LLM-based extraction and prediction of Slovak criminal court decisions.

Matúš Kundracik[1], Zoltán Szoplák[1], Peter Gurský[1] and Dávid Varga[1]

[1]Institute of Computer Science, Faculty of Science, Pavol Jozef Šafárik University in Košice, Jesenná 5, 040 01 Košice, Slovakia

Abstract

This paper presents a system for automated extraction of structured data from Slovak criminal court decisions and their usage in subsequent verdict prediction using Large Language Models. The data extracted by prompt engineering combined with domain knowledge is then used by a cascading prediction framework combining deterministic rules with Retrieval-Augmented Generation (RAG) based on similar cases. The system achieved 100% accuracy in binary verdict classification (acquittal vs. conviction) and delivered mixed results when it came to predict sentence types and their magnitude (be it length, sum, etc.) while also revealing some limitations of generative AI in legal decision analysis.

Keywords

legal texts, information extraction, verdict prediction, large language models, natural language processing, prompt engineering

1. Introduction

The existing verdicts of past judicial decisions constitute an important source of information on court decision-making and the practical application of criminal law. However, most of these documents are available only as unstructured text, which significantly limits automated analysis, statistical processing, and efficient retrieval of relevant case law. Transforming judicial decisions into a structured format and creating subsequent predictions can therefore assist legal experts in analyzing case law, explore hypothetical cases as well as detect anomalies by comparing where actual past verdicts and automated models different.

Traditional approaches based on regular expressions and manually designed rules are unable to reliably capture the intricate multivariate features unique to every case that affects final decision making, but recent advances in large language models (LLMs) offer us another avenue.

The aim of this study is to design and implement a system for the automated extraction of structured information from Slovak criminal court decisions and to utilize the extracted data for predicting judicial outcomes and imposed sanctions.

The proposed solution employs the Llama-3.3 70B generative model and combines full-text analysis with deterministic rules and the Retrieval-Augmented Generation (RAG) framework.

2. Related Work

Early Legal Judgment Prediction (LJP) systems relied on traditional machine learning algorithms combined with sparse text representations such as Bag-of-Words and TF-IDF. Case descriptions or judicial opinions were transformed into feature vectors and classified using Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Random Forests, and Gradient Boosting, explored in [2]

ITAT'26: Information Technologies – Applications and Theory, September 25–29, 2026, Biele Karpaty, Slovakia

*Corresponding author.

†These authors contributed equally.

matus.kundracik@student.upjs.sk. (M. Kundracik); zoltan.szoplak@student.upjs.sk (Z. Szoplák); peter.gursky@upjs.sk (P. Gurský); david.varga@student.upjs.sk (D. Varga)

0000-0003-1823-0536 (Z. Szoplák); 0000-0002-4744-7390 (P. Gurský); 0000-0002-3176-8106 (D. Varga)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-09

CEUR-WS.org/Vol-4323/paper21.pdf

These approaches achieved competitive performance on relatively simple tasks, such as binary judgment prediction or basic offense classification. However, their effectiveness was limited by their dependence on manually engineered features and their inability to capture the rich semantic and contextual information inherent in legal documents. Furthermore, many early studies were conducted on relatively small datasets from a single jurisdiction, limiting their generalizability.

Traditional classifiers also struggled with multi-output legal prediction, where a single case may involve multiple defendants, offenses, penalties, and mitigating or aggravating circumstances.

As stated in [9], the introduction of deep learning significantly improved legal text analysis. Recurrent neural networks (LSTM and GRU), convolutional neural networks, and hierarchical attention architectures enabled models to capture contextual dependencies across words, sentences, and document sections more effectively than traditional machine learning methods.

Subsequently, transformer-based language models, particularly domain-specific variants such as LegalBERT, established new state-of-the-art performance on legal text classification, citation extraction, and judgment prediction tasks. Nevertheless, standard Transformer architectures remain constrained by quadratic computational complexity with respect to input length, making the processing of lengthy judicial decisions computationally expensive. As a result, long documents are typically segmented into smaller chunks, which may disrupt logical dependencies between distant sections of the text and negatively affect downstream reasoning.

Recent advances have shifted the focus toward Large Language Models (LLMs), including GPT and LLaMA families, which demonstrate substantially stronger reasoning capabilities and a superior understanding of complex legal language. Their ability to process long contextual information and perform instruction-following makes them particularly attractive for information extraction and legal reasoning. However, standalone LLMs remain susceptible to factual hallucinations, inconsistent reasoning, and domain-specific inaccuracies.

Consequently, recent research increasingly favors hybrid architectures that integrate LLMs with external legal knowledge bases, rule-based components, or Retrieval-Augmented Generation (RAG) techniques to improve factual consistency and reliability, such as described in [8]

One representative hybrid framework is LegalReasoner, described in [1] which combines domain-adapted language models with graph neural networks and multi-stage reasoning. The framework first performs legal knowledge infusion through additional pre-training on statutes, case law, and legal literature. It then represents legal entities and judicial decisions as a knowledge graph, allowing graph neural networks to retrieve relevant precedents and statutory provisions. Finally, hierarchical transformer reasoning synthesizes factual information, retrieved evidence, and applicable legal rules to produce legally coherent predictions. Experimental evaluation on the ECHR [6] and CAIL2018 [7] datasets demonstrates consistent improvements over previous state-of-the-art approaches.

Beyond transformer-based models, recent research has also explored State Space Models (SSMs) for efficient long-sequence processing, such as described in [10] The MambaEffNet architecture combines EfficientNet-inspired feature extraction with the Mamba state-space model, convolutional layers, multi-head attention, and multilayer perceptrons to process structured judicial data with linear computational complexity.

Unlike standard Transformers, Mamba-based architectures scale efficiently to long sequences while preserving relevant contextual information. The proposed framework further addresses missing data through GAN-based data imputation, resulting in higher prediction accuracy and F1 scores compared with conventional machine learning baselines.

Another important research direction focuses on improving the reasoning capabilities and explainability of LLMs through prompt engineering. Rather than directly predicting judicial outcomes, Legal Syllogism Prompting (LoT), described in [4] guides LLMs to follow the classical structure of legal reasoning by explicitly identifying the applicable legal rule (major premise), analyzing the facts of the case (minor premise), and deriving the final legal conclusion.

This structured reasoning process improves prediction accuracy while providing transparent intermediate reasoning steps that legal practitioners can verify. Notably, LoT achieves these improvements without requiring additional model fine-tuning, making it an efficient zero-shot prompting strategy.

Despite these advances, recent benchmarking studies indicate that general-purpose LLMs still face important limitations in Legal Judgment Prediction. Their performance degrades in zero-shot settings without relevant contextual examples, and retrieval systems alone may occasionally outperform LLM-enhanced pipelines when high-quality similar cases are available.

Moreover, domain-specific transformer models, such as LegalBERT, described in [5] continue to achieve more stable performance than general-purpose LLMs in supervised settings with sufficient labeled training data. These findings suggest that hybrid architectures combining retrieval, structured reasoning, and domain adaptation currently represent the most promising direction for practical legal AI systems.

3. Methods

This section describes the methodology used to transform unstructured Slovak criminal court decisions into structured representations suitable for downstream legal analytics and the subsequent usage of this representation to perform Legal Judgment Prediction (LJP). The proposed framework consists of two sequential components. The first component extracts legally relevant information from judicial decisions using a Large Language Model (LLM) supported by deterministic rules and retrieval mechanisms. The second component utilizes the extracted information to predict judicial outcomes and imposed criminal sanctions through a hybrid reasoning pipeline. The framework combines full-document processing, domain-specific prompt engineering, deterministic post-processing, and Retrieval-Augmented Generation (RAG) to improve both extraction quality and prediction reliability.

3.1. Information Extraction Framework

The primary objective of the extraction framework is to transform lengthy, unstructured judicial decisions into a structured representation while preserving legally significant relationships between extracted entities. Traditional information extraction systems rely heavily on regular expressions, manually engineered heuristics, or deterministic parsing rules. Although these approaches perform well for highly standardized documents, they struggle with the semantic variability characteristic of judicial writing. Criminal judgments frequently express identical legal concepts using substantially different linguistic formulations, making purely rule-based extraction unreliable. To overcome these limitations, the proposed framework employs a Large Language Model as the principal semantic extraction engine. Rather than attempting to identify isolated textual patterns, the language model processes the judicial decision as a coherent legal document and infers relationships between offenses, statutory references, criminal sanctions, evidence, and judicial reasoning. Nevertheless, unrestricted LLM generation introduces additional challenges, including factual hallucinations, inconsistent terminology, and structurally invalid outputs. Consequently, the proposed architecture combines neural reasoning with deterministic validation rules and constrained prompting. Ideally, this hybrid design preserves the semantic capabilities of the language model while ensuring that the extracted information remains legally consistent.

3.2. Evolution of the Extraction Strategy

Several extraction strategies were investigated during system development before selecting the final architecture. The evolution of the extraction pipeline was primarily driven by the need to preserve semantic consistency across lengthy judicial decisions while minimizing computational complexity. Initial experiments followed the conventional approach adopted by many transformer-based document processing systems, in which long documents are divided into smaller segments before inference. Although computationally efficient, this strategy consistently produced incomplete and internally inconsistent extractions because important legal relationships frequently spanned multiple text segments. Subsequent experiments investigated semantic retrieval mechanisms designed to reduce context length while preserving the most legally relevant portions of the document. Although retrieval substantially

improved efficiency, important dependencies occasionally remained fragmented across retrieved passages. Based on these observations, the final framework adopted full-document processing whenever it was feasible and only used relevant part extraction if the decision didn't fit into the context window of the language model. Processing the complete judicial decision enabled the language model to construct a unified semantic representation of the case, significantly improving extraction consistency while reducing logical contradictions between extracted entities. The following subsections describe each investigated strategy in detail together with the motivation for the final architectural decisions.

3.2.1. Chunk-Based Processing

The initial implementation employed a chunk-based processing strategy in which every judicial decision was divided into smaller text segments compatible with the language model's context window. Each chunk was processed independently, after which the extracted information was merged into a single structured representation. This approach significantly reduced computational requirements and enabled processing of arbitrarily long judicial decisions. However, experimental evaluation revealed several important limitations. Many legally significant relationships extend across multiple paragraphs or document sections. Monetary fines, substitute imprisonment sentences, probation periods, aggravating circumstances, and statutory references frequently appear in different parts of the judicial decision. Because individual chunks were processed independently, the language model often failed to associate related legal concepts correctly. For example, substitute imprisonment sentences were frequently extracted as independent punishments instead of conditional sanctions associated with unpaid monetary fines. Similar inconsistencies occurred when linking criminal offenses with mitigating circumstances or judicial reasoning distributed across multiple sections of the decision. Several improvements to chunk-based processing were evaluated, including overlapping windows, hierarchical summarization, and multi-stage extraction using specialized prompts. Although these techniques slightly improved extraction quality, none consistently preserved the complex semantic relationships characteristic of criminal court decisions. These observations motivated the investigation of retrieval-based alternatives capable of preserving greater contextual coherence while reducing the amount of processed text.

3.2.2. Semantic Retrieval of Relevant Context

To reduce the amount of text processed by the language model while preserving semantic coherence, a retrieval-based preprocessing strategy was investigated. Rather than embedding complete text segments directly, each document chunk was first transformed into a compact semantic representation using statistically significant keywords. The proposed retrieval pipeline differs from conventional document retrieval approaches based on TF–IDF. Instead of comparing multiple independent documents within an external corpus, the complete judicial decision is treated as the corpus, while each document chunk represents an individual document. This formulation allows the retrieval process to identify the portions of a judicial decision that contain the highest concentration of legally relevant information. Following lemmatization and the removal of common Slovak stop words and frequent legal boilerplate expressions, candidate unigrams, bigrams, and trigrams are subsequently ranked according to their TF–IDF scores. The highest-ranked keywords are embedded using a sentence embedding model. Then, individual keyword embeddings are averaged into a single dense vector representing each document chunk. During retrieval, user-defined extraction queries are embedded in the same semantic space, and cosine similarity is used to identify the most relevant portions of the judicial decision.

Compared with exhaustive full-document processing, this retrieval stage substantially reduced computational cost while preserving most legally relevant information. Nevertheless, empirical evaluation revealed that retrieval alone could not completely eliminate the fragmentation of legal reasoning. Important relationships occasionally remained distributed across multiple retrieved passages, particularly when judicial reasoning referred to evidence, statutory provisions, and sentencing decisions located in different sections of the judgment. Merging the retrieved segments into a single text offered some improvements, it remained inferior to processing the complete judicial decision whenever the context

size allowed for it. Regardless, for many decisions not fitting the aforementioned criteria, this approach offered a viable alternative.

3.2.3. Hybrid Offense Identification

Allowing the language model to generate offense names freely resulted in inconsistent terminology and occasional hallucination. Similar criminal offenses were frequently expressed using different linguistic formulations, making statistical analysis difficult. To improve consistency, offense identification was reformulated as a constrained selection task. Instead of producing arbitrary offense names, the framework first retrieves a short list of candidate criminal offenses from an external database of Slovak criminal legislation using semantