large-scale testing (Wang et al., 2024; He et al., 2025).

The outcome of this evaluation, including the final prompt choice, is presented in Subchapter 4.1.1.

3.4.2. Prompt selection for prediction

Before proceeding with fine-tuning, an effective prompt for the classification task had to be identified. Four candidate prompts were developed to instruct the model to classify each petition summary as either violation (1) or no violation (0), with variations in structure, tone, and emphasis on legal reasoning.

The prompts were first tested with GPT-4o in zero-shot mode (i.e., without fine-tuning) on a balanced validation set of 10 cases. Based on these preliminary results, the evaluation was extended to 50 cases, allowing for a more reliable comparison of performance. The outcome of this evaluation, including the selection of the final prediction prompt, is presented in Subchapter 4.1.2.

3.4.3. Revalidating the use of summaries

The decision to rely on summaries rather than full petitions was initially made during data preparation for reasons of cost and scalability. However, given the iterative nature of the process, this choice was subsequently re-examined. After the most effective prediction prompt had been identified, an additional evaluation was performed to verify whether the use of summaries entailed any performance loss when compared to full-text inputs. The results of this reassessment are reported in Subchapter 4.2.

3.5. Modelling

The modelling phase aimed to develop a classification system capable of predicting the outcome of ECtHR cases based solely on the petition summary. Given the textual nature of the task, LLMs were employed, leveraging their capacity to process complex legal language and to reason over abstract patterns.

All models were fine-tuned and deployed using the OpenAI API, which supports both zero-shot inference and supervised fine-tuning. The modelling approach followed three main stages:

- Provider Selection: OpenAI was selected as the model provider due to its accessible interface, support for fine-tuning, and strong performance in natural language tasks. During the modelling phase, three model variants – GPT-3.5-turbo, GPT-4o-mini and GPT-4o – were experimentally compared through fine-tuning to assess their performance in the classification task.
- Inference vs. Fine-tuning: zero-shot inference was used to test different prediction prompts and to select the most effective instruction format. In this setting, the model receives only prompt and input text at inference time, without prior exposure to labelled

fine-tuning examples. For the final classifier, a fine-tuning approach was adopted, in which the model was fine-tuned on supervised examples consisting of summarised petitions and their corresponding binary outcomes. Fine-tuning enables the model to internalise domain-specific patterns and improve performance through repeated exposure (Bergmann, D., 2024).

    - Implementation: zero-shot inference is executed via a standard API call to a pre-trained model, supplying a prompt and receiving a direct response without altering the model's internal weights. Fine-tuning requires preparing a .jsonl dataset in chat-style format, uploading it to OpenAI's platform, and running supervised training on a selected base; upon completion, a custom model ID is produced for deployment through the same API endpoints.

- Fine-tuning Procedure: the fine-tuning process was implemented incrementally to assess how performance evolved with increasing amounts of data. Five cumulative datasets were prepared, ranging from 50 to 400 cases (see 3.5.2). All examples followed a consistent input format (see 3.5.3), embedding instructions and labels within OpenAI's chat-style message schema. Each fine-tuning run was executed using OpenAI's create fine-tune job endpoint, with default parameters except the final experiment.

To provide a systematic evaluation, the experimental workflow was structured as shown in Figure 3.11. Each fine-tuning run combined a training[2] set and a separated test set, producing a distinct set of results. By incrementally increasing the training size across nine runs, the study traced the effect of data volume on predictive performance. After the ninth run, an additional experiment with alternative hyperparameter settings was executed, to assess the sensitivity of the results to training configuration. This framework ensured that the modelling process was both transparent and replicable, while enabling the comparative analysis presented in Subchapter 4.3.

[2] Throughout this thesis, "training" denotes fine-tuning of pre-trained LLMs.

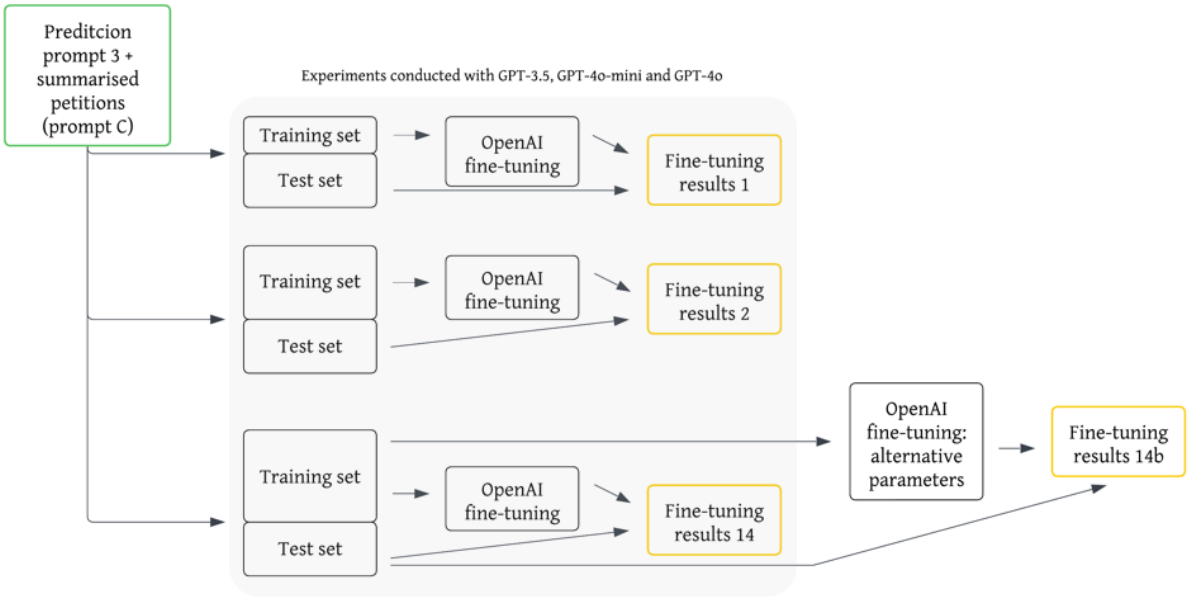

Figure 3.11 - Workflow of fine-tuning experiments with GPT models

3.5.1. Model selection

The fine-tuning experiments were conducted using three models provided by OpenAI: GPT-3.5, GPT-4o-mini, and GPT-4o. The selection of these models was motivated by their availability for fine-tuning, their different trade-offs between performance and computational cost, and their relevance for natural language processing tasks involving complex legal texts.

GPT-3.5 was chosen as a baseline, given its wide use and lower computational cost, making it a suitable starting point for experimentation (Zane, X., 2025). GPT-4o-mini was included as a lighter variant of the GPT-4 family, designed to balance efficiency and performance (Ong, R., 2024). Finally, GPT-4o was selected as the most advanced option, expected to achieve higher performance due to its enhanced reasoning capabilities (OpenAI, 2024).

The use of these three models allowed for testing the methodology across different levels of model capacity, ensuring that the fine-tuning process was not dependent on a single model family.

3.5.2. Training and test sets

To assess the impact of training size on model performance, five cumulative datasets were constructed: 50, 100, 200, 300, and 400 randomly selected examples. Each set included all the cases from the smaller subsets and was balanced across the two classes. At the outset, the number of examples required for the model to achieve stable performance was unknown, so the experiment was designed to incrementally increase training size and observe its effect. The

choice of a maximum of 400 training examples was therefore not intended to exploit the full dataset, but rather to strike a balance between empirical insight and computational feasibility. This design highlights the learning behaviour of LLMs in settings with limited data, in contrast to traditional models that typically require larger datasets to achieve comparable results.

Moreover, fine-tuning LLMs is substantially more expensive than inference (OpenAI, 2025), both in terms of token usage and processing time, which further reinforced the need to limit training size while still enabling a meaningful evaluation of performance.

To evaluate the model's predictive performance, 518 more recent cases were chosen for the test set to better approximate real predictive use, ensuring that the model is assessed on cases it has not seen during training.

3.5.3. Input formatting

Each training example was formatted as a chat-style interaction, following OpenAI's fine-tuning structure. Specifically, the input consists of two messages:

- a user message containing the classification prompt (Prompt 3) followed by the petition summary (generated using Prompt C).
- an assistant message with the binary label ('0' or '1'), representing the Court's decision.


```
{
 "messages": [
  {
   "role": "user",
   "content": "[Prediction prompt instructions]\n\n[Petition summary]"
  },
  {
   "role": "assistant",
   "content": "1"
  }
 ]
}
```


Figure 3.12 - Example JSONL record (label: violation = 1)

The instructions were embedded directly into the user message to ensure consistency between fine-tuning and inference, since in practical applications user inputs constitute the primary channel through which models receive instructions. The resulting dataset was saved in .jsonl format, as required by the OpenAI fine-tuning API. This format was also adopted for the test set used during inference, ensuring consistency across all phases of modelling.

3.5.4. Fine-tuning configuration and hyperparameters

All fine-tuning experiments were initially conducted using the default configuration parameters (3 training epochs, batch size 1, learning rate multiplier 2). In addition to this baseline, a second configuration with slightly modified parameters was also tested, to explore potential improvements in learning stability and efficiency. The outcomes of this comparison are presented in Subchapter 4.4.

3.6. Evaluation

The evaluation phase assessed the predictive performance of each fine-tuned model on the held-out test set, which had not been seen during training.

3.6.1. Evaluation metrics

To evaluate the performance of the classification models, a set of standard metrics commonly used in binary classification – accuracy, recall, precision, and F1-score – was adopted.

Since the test set is moderately balanced between violation (315 cases) and non-violation (203 cases), accuracy was considered a valid measure of overall model correctness. However, additional metrics were regarded as essential to provide a more complete assessment of model performance. The metrics were defined as follows, based on the four possible prediction outcomes:

- True Positives (TP): cases in which a violation was correctly predicted.
- False Positives (FP): cases in which a violation was predicted when none occurred.
- False Negatives (FN): cases in which an actual violation was not detected.
- True Negatives (TN): cases in which non-violation was predicted and none occurred.

Precision (P) refers to the proportion of predicted violations that were correctly classified. It reflects how reliable the model is when it predicts a human rights violation:

$$P = TP / (TP + FP)$$

Recall (R) measures the proportion of actual violation cases that the model correctly identified. In this context, it was prioritized due to the ethical concern of minimizing false negatives, i.e., actual human rights violations that go undetected.

$$R = TP / (TP + FN)$$

F1-score is the harmonic mean of precision and recall. It provides a balanced view of model performance, especially useful when precision and recall must be traded off:

$$F1 = 2 \cdot (P \cdot R) / (P + R)$$

Accuracy (A) indicates the overall proportion of correct predictions, considering both classes:

$$A = (TP + TN) / (TP + TN + FP + FN)$$

Finally, a confusion matrix was employed as a diagnostic tool to visualise the distribution of correct and incorrect predictions across both classes, facilitating a more intuitive understanding of model errors.

Table 3.2 - Confusion Matrix

	Predicted: Violation (1)	**Predicted: No Violation (0)**
Actual: Violation (1)	TP	FN
Actual: No Violation (0)	FP	TN

3.6.2. Ethical relevance of recall

Recall for the violation class was prioritised for ethical reasons. From a legal perspective, the failure to identify a true violation of human rights was regarded as more consequential than issuing a false alarm. A high recall was considered essential to minimise false negatives, which, in this context, correspond to undetected human rights violations.

3.7. Deployment

Although no production deployment was carried out in this study, the final phase of the CRISP-DM methodology is addressed here to reflect on the model's potential applicability in real-world settings and to identify the steps required for operational use.

From a technical perspective, the fine-tuned model could be deployed via API using OpenAI's infrastructure, allowing legal practitioners or researchers to input petition summaries and receive outcome predictions. However, the deployment of such a system in a judicial or institutional setting would raise several concerns.

First, any practical deployment would need to be preceded by rigorous validation in collaboration with legal experts that the model's outputs are interpretable, consistent, and legally sound. Second, considerations relating to explainability, accountability, and fairness would have to be addressed before integration into decision-making workflows, particularly in sensitive areas such as human rights litigation.

Accordingly, the current model should be regarded as a proof of concept, demonstrating the technical feasibility of using LLMs to predict ECtHR outcomes. In the future, similar models could be used to support decision-making processes in various ways, such as:

- Assisting legal professionals in prioritising which cases to analyse or escalate.
- Supporting academic research into judicial patterns and decision trends.
- Serving as an internal tool for human rights organisations to estimate the likely outcome of petitions before the ECtHR.

4. RESULTS AND DISCUSSION

This chapter presents and discusses the results obtained during the modelling phase. The analysis is structured to reflect the key decisions made during model development, from prompt selection to input format validation and model fine-tuning. The performance of each model variant is compared across different training sizes and input configurations.

4.1. Prompt selection

4.1.1. Summarisation prompts

When comparing the three prompts (Figure 4.1), Prompt A produced very short and superficial summaries, often omitting legal arguments and references to national decisions, which made it unsuitable for the task. Prompt B generally generated longer output but often included redundancy and weaker structure. By contrast, Prompt C consistently produced concise yet informative summaries, systematically incorporating relevant articles of the European Convention, legal reasoning, and references to national decisions. These characteristics made Prompt C the most suitable for this project, and it was therefore selected to generate the complete set of summaries used in all subsequent prediction experiments.

Prompt C is shown below, while Prompts A and B provided in Appendix B.

Table 4.1 - Qualitative comparison of summarisation prompts

Criterion / Prompt	A	B	C
Conciseness	Short	Long	Medium
Coverage of the facts	Superficial	Good	Complete
Articles included	Almost never	Often	Almost always
Legal arguments	General	Present	Detailed
National decisions mentioned	No	Sometimes	Yes
Structured style	Free	Partial	Yes
Overall legal relevance	Limited	Good	High
Reference to the country	Not always	Sometimes	Always

Summarize the following petition filed with the European Court of Human Rights, without starting with the word "summary". The summary should be concise but comprehensive, highlighting the main points, including:
- the country against which the complaint is made (the respondent);
- information about the complainants, like name, age, nationality;
- previous decisions by national courts on the case;
- articles of the European Convention on Human Rights that are at stake;
- the relevant facts of the case;
- the parties' arguments.
Make sure the summary maintains the clarity and accuracy of the information provided in the original text and ensure that the country against which the complaint is made (the respondent) is clearly mentioned.

Figure 4.1 - Prompt C: instruction for summarising petition texts

4.1.2. Prediction prompts

The results of the zero-shot inference experiments conducted to identify the most effective prompt for the classification task are presented in Table 4.. All prompts were tested using GPT-4o and summarised petitions, initially on a small balanced validation set of 10 examples, and subsequently on a larger set of 50 examples. The objective was to evaluate how variations in prompt structure and emphasis influenced classification performance, particularly in terms of accuracy and recall for the violation class.

In the initial evaluation, Prompt 3 was found to achieve the highest accuracy (0.70) and to provide a strong balance between the two classes, with recall values of 0.60 for class 0 and 0.80 for class 1. By contrast, Prompts 1 and 2 displayed limited ability to detect non-violations (recall of 0.20), while Prompt 4 underperformed on both classes. Based on these results, Prompt 3 was selected for use in all subsequent modelling phases.

You are a legal assistant helping to classify petitions submitted to the European Court of Human Rights. Your task is to classify each petition as '0' (non-violation) or '1' (violation) based on the criteria and facts presented.
To make your decision, consider the following points:
Legal Criteria: the court bases its decisions on the European Convention of Human Rights;
Case Details: analyze the facts, including the nature of the allegation, the evidence provided, and the historical and factual context;
Previous Agreements: check whether there are prior agreements between the parties involved; the existence of such agreements may indicate that the right has not been violated, unless there is clear proof that the agreement has been violated;
Objectivity: ignore the petitioner's subjectivity and emotional claims; focus on objective facts and the application of legal criteria;
Documentary and Testimonial Evidence: evaluate the strength of the documentary and testimonial evidence presented;
National Court Decisions: consider the decisions made by national courts.
Based on these points, classify the petition as follows:
'0' for non-violation
'1' for violation
The output must contain only 0 or 1.

Figure 4.2 - Prompt 3: instruction for prediction tasks

Although this decision was made early in the modelling process, a second evaluation was later conducted using a larger validation set of 50 balanced cases (25 violations and 25 non-violations), with the aim of retrospectively confirming the robustness of the prompt choice. At that stage, fine-tuning had already been completed using Prompt 3, and retraining with an alternative prompt was not feasible due to time and resource constraints. In this later evaluation, Prompt 3 did not achieve the highest recall for the violation class or the highest overall accuracy, although it remained close to the top results and continued to provide a favourable balance between sensitivity to violations and overall classification performance. Given its strong performance in the initial evaluation, alignment with the project's priorities, and the need for methodological consistency, Prompt 3 was retained for all subsequent modelling stages.

Prompt 3 is shown below, while Prompts 1, 2, and 4 are provided in Appendix C.

Table 4.2 – Prompt performance in zero-shot GPT-4o with summaries (10 vs. 50 cases)

Model	Prompt	Set size	Input Format	Recall		Acc
				0	1	
GPT 4o	1	10	Summary	0.20	0.80	0.50
	2			0.20	0.80	0.50
	3			0.60	0.80	0.70
	4			0.40	0.60	0.50
GPT 4o	1	50		0.28	0.92	0.60
	2			0.28	0.92	0.60
	3			0.28	0.88	0.58
	4			0.56	0.64	0.60

4.2. Revalidating the use of summaries

As part of the iterative nature of the CRISP-DM process, the design choice of using summarised petitions was retrospectively revisited. To ensure that this decision was not compromised in terms of predictive quality, a validation experiment was conducted using zero-shot inference with the original full texts – that is, without any prior training or exposure to labelled examples. This retrospective test was performed on the same 50 cases (25 violations and 25 non-violations) previously employed during the prompt selection phase. The objective was to verify whether any key legal information relevant for classification had been removed by the summarisation process.

Using Prompt 3 and the GPT-4o model in zero-shot mode, the exact same classification logic was applied to the full-length petitions. The results were subsequently compared with those obtained using the summarised versions of the same 50 cases (Table 4.).

Interestingly, whilst an accuracy of 58% was achieved with the summaries, a slightly lower accuracy of 56% was yielded by the full texts – despite their containing the complete legal narrative. More notably, a recall of 100% for the violation class was obtained when the full texts were used. At first glance, this might be considered preferable, especially given the project's priority to minimise false negatives. However, such perfect recall may indicate overfitting to specific linguistic or stylistic cues present in the longer documents.

From a generalisation perspective, concerns are thus raised. Full petitions include a wide range of case-specific details and narrative styles, which may introduce noise or variability that is difficult to learn from in small datasets. In contrast, the summaries were generated using a consistent instruction format (Prompt C), by which the language model was guided to extract and organise key legal elements – such as facts, legal arguments, and references to articles – in a structured and standardised manner. This regularity is considered to help models learn more stable patterns, rendering them less prone to overfitting and potentially more generalisable to unseen cases.

Therefore, despite the lower recall of the summaries in this sample, it is believed that the summarisation step did not impair the model's capacity to generalise. The trade-off between perfect recall on a small sample and more controlled, consistent inputs on a larger scale supports the original design choice.

Table 4.3 – Revalidating summaries vs. full texts in GPT-4o (Prompt 3, 50 cases)

Model	Prompt	Set size	Input Format	Recall		Acc
				0	1	
GPT 4o	3	50	Summary	0.28	0.88	0.58
			Text	0.12	1.00	0.56

4.3. Performance of fine-tuned models

The performance of all fine-tuned models on the held-out test set of 518 petitions is reported in Table 4.. A consistent prompting regime was used across experiments, and three model families were evaluated (GPT-3.5-turbo, GPT-4o-mini, GPT-4o) over cumulative training sizes from 50 to 400 examples. Metrics are reported per class (recall, precision, F1) together with overall accuracy, execution time, and monetary cost.

A clear pattern emerges: both model choice and training size materially influence outcomes. Smaller models display greater variability across set sizes, with phases of improvement that are not strictly monotonic, whereas the largest model (GPT-4o) offers a more stable profile and the strongest top-line results. In practical terms, the best accuracy is obtained

by GPT-4o at medium training sizes (73% at 200 and 300 examples), while GPT-3.5-turbo and GPT-4o-mini remain below that ceiling despite occasional gains with additional data. These observations indicate that architecture capacity and data regime jointly shape effectiveness, even under a uniform experimental setup.

For GPT-3.5-turbo, an initial accuracy of 63% is achieved with 50 training examples, but performance declines to 60% at 100 before partially recovering to 65% at 200; this fluctuation suggests instability in how the model adapts to the data, likely reflecting sensitivity to small changes in training size rather than a consistent learning trend – an effect that is plausible in legal language, where subtle textual nuances can strongly affect performance. GPT-4o-mini follows a different trajectory: it underperforms initially (below 40% at 50 examples), improves at 100 (57%) and 200 (62%), peaks at 68% with 300, and declines again to 64% at 400, reflecting an overall unstable pattern that is consistent with limited capacity or high sensitivity to data distribution in complex legal summaries. By contrast, GPT-4o achieves the most balanced and robust performance, with the highest accuracy among the tested models, reaching 73% at both 200 and 300 examples while also delivering superior recall for the violation class; at 200 examples, recall for violations is 0.77, already surpassing the project's success threshold. The corresponding confusion matrix (Figure 4.4) illustrates the distribution of correct and incorrect classifications; although absolute counts reflect the larger number of violation cases in the test set, the associated metrics confirm higher recall for violations (0.77) than for non-violations (0.67), indicating greater sensitivity to detecting infringements.

Within each family, scaling the training set improves performance up to a point, while monetary and time costs rise more steadily. In GPT-3.5, expanding from 50 to 400 examples increased spend by 4.75 USD and runtime by 41 minutes; accuracy ended essentially flat (63% to 62%) and violation-class recall fell by 10 percentage points (pp), suggesting that most useful signal is already captured around 200 examples. In GPT-4o-mini, the jump from 50 to 300 examples raised accuracy by 29 pp and violation recall by 71 pp for an extra 1.27 USD and more 16.8 minutes; moving to 400 added 0.51 USD and more 0.2 minutes but accuracy dropped by 4 pp and recall eased slightly. In GPT-4o, advancing from 50 to 300 examples lifted accuracy by 12 pp and violation recall by 21 pp at an additional 12.64 USD, with latency roughly stable; beyond 200 examples, 4.92 USD more bought no further accuracy gain and left violation recall unchanged.

Overall, GPT-4o typically entails a cost between 3.5 and 10 times higher than that of the smaller models at comparable training sizes, trading higher spend and longer latency for superior accuracy and improved detection of violations. These findings underscore the joint

influence of model architecture and training size on predictive performance: the smaller models (GPT-3.5-turbo and GPT-4o-mini) tended to be more unstable and to generalise less well, whereas evidence from the tested configurations indicates that GPT-4o offers a more stable profile at medium training sizes, aligning with the study's emphasis on identifying rights violations.

Table 4.4 - Performance of fine-tuned GPT models with summaries across set sizes

Model	Set size	Input Format	Recall		Precision		F1-score		Acc	Execution time (min)	Cost (USD)
			0	1	0	1	0	1			
GPT 3.5	50	Summary	0.49	0.71	0.52	0.68	0.51	0.70	0.63	7.5	0.68
	100		0.70	0.53	0.49	0.73	0.58	0.62	0.60	10.9	1.36
	200		0.53	0.73	0.56	0.71	0.54	0.72	0.65	18.6	2.73
	300		0.58	0.71	0.57	0.72	0.57	0.72	0.66	34.3	4.08
	400		0.64	0.61	0.51	0.73	0.57	0.66	0.62	48.5	5.43
GPT 4o-mini	50		1.00	0.00	0.39	0.00	0.57	0.00	0.39	5.1	0.25
	100		0.81	0.42	0.47	0.77	0.60	0.54	0.57	7.6	0.51
	200		0.51	0.70	0.52	0.69	0.51	0.69	0.62	12.8	1.01
	300		0.62	0.71	0.58	0.74	0.60	0.73	0.68	21.9	1.52
	400		0.56	0.68	0.53	0.71	0.55	0.69	0.64	22.1	2.03
GPT 4o	50		0.69	0.56	0.50	0.74	0.58	0.63	0.61	27.7	2.12
	100		0.65	0.73	0.60	0.76	0.62	0.74	0.70	24.9	4.25
	200		0.67	0.77	0.65	0.79	0.66	0.78	0.73	30.8	9.84
	300		0.66	0.77	0.65	0.78	0.66	0.78	0.73	28.5	14.76

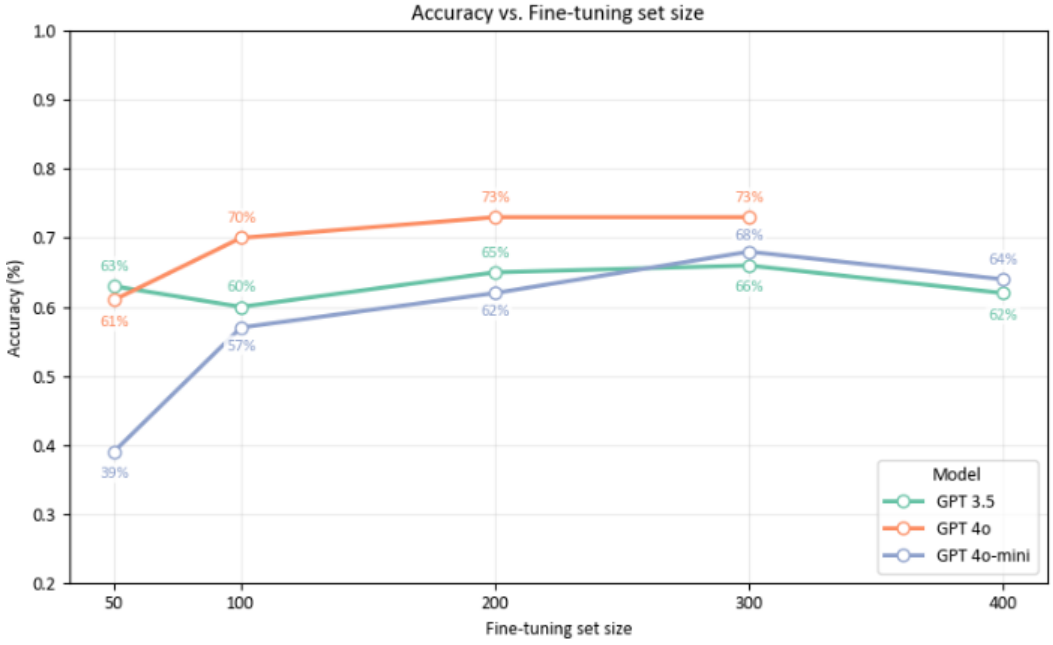

Figure 4.3 - Accuracy of fine-tuned GPT models across set sizes

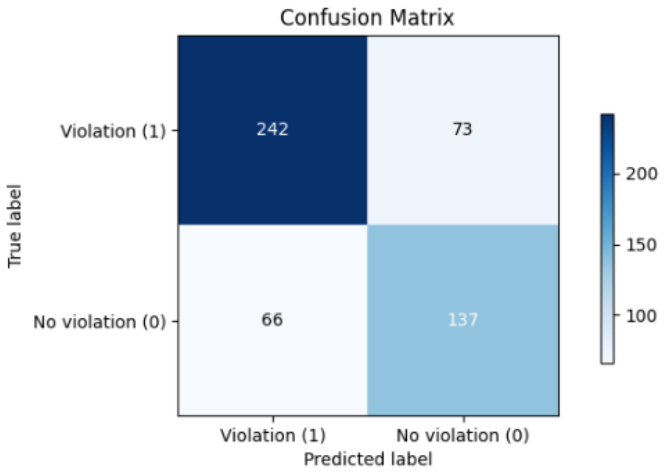

Figure 4.4 - Confusion matrix of GPT-4o fine-tuned on 200 case summaries

4.4. Hyperparameter sensitivity analysis

To further explore the baseline configuration, a final fine-tuning round was conducted with adjusted hyperparameters. As a reference point, the GPT-4o model fine-tuned on 300 cases was selected as the baseline. This choice was made because the results at 200 and 300 cases were very similar, even achieving the same accuracy, and therefore the larger dataset was considered the most reasonable basis for testing refinements. Until this point, all models were trained using OpenAI's default configuration: 3 epochs (i.e., passes through the dataset), a batch size of 1 (the number of examples processed at once before updating weights), and a learning rate multiplier of 2 (which controls the magnitude of adjustments during each update). These defaults are often sufficient for general tasks, but given the model's promising performance, training was refined to allow more stable and efficient learning.

In the adjusted configuration, the number of epochs was increased to 4 (allowing the model slightly more exposure to the training data), the batch size was doubled to 2 (enabling the model to capture better gradients by seeing more data per update), and the learning rate multiplier was reduced to 1.5 (to avoid overshooting optimal weights). These changes, though small, were meaningful in the context of fine-tuning large models on medium-sized datasets.

As shown in Table 4., the adjusted configuration produced results broadly comparable to the baseline, with accuracy remaining at 73% and only minor variations across recall, precision, and F1-scores. The most notable improvement was observed in recall for the violation class (0.77 to 0.81), aligning with the project's priority of reducing false negatives.

Beyond predictive performance, Table 4. also reports execution time and monetary cost. The baseline configuration required approximately 28.5 minutes and USD 14.76, while the adjusted experiment took 23.5 minutes and USD 13.31. The reduction in time and cost can be

explained by the larger batch size, which reduced the number of weight updates and API calls, outweighing the additional epoch introduced (Devansh, 2022). Although the refined configuration was slightly cheaper and faster, the overall cost of fine-tuning runs remained substantial. This financial constraint limited the scope of additional hyperparameter experiments.

Table 4.5 – GPT-4o performance with summaries: baseline vs. adjusted fine-tuning

Model	Configuration	Recall		Precision		F1-score		Acc	Execution time (min)	Cost (USD)
		0	1	0	1	0	1			
GPT-4o	Baseline	0.66	0.77	0.65	0.78	0.66	0.78	0.73	28.5	14.76
GPT-4o	Experiment 1	0.60	0.81	0.67	0.76	0.63	0.78	0.73	23.5	13.31

4.5. Post-hoc analysis of model predictions

When AI and ML models are discussed, the terms interpretability and explainability are often mentioned. However, distinct concepts with important implications are represented by these terms. Interpretability is defined as the degree to which a model's decisions can be directly understood by a human, as is observed in models such as decision trees or linear regressions, where the influence of each input on the output can be clearly identified. In contrast, explainability is concerned with the ability to provide understandable explanations for a model's predictions, even when the model itself is highly complex or operates as a "black box." For example, although a neural network may not be inherently interpretable, explainability can still be provided by identifying the features that contributed most to a specific prediction. Thus, intrinsic transparency is emphasized in interpretability, whereas the reasoning behind outcomes is made comprehensible in explainability, regardless of the model's complexity (IBM, 2024).

To provide some degree of explainability of the model's predictions, a post-hoc analysis of bigrams in the predicted classes was conducted. The most frequent bigrams across the predicted classes of violation and non-violation were initially identified; however, they appeared to convey limited interpretative value (see Appendix D).

A deeper examination was subsequently performed by identifying bigrams that were exclusive to each predicted class. In the violation class, exclusive bigrams such as *fair hearing*, *domestic courts*, *raises questions*, and *pre detention* were observed. The bigram *pre detention*, previously noted in the exploratory analysis, may indicate cases involving issues of procedural fairness before detention, highlighting procedural rights concerns. The bigram *fair hearing* is particularly relevant, as it relates directly to Article 6 of the ECHR, suggesting cases where the adequacy, impartiality, or timeliness of judicial proceedings was contested. Similarly, *domestic*

courts emphasise the involvement of national courts in assessing these rights, while *raises questions* could indicate judicial doubt or the need for further scrutiny regarding the application of rights in specific cases. Other relevant bigrams include *rights protocol* or *property rights*, which highlight the invocation of specific rights provisions under the Convention.

In the non-violation class, exclusive bigrams such as *right respect*, *private family*, *national born*, and *supreme administrative* were identified. The bigram *right respect*, also noted in the exploratory analysis, may suggest references to rights that were considered respected, though its precise implication remains less clear. *Private family* likely relates to considerations under Article 8 of the Convention, where the rights to privacy and family life were assessed and ultimately not found to be violated. *Supreme administrative* points to the involvement of high-level administrative or judicial review bodies in resolving these cases, while *national born* may reflect arguments tied to nationality and Article 14 (non-discrimination). Additionally, the appearance of the bigram Bosnia Herzegovina may reflect a country-specific pattern in the dataset. Like the earlier observation of Russia in the violation class, this highlights a potential bias where certain countries might be overrepresented in one predicted class, which should be interpreted with caution.

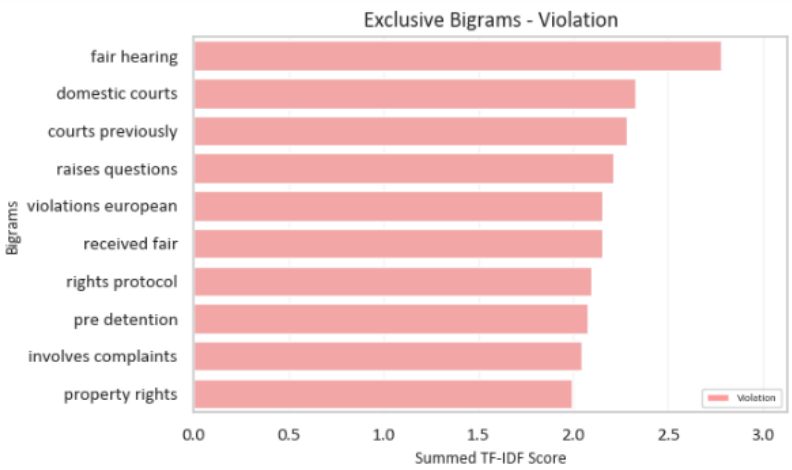

Figure 4.5 - Top exclusive bigrams in predicted violation cases

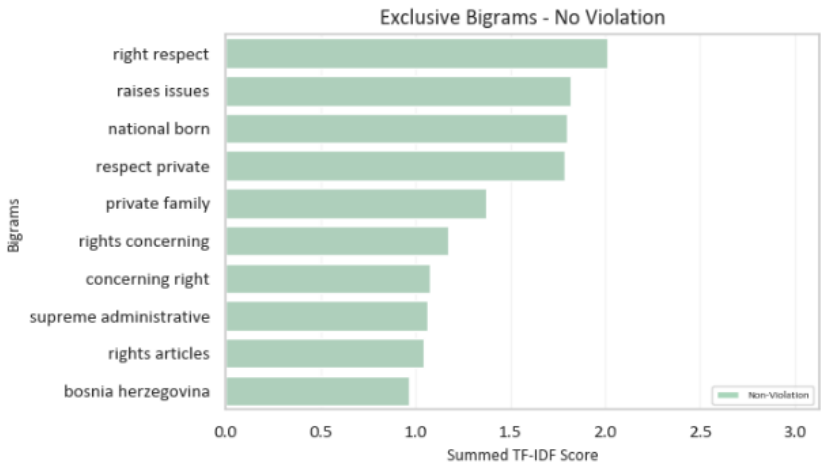

Figure 4.6 - Top exclusive bigrams in predicted non-violation cases

These findings suggest that class-specific linguistic patterns partially account for the model's predictions. Analysing exclusive bigrams offers an initial, human-readable view of features that differentiate violation from non-violation cases, but it should be regarded as a first step toward explainability rather than a comprehensive account of the model's behaviour.

5. ETHICAL AND LEGAL CHALLENGES

The increasing integration of AI into judicial processes raises profound questions about the ethical, legal, and institutional implications of such technologies. While AI offers the potential to enhance efficiency, consistency, and accessibility in courts, it simultaneously introduces risks that cannot be overlooked. Decisions in judicial contexts are inherently complex, requiring careful consideration of laws, precedents, and social circumstances – factors that may resist straightforward algorithmic codification. Ethical concerns emerge regarding bias, explainability, accountability, and the potential erosion of impartiality, while legal frameworks impose clear limits on the use of automated systems in decision-making (Teigão & Fogaça, 2025).

In this context, frameworks such as the European Ethical Charter on the use of AI in judicial systems, adopted by the Council of Europe in 2018, provide guiding principles to ensure that AI tools respect fundamental rights, maintain fairness, and support rather than replace human judgment. Moreover, the introduction of AI tools impacts the institutional functioning of courts, influencing public trust and the role of judges in safeguarding justice. Understanding these implications is crucial to ensure that AI serves as a supportive instrument, augmenting human judgment without undermining the fundamental principles of the judicial system (Junior & Nogueira, 2025).

5.1. Judicial decision-making theory and discretion

Judicial decision-making has traditionally been understood as a process that cannot be reduced to the mechanical application of legal rules. Rather, it is shaped by reasoning and interpretation, requiring that decisions be justified through arguments that are coherent and principled. In this perspective, legal validity is not derived merely from formal authority but from the capacity of judicial reasoning to withstand rational scrutiny. Alexy (2010) has argued that judicial decisions must be grounded in a discourse of justification, in which principles and rules are balanced through rational argumentation rather than arbitrary choice. This argumentative structure places limits on discretion, as judges are compelled to present reasons that can be critically assessed within the legal community.

The existence of discretion, however, remains an inherent feature of adjudication. In contexts where rules leave gaps or where competing rights must be balanced, judges exercise a margin of interpretative freedom. Within the ECtHR, this phenomenon has been conceptualized through the doctrine of the “margin of appreciation,” which allows national authorities a certain

latitude in applying the ECHR, while ensuring that ultimate supervision is retained by the Court (Greer, 2000). The doctrine highlights the tension between judicial universality and contextual sensitivity: rights are to be upheld consistently, yet their application must account for cultural, social, and political differences across member states.

Dworkin (1978) further reinforced that judicial discretion is not unlimited, since rights must be treated as "trumps" that constrain both legislative and judicial choices. This perspective underscores that adjudication cannot be guided by utilitarian calculations alone but must be anchored in the principled protection of individual rights.

Considering these insights, the application of AI to judicial contexts must be carefully framed. While computational systems may assist courts in tasks such as case prioritisation or the identification of patterns across large datasets, they cannot replace the interpretative and principled reasoning that underpins judicial discretion. Substituting human judgement with algorithmic decision-making would risk reducing adjudication to utilitarian efficiency, undermining the argumentative, normative, and rights-based foundations of justice. AI can therefore be conceived only as a supportive instrument – enhancing efficiency while leaving the core task of principled judgement to human judges.

5.2. Ethical implications of AI

The ethical evaluation of AI in courts begins with transparency and responsibility for reasons. In judicial settings, reason-giving is part of legitimacy. When complex models operate opaquely, meaningful contestation and appeal become difficult. Research shows that machine-learning systems can be both inscrutable and non-intuitive, and each problem demands a different remedy. Explaining mechanics does not, by itself, justify the use of a model in adjudication (Selbst & Barocas, 2018). Interpretability should therefore be treated as a design choice. Systems used in rights-sensitive decisions ought to be selected or constrained so that their reasons can be understood by affected parties (Molnar, 2020). Accountability also matters. Algorithmic systems need audit trails, traceability, and enforceable oversight so that responsibility can be located when errors occur (Kroll et al., 2017). In European data-protection law, debate over a stand-alone "right to explanation" warns against overclaiming what the GDPR provides, yet rules on automated decision-making, notice, and access still create a meaningful accountability regime (Wachter et al., 2017; Kaminski, 2019).

Fairness is a second axis of ethical concern. Training data often reflect historical inequalities; as a result, model outputs can differentially burden protected groups even without discriminatory intent. There is no single fairness criterion that can be satisfied across all

contexts; parity of errors, calibration, and independence frequently trade off, and choices among them are normative rather than purely technical (Barocas et al., 2023; Binns, 2020). In the judicial sphere, these trade-offs are acute: a metric that optimizes aggregate accuracy can still erode equality before the law if it systematically shifts risks onto the least advantaged (Binns, 2020).

To address these risks, widely cited ethical frameworks converge on core requirements. The European Commission's Ethics Guidelines for Trustworthy AI promote human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity/non-discrimination/fairness, societal well-being, and accountability (European Commission HLEG, 2019). Mapping studies show broad global agreement around such principles, while warning that high-level guidance must be operationalized to be effective (Jobin et al., 2019). Complementary work stresses that ethics must translate into concrete practices that secure safety, justice, and respect for rights (Leslie, 2019).

A rights-based lens is therefore indispensable in judicial settings. International human-rights law has been proposed as a workable accountability frame because it defines protected interests, allocates responsibility across actors, and applies across the AI lifecycle – from design choices to deployment and redress (McGregor et al., 2019). Ethical and legal considerations converge here: the legitimacy of any judicial AI depends on whether it supports rather than displaces the human responsibility to reason about and justify outcomes (Sartor, 2020). In practical terms, AI should be confined to assistive roles – such as case prioritization, document analysis, and pattern detection – where interpretability and auditability can be assured, and final judgments remain human-reasoned and rights-constrained (Selbst & Barocas, 2018; Molnar, 2020; Kroll et al., 2017). Delegating adjudicative discretion to opaque or purely consequentialist models would risk collapsing adjudication into utilitarian optimization, undermining impartiality, public trust, and the rights-based foundations of justice (European Commission HLEG, 2019; Jobin et al., 2019).

5.3. Legal and institutional implications

Legal constraints and institutional design need to be considered together when courts evaluate AI. In EU law, a risk-based framework has been established that imposes specific duties where systems are used in the administration of justice. High-risk classification brings obligations on risk management, data and model quality, logging, transparency to users, and real human oversight. In the public sector, fundamental-rights impact assessment has been proposed as a condition for first use. These features shape procurement and deployment choices and link court

practice to external conformity assessment and post-market supervision (European Parliamentary Research Service, 2021; Veale, M. & Borgesius, F., 2021).

Council of Europe instruments complement this architecture. The European Ethical Charter on the Use of AI in Judicial Systems sets five principles: respect for fundamental rights, non-discrimination, quality and security, transparency with the possibility of external audit, and user control to preserve human authority. Special caution is urged in criminal matters to avoid discriminatory or determinist uses that would compromise fair-trial guarantees (CEPEJ, 2018). The Framework Convention on AI requires Parties to align AI lifecycles with human rights, democracy, and the rule of law. It calls for documentation, risk monitoring, ex-ante testing where appropriate, independent oversight, and, where necessary, moratoria or bans on uses incompatible with those values (Council of Europe, 2024).

Transparency and explanation duties remain a core legal issue. Debate under EU data-protection law shows that a general "right to explanation" should not be overstated. Explanation of model mechanics is not the same as providing reasons that justify a legal outcome, and current rules distribute duties across information, access, and review rights rather than a single uniform claim to explanation. In judicial settings, reason-giving remains central to legitimacy and appealability, so explainability must be oriented to meaningful contestation rather than to purely technical disclosure (Edwards & Veale, 2017).

These legal limits translate into concrete institutional choices. Responsibilities need to be allocated between judges, court administration, and technical teams. Human oversight must be real, with authority to accept, question, or reject system outputs and a duty to give reasons when AI is relied upon. Data provenance should be documented, logs preserved, and models monitored for drift and error. Independent testing and audit should be enabled, together with clear channels for litigants to contest AI-assisted steps and seek review. Staff training and vendor management are necessary, particularly where tools are procured rather than developed in-house (European Parliamentary Research Service, 2021; CEPEJ, 2018; Council of Europe, 2024).

A rights-based orientation provides the common normative core for these choices. Legal responsibility is clarified when design and governance are anchored in the protection of fundamental rights and in the duty to justify outcomes. In courts, that orientation sustains institutional legitimacy and ensures that AI supports rather than displaces adjudication (Sartor, 2020). In practice, assistive roles are most consistent with these constraints. Case triage, document analysis, and the surfacing of comparators can be enabled where auditability is maintained, and judges retain responsibility for reasons and outcomes. Substitution of

adjudicative discretion by opaque or consequentialist models would risk undermining impartiality, equality of arms, and public trust (CEPEJ, 2018; Council of Europe, 2024).

5.4. Critical synthesis

Across normative theory and practice, the lesson that emerges is that AI in courts can assist but cannot replace reason-giving adjudication. Judicial outcomes are not the product of mechanical rule application but of argumentative justification under principles and rules that can be criticised within a professional community. Any computational aid must therefore be evaluated by its capacity to support, rather than supplant, the disciplines of legal reasoning and institutional legitimacy.

From an ethical perspective, prioritising recall for the violation class was defensible where errors of omission carry heightened moral cost. This choice coheres with rights-based approaches to AI governance, where the primary question is whether protected interests are adequately safeguarded across the AI lifecycle, and whether recourse is meaningful when automation errs. In judicial contexts, such an orientation is reinforced by sector-specific guidance: the CEPEJ Ethical Charter requires respect for fundamental rights, non-discrimination, quality and security, transparency and fairness, and "under user control", which together imply that any predictive system should be advisory, auditable, and embedded in human oversight with documented data provenance and limits.

Legally, the European AI Act establishes a risk-based regime that will treat many court-adjacent systems as high-risk, triggering requirements for risk management, data governance, technical documentation, logging, human oversight, and post-market monitoring; it also introduces obligations for general-purpose models that may be fine-tuned for legal decision support. These provisions entered into force in August 2024 and will shape any path to operational use within European institutions. In parallel, the Council of Europe's Framework Convention on AI anchors deployment to human rights, democracy and the rule of law, and calls for effective, independent oversight mechanisms – expectations that fit naturally with internal audit functions and external review for judicial tools. The broader accountability debate also cautions against over-reliance on a narrow "right to explanation", emphasising system-level accountability, contestability, and better decisions rather than solely post-hoc explanations.

Institutionally, the discussion supports a limited but valuable role for AI: triage, research support, and pattern analysis. Such uses keep adjudicative discretion with judges while helping

cope with scale. Any real deployment would require expert validation, interpretability checks, and fairness assessments before integration into workflows.

Overall, the evidence recommends a conservative integration pathway. Predictors should be confined to advisory roles with clear boundaries; sensitive attributes and proxies should be minimised or neutralised; performance should be monitored via class-specific metrics that align with rights-protection goals; and independent oversight should be instituted from design through decommissioning. Ethical charters and human-rights instruments provide the north star, while binding regulation specifies the floor for compliance.

6. CONCLUSION

6.1. Objectives and findings

This thesis sets out to assess whether LLMs can forecast the outcomes of ECtHR cases based solely on applicants' petitions. The project focused on prediction rather than retrospective analysis, testing performance on more recent cases to approximate real-world deployment. Crucially, the work did not consist of merely submitting petitions to a model and requesting a violation/non-violation label. Instead, a lean but deliberate pipeline was established: prediction instructions were tested and then fixed, recognising that prompt formulation is a critical determinant of LLM behaviour; alternative summarisation prompts were evaluated, with the resulting summaries reviewed manually to judge whether they retained the legal and factual elements considered relevant for decision-making; and controlled comparisons were conducted across LLM variants and training sizes to select configurations for fine-tuning and evaluation.

After fine-tuning, GPT-4o achieved 73% accuracy on a held-out test set of 518 petitions. For the violation class, recall reached 81%, precision 76%, and the F1-score 78%, with 242 of 315 true violations correctly flagged as likely violations. These outcomes surpassed the pre-set success criteria of an F1-score of at least 67% and recall above 70% and aligned with the project's priority to minimise false negatives, given the higher ethical cost of overlooking genuine human-rights breaches.

Beyond headline metrics, the results clarify how effectiveness relates to efficiency. Moving from smaller, cheaper configurations to GPT-4o typically raised accuracy by about 5 to 8 pp and improved violation recall by roughly 4 to 6 points, but at a monetary cost that was between 3.5 and 10 times higher and with longer latency, depending on the training size. Put differently, increasing F1-score from about 0.73 in the strongest lower-cost setting (GPT-4o-mini) to 0.78 with GPT-4o required a cost premium that must be weighed against the operational priority of detecting potential violations at intake.

These findings answer the central question of why and how large language models can be applied effectively in this context. They are effective because petition texts contain regularities – fact patterns, allegations and legal cues – that pre-trained models can exploit even with limited labelled data. They can be applied effectively by constraining inputs to the petition at filing, fixing a robust prompting regime, selecting among LLM variants and training sizes with class metrics that emphasise the identification of violations, and assessing cost alongside accuracy, recall and latency so that improvements in predictive quality are balanced against real deployment constraints.

6.2. Contributions and implications

The research makes several contributions to academic and practical debates on judicial decision prediction. First, it demonstrates that LLMs can be applied beyond their common role in text generation, functioning effectively as classifiers in outcome forecasting tasks. Second, the study validates a simplified and reproducible pipeline that reduces dependence on handcrafted features, thereby lowering barriers for replication in other domains. Third, it foregrounds the higher cost of false negatives in this setting and reports results accordingly, prioritising recall where appropriate. From an institutional perspective, the findings suggest that predictive tools may help manage the ECtHR's chronic caseload by supporting early triage and prioritisation of petitions. More broadly, the study highlights the potential of AI to complement judicial processes, while recognising that significant challenges remain regarding transparency, fairness, and accountability.

6.3. Limitations and future research

Several limitations must be acknowledged. Generalisability is limited by the focus on a single court and dataset, although the pipeline is sufficiently simple to be adapted elsewhere. Automatically generated summaries were found not to harm performance in the experiments, although nuances present in full petitions may still be removed. Sensitivity to prompt wording and configuration was observed. In addition, the LLMs used remain "black boxes": only a first, human-readable glimpse of features potentially used by the model is provided by post-hoc analyses, rather than a faithful account of internal reasoning. Finally, evaluation was conducted on a test set limited to cases decided no later than 2019. Given the six-year gap to the present (2025), distribution shifts in more recent cases may not be captured, and conclusions about temporal stability and generalisability across legal domains are therefore constrained.

Future work should be extended to cross-jurisdictional settings and to larger, temporally diverse test sets, so that robustness and drift can be assessed. Greater attention should be given to explainability, with methods being adopted to identify and clearly summarise the text segments that influenced each decision. Additional priorities include error analysis by Convention article and by country, together with assessment of sensitivity to prompt wording. A useful next step is to combine petition text with structured features (e.g., respondent state, articles invoked) and test whether any accuracy gains justify the added complexity and cost. Reliability and accountability can be improved through expert review and transparent

documentation of data and prompts. If robustness and explainability are strengthened, use as a decision-support tool may be envisaged to ease caseloads while respecting judicial discretion.

7. REFERENCES

Aguiar, G. A., Rosa, D. S., & Hoch, P. A. (2024). Uso de Inteligência Artificial em Decisões Judiciais: Perspetivas, Desafios e Limites Éticos. *Anais do 7º Congresso Internacional de Direito e Contemporaneidade.* UFSM. https://www.ufsm.br/cursos/pos-graduacao/santa-maria/ppgd/congresso-direito-anais

Ahmad, S., Asghar, M. Z., Alotaibi, F. M., & Al-Otaibi, Y. D. (2022). A hybrid CNN + BILSTM deep learning-based DSS for efficient prediction of judicial case decisions. *Expert Systems with Applications*, *209*. https://doi.org/10.1016/j.eswa.2022.118318

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science, 2*. https://doi.org/10.7717/peerj-cs.93

Almuslim, I. & Inkpen, D. (2022). Legal Judgment Prediction for Canadian Appeal Cases. *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)*. https://doi.org/10.1109/cdma54072.2022.00032

Ariai, F., Mackenzie, J., & Demartini, G. (2024). *Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges*. Arxiv.org. https://doi.org/10.48550/arXiv.2410.21306

Badr, Y. (2024). Editorial: Evolution of large language models and their role in shaping general artificial intelligence. *Digital Transformation and Society, 3*(1), 1–2. https://doi.org/10.1108/dts-02-2024-088

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press. https://fairmlbook.org/

Belzuz Abogados. (2024). *The dangers of using artificial intelligence in the field of law.* https://www.belzuz.net/en/publications/in-english/item/12183-the-dangers-of-using-artificial-intelligence-in-the-field-of-law.html

Bergmann, D. (2024). *What is fine-tuning?* Ibm. https://www.ibm.com/think/topics/fine-tuning

Bertalan, V. G. F., & Ruiz, E. E. S. (2022). Using attention methods to predict judicial outcomes. *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-022-09342-7

Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of Machine Learning Research, 81,* 149-159. https://proceedings.mlr.press/v81/binns18a.html

Blank, I. A. (2023). What are large language models supposed to model? *Trends in Cognitive Sciences, 27*(11), 987–989. https://doi.org/10.1016/j.tics.2023.08.006

Bowie, N. E. (1977). Taking rights seriously [Review of the book Taking rights seriously, by R. Dworkin]. *Catholic University Law Review, 26*(4). https://scholarship.law.edu/lawreview/vol26/iss4/10

Budhiraja, A. & Sharma, K. (2022). Correlation of Language Processing and Learning Techniques for Legal Support System. *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, 305–310. https://doi.org/10.1109/com-it-con54601.2022.9850546

Canalli, R. L. (2024). Interpretable AI models for judicial decision making: beyond explicability towards legal due process. *E-Publica, 11*(1). https://doi.org/10.47345/v11n1art6

Chahal, H., Abdulla, S., Murdick, J., & Rahkovsky, I. (2021). *Mapping India's AI Potential*. Center for Security and Emerging Technology. https://doi.org/10.51593/20200096

Chen, D., Huang, Y., Ma, Z., Chen, H., Pan, X., Ge, C., Gao, D., Xie, Y., Liu, Z., Gao, J., Li, Y., Ding, B., & Zhou, J. (2023). *Data-Juicer: A One-Stop Data Processing System for Large Language Models.* ArXiv.org. https://doi.org/10.48550/arXiv.2309.02033

Council of Europe. (1950). *European Convention of Human Rights*. Rome.
Council of Europe. (2021). *The ECHR in 50 questions*.
Council of Europe. (2023). *Annual Report 2023*. https://www.coe.int/en/web/execution/annual-reports
Council of Europe. (2024). *Rules of Court*.
Council of Europe. (2024). *The Framework Convention on Artificial Intelligence.*
Council of Europe. (2025). *ECHR - Analysis of Statistics 2024*.
Devansh. (2022). *How does Batch Size impact your model learning*. Geek Culture. Medium. https://medium.com/geekculture/how-does-batch-size-impact-your-model-learning-2dd34d9fb1fa
Dikov, G. & Chernishova, O. (2023). *The European Human Rights System*. GlobaLex. https://www.nyulawglobal.org/globalex/european_human_rights_system1.html
Dittmar, L. (2024,). *What Does Transparency Really Mean in the Context of AI Governance?* OCEG. https://www.oceg.org/what-does-transparency-really-mean-in-the-context-of-ai-governance/
Dressel, J., & Farid, H. (2021). The Dangers of Risk Prediction in the Criminal Justice System. *MIT Case Studies in Social and Ethical Responsibilities of Computing.* https://doi.org/10.21428/2c646de5.f5896f9f
Edwards, L., & Veale, M. (2017). Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For. *Duke Law & Technology Review, 16*(1), 18 – 84. https://scholarship.law.duke.edu/dltr/vol16/iss1/2/
European Commission for the Efficiency of Justice (CEPEJ). (2018). *European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment*.
European Commission. (2019). *Ethics guidelines for trustworthy AI.*
European Law Institute. (2025). *ELI Case-Overload at the European Court of Human Rights: An Update*. https://www.europeanlawinstitute.eu/projects-publications/current-projects/current-projects/eli-case-overload-at-the-european-court-of-human-rights-an-update/
European Parliament. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj
Frankenreiter, J. & Nyarko, J. (2023). Natural Language Processing in Legal Tech. *Cambridge University Press EBooks,* 70–90. https://doi.org/10.1017/9781009255301.005
Greer, S. (2010). The Interpretation of the European Convention on Human Rights: Universal Principle or Margin of Appreciation? *UCL Human Rights Review, 3,* 1-14. https://hdl.handle.net/1983/1a7eca53-4a8d-4737-bcaa-95b28bf63f58
Guha, N., Nyarko, J., Ho, D. E., & Ré, C. (2025). Building GenAI Benchmarks. *Oxford University Press EBooks*. https://doi.org/10.1093/oxfordhb/9780198940272.013.0007
Hacker, P. (2020). AI Regulation in Europe. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3556532
Hagos, D. H., Battle, R., & Rawat, D. B. (2024). *Recent Advances in Generative AI and Large Language Models: Current Status, Challenges, and Perspectives.* ArXiv.org. https://doi.org/10.48550/arXiv.2407.14962
He, Z., Naphade, S., & Huang, T.-H. K. (2025). Prompting in the Dark: Assessing Human Performance in Prompt Engineering for Data Labeling When Gold Labels Are Absent. *ArXiv (Cornell University)*. https://doi.org/10.1145/3706598.3714319
Jobin, A., Ienca, M., & Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, E., Holt, X., & Wilson, N. (2025). *Improving the Accuracy and Efficiency of Legal Document Tagging with Large Language Models and Instruction Prompts*. Arxiv.org. https://arxiv.org/html/2504.09309v1

Junior, L. L. & Nogueira, A. B. (2025). O Uso da Inteligência Artificial no Processo Judicial. *Lumen et Virtus, 16*(51). https://doi.org/10.56238/levv16n51-003

Kaminski, M. E. (2018). The Right to Explanation, Explained. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.3196985

Kanaya, S. & Taylor, L. (2020). *Type I and Type II Error Probabilities in the Courtroom.* Munich Personal RePEc Archive. https://mpra.ub.uni-muenchen.de/100217/

Keeling, R., Rishi Chhatwal, Huber-Fliflet, N., Zhang, J., & Zhao, H. (2020). *Using Machine Learning on Legal Matters: Paying Attention to the Data Behind the Curtain.* UC Law SF Scholarship Repository. https://repository.uclawsf.edu/hastings_science_technology_law_journal/vol11/iss1/3/

Kitchenham, B. & Brereton, P. (2013). A systematic review of systematic review process research in software engineering. *Information and Software Technology, 55*(12), 2049–2075. https://doi.org/10.1016/j.infsof.2013.07.010

Kroll, J., Huey, J., Barocas, S., Felten, E., Reidenberg, J., Robinson, D., & Yu, H. (2017). *Accountable Algorithms. University of Pennsylvania Law Review, 165*(3), 633. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3/

Lage-Freitas, A., Allende-Cid, H., Santana, O., & Oliveira-Lage, L. (2022). Predicting Brazilian Court Decisions. *PeerJ Computer Science, 8*. https://doi.org/10.7717/peerj-cs.904

Lai, J., Gan, W., Wu, J., Qi, Z., & Yu, P. S. (2024). Large language models in law: A survey. *AI Open, 5,* 181–196. https://doi.org/10.1016/j.aiopen.2024.09.002

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529

Ma, W. (2022). Artificial Intelligence-Assisted Decision-Making Method for Legal Judgment Based on Deep Neural Network. *Mobile Information Systems*, *2022*, 1–9. https://doi.org/10.1155/2022/4636485

Management Solutions. (2023). *A ascensão dos Large Language Models: dos fundamentos à aplicação*. https://www.managementsolutions.com/pt-br/microsites/whitepapers/llm

McConnell, D. J., Zhu, J., Pandya, S., & Aguiar, D. (2021). Case-level prediction of motion outcomes in civil litigation. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law (ICAIL 2021),* 99-108. https://doi.org/10.1145/3462757.3466101

McGrath, A. & Jonker, A. (2024). *Interpretabilidade da IA*. IBM. https://www.ibm.com/br-pt/think/topics/interpretability

McGregor, L., Murray, D., & Ng, V. (2019). International human rights law as a framework for algorithmic accountability. *International and Comparative Law Quarterly, 68*(2), 309–343. https://doi.org/10.1017/s0020589319000046

Medvedeva, M. & Mcbride, P. (2023). Legal Judgment Prediction: If You Are Going to Do It, Do It Right. *Proceedings of the Natural Legal Language Processing Workshop 2023,* 73-84. https://doi.org/10.18653/v1/2023.nllp-1.9

Medvedeva, M., Üstün, A., Xu, X., Vols, M., & Wieling, M. (2021). Automatic judgement forecasting for pending applications of the European Court of Human Rights. *Proceedings of the Fifth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2021)*, 12-23. https://research.rug.nl/en/publications/automatic-judgement-forecasting-for-pending-applications-of-the-e

Medvedeva, M., Vols, M., & Wieling, M. (2019). Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law, 28,* 237-266. https://doi.org/10.1007/s10506-019-09255-y

Medvedeva, M., Wieling, M., & Vols, M. (2022). Rethinking the field of automatic prediction of court decisions. *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-021-09306-3

Mentzingen, H., Antonio, N., & Lobo, V. (2023). Joining metadata and textual features to advise administrative courts decisions: a cascading classifier approach. *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-023-09348-9

Molnar, C. (2019). *Interpretable Machine Learning: A guide for making black box models explainable*. https://christophm.github.io/interpretable-ml-book/

Muñoz Soro, J. F. & Serrano-Cinca, C. (2021). A model for predicting court decisions on child custody. *PLOS ONE*, *16*(10). https://doi.org/10.1371/journal.pone.0258993

N. Sivaranjani, Jayabharathy, J., & Teja, P. C. (2021). Predicting the supreme court decision on appeal cases using hierarchical convolutional neural network. *International Journal of Speech Technology*, *24*(3), 643–650. https://doi.org/10.1007/s10772-021-09820-4

Nuranti, E. Q., Yulianti, E., & Husin, H. S. (2022). Predicting the Category and the Length of Punishment in Indonesian Courts Based on Previous Court Decision Documents. *Computers, 11*(6). https://doi.org/10.3390/computers11060088

Nwanne, W. (2025). *General-purpose vs. reasoning models in Azure OpenAI.* Azure AI Foundry Blog – Microsoft Tech Community. https://techcommunity.microsoft.com/blog/azure-ai-foundry-blog/general-purpose-vs-reasoning-models-in-azure-openai/4403091

Ong, R. (2024). What is GPT-4o mini? How It Works, Use Cases, API & More. DataCamp. https://www.datacamp.com/blog/gpt-4o-mini

OpenAI. (2024). Hello GPT-4o. https://openai.com/index/hello-gpt-4o/

OpenAI. (2025). *API Pricing*. https://openai.com/api/pricing/

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., & McGuinness, L. A. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *British Medical Journal, 372*(71). https://doi.org/10.1136/bmj.n71

Pande, R., & Alam, S. (2020). Predicting the outcome of judicial cases using semantic analysis. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1757-1761. https://doi.org/10.1109/ssci47803.2020.9308506

Park, M., & Chai, S. (2021). AI Model for Predicting Legal Judgments to Improve Accuracy and Explainability of Online Privacy Invasion Cases. *Applied Sciences*, *11*(23), 11080. https://doi.org/10.3390/app112311080

Picinali, F. (2018). Do theories of punishment necessarily deliver a binary system of verdicts? An exploratory essay. *Criminal Law and Philosophy, 12*(4), 555–574. https://doi.org/10.1007/s11572-017-9440-y

Pillai, V.G. & Chandran, L. R. (2020). Verdict Prediction for Indian Courts Using Bag of Words and Convolutional Neural Network. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. https://doi.org/10.1109/icssit48917.2020.9214278

Preston, B. (2024). *How ChatGPT really works: What you need to know about vector embeddings*. Applied Innovation Exchange. https://blog.appliedinnovationexchange.com/how-chatgpt-really-works-what-you-need-to-know-about-vector-embeddings-7985eb26a4cf

Raza, M., Jahangir, Z., Riaz, M. B., Saeed, M. J., & Sattar, M. A. (2025). Industrial applications of large language models. *Scientific Reports, 15*(1). https://doi.org/10.1038/s41598-025-98483-1

Rodrigues, J. (2021). *Modelação e previsão de decisões judiciais utilizando um repositório de sentenças* [Master's thesis, Universidade do Minho]. Repository.

Rosili, N. A. K., Hassan, R., Zakaria, N. H., Kasim, S., Rose, F. Z. C., & Sutikno, T. (2021). A systematic literature review of machine learning methods in predicting court decisions. *IAES International Journal of Artificial Intelligence, 10*(4), 1091. https://doi.org/10.11591/ijai.v10.i4.pp1091-1102

Sartor, G. (2020). Artificial intelligence and human rights: Between law and ethics. *Maastricht Journal of European and Comparative Law, 27*(6), 705–719. https://doi.org/10.1177/1023263x20981566

Schmallenbach, L., Bärnighausen, T. W., & Lerchenmueller, M. J. (2024). The global geography of artificial intelligence in life science research. *Nature Communications, 15*. https://doi.org/10.1038/s41467-024-51714-x

Selbst, A., & Barocas, S. (2018). The Intuitive Appeal of Explainable Machines. *Fordham Law Review, 87*(3), 1085. https://ir.lawnet.fordham.edu/flr/vol87/iss3/11

Sert, M. F., Yıldırım, E., & Haşlak, İ. (2021). Using Artificial Intelligence to Predict Decisions of the Turkish Constitutional Court. *Social Science Computer Review*. https://doi.org/10.1177/08944393211010398

Shaikh, R. A., Sahu, T. P., & Anand, V. (2020). Predicting Outcomes of Legal Cases based on Legal Factors using Classifiers. *Procedia Computer Science*, *167*, 2393–2402. https://doi.org/10.1016/j.procs.2020.03.292

Shu, D., Zhao, H., Liu, X., Demeter, D., Du, M., & Zhang, Y. (2024). *LawLLM: Law Large Language Model for the US Legal System.* ArXiv.org. https://arxiv.org/html/2407.21065v1

Silva, R. L. A., Toledo, C., & Duarte, L. G. M. (2025). Teoria da Argumentação Jurídica, Teoria dos Direitos Fundamentais e Teoria dos Princípios em Robert Alexy. *Aracê, 7*(7), 38847-38864. https://doi.org/10.56238/arev7n7-211

Sukanya, G. & Priyadarshini, J. (2024). Hybrid CNN: An Empirical Analysis of Machine Learning Models for Predicting Legal Judgments. *International Journal of Advanced Computer Science and Applications, 15(7)*. https://doi.org/10.14569/ijacsa.2024.01507124

Teigão, R. C. & Fogaça, L. A. F. (2025). O Uso Ético e Responsável da Inteligência Artificial no Judiciário Brasileiro: Impactos para o Jurisdicionado e Balizas Regulatórias. *Gralha Azul, 1*(28). https://revista.tjpr.jus.br/gralhaazul/article/view/189

Umamaheswari, S., Aartisha, S., Kanimozhi, J., & Suhashini, R. (2023). Building accurate legal case outcome prediction models. *2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*. 1–6. https://doi.org/10.1109/icaeca56562.2023.10200651

Veale, M. & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act: Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International, 22*(4), 97–112. https://doi.org/10.9785/cri-2021-220402

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law, 7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Wang, L., Chen, X., Deng, X., Wen, H., You, M., Liu, W., Li, Q., & Li, J. (2024). Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *Npj Digital Medicine*, *7*(1), 1–9. https://doi.org/10.1038/s41746-024-01029-4

Wang, Y., Gao, J., & Chen, J. (2020). Deep Learning Algorithm for Judicial Judgment Prediction Based on BERT. *2020 5th International Conference on Computing, Communication and Security (ICCCS)*, *46*, 1–6. https://doi.org/10.1109/icccs49678.2020.9277068

Wildhaber, L. (2025). The European Court of Human Rights: The Past, The Present, The Future. *American University International Law Review, 22*(4), 521-538. https://digitalcommons.wcl.american.edu/auilr/vol22/iss4/2/

Yang, Z. (2024). *Comparing Traditional Machine Learning and Large Language Models: An Application to Mental Health Text Classification* [Master's thesis, University of California]. Escholarship.org. https://escholarship.org/uc/item/0d63p0jj

Zahir, J. (2023). Prediction of court decision from Arabic documents using deep learning. *Expert Systems*, *40*(6). https://doi.org/10.1111/exsy.13236

Zane, X. (2025). GPT-3.5: *The comprehensive guide to OpenAI's advanced language model in 2025*. BytePlus. https://www.byteplus.com/en/topic/514132?title=gpt-3-5-the-comprehensive-guide-to-openai-s-advanced-language-model-in-2025

Zheng, X., Liu, Y., Jiang, J., Thomas, L. M., & Su, N. (2021). PREDICTING THE LITIGATION OUTCOME OF PPP PROJECT DISPUTES BETWEEN PUBLIC AUTHORITY AND PRIVATE PARTNER USING AN ENSEMBLE MODEL. *Journal of Business Economics and Management*, *22*(2), 320–345. https://doi.org/10.3846/jbem.2021.13219

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020). How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.466

8. APPENDICES

Appendix A: Dataset variables

Variable	Type	Role	Description
itemid	Identifier	Metadata	Identifier assigned to each communicated case
docname	String	Metadata	Official name of the case as published
appno	Identifier	Metadata	Application number assigned by the ECtHR
kpdate	Date	Metadata	Date when the case was communicated by the Court (key procedural date)
respondent	Categorical	Metadata	Country or countries against which the complaint was filed
url	String	Metadata	Link to the official HUDOC page of the communicated case
text	Text	Input	Full text of the application (includes both facts and questions)
violation	Binary	Target	1 = Violation; 0 = No violation
date_label	Date	Metadata	Date of the final judgment issued by the Court
facts	Text	Contextual	Description of facts presented by the applicant
questions	Text	Contextual	Legal complaints or articles invoked by the applicant
itemid_final_judgement	Identifier	Metadata	Identifier of the corresponding final judgment
facts_final_judgement	Text	Contextual	Description of facts as stated by the Court in its final judgment

Appendix B: Alternative summarisation prompts

Prompt A:

> Summarize the following petition submitted to the European Court of Human Rights. The summary should be concise and clear, capturing the essential facts and legal concerns raised by the applicant. Avoid excessive detail and focus on the overall context and claims.

Prompt B:

> Summarize the main legal and factual aspects of the following petition submitted to the European Court of Human Rights. Emphasize the key facts of the case, the rights alleged to have been violated under the European Convention on Human Rights, and the core legal arguments made by the applicant. Keep the summary informative yet concise.

Appendix C: Alternative prediction prompts

Prompt 1:

> You are a legal assistant classifying petitions submitted to the European Court of Human Rights.
> Base your decision on:
> - Legal criteria from the European Convention of Human Rights;
> - The nature of the allegation, factual and historical context;
> - National court decisions;
> - Objective facts and strength of evidence.
> Respond with:
> 0 – no violation
> 1 – violation
> Only output 0 or 1.

Prompt 2:

> Analyze the content of the input text to determine whether it indicates a violation of a human right.
> Please note that not all cases constitute human rights violations and that the text also takes the subjective perspective of the applicant. So, use your judgment carefully, considering the facts of the case.
> For example, an application that details arbitrary detention without due process may constitute a violation (1), whereas an application concerning the enforcement of previously agreed upon international or bilateral agreements might not constitute a violation (0).
> The output should be only 1 (violation) or 0 (no violation), not text.

Prompt 4:

> You are a legal assistant helping to classify petitions submitted to the European Court of Human Rights. Your task is to classify each petition as '0' (non-violation) or '1' (violation) based on the criteria and facts presented. To make your decision, consider the following points:
> - Legal Criteria: the court bases its decisions on the European Convention of Human Rights;
> - Case Details: analyze the facts, including the nature of the allegation, the evidence provided, and the historical and factual context;
> - Previous Agreements: check whether there are prior agreements between the parties involved; the existence of such agreements may indicate that the right has not been violated, unless there is clear proof that the agreement has been violated;
> - Objectivity: ignore the petitioner's subjectivity and emotional claims; focus on objective facts and the application of legal criteria;
> - Documentary and Testimonial Evidence: evaluate the strength of the documentary and testimonial evidence presented;
> - National Court Decisions: consider the decisions made by national courts.
> When making your classification, keep in mind that not all petitions result in a finding of violation. Carefully weigh the evidence for both potential outcomes.
> Based on these points, classify the petition as follows:
> '0' for non-violation
> '1' for violation
> The output must contain only 0 or 1.

Appendix D: Top Bigrams by predicted class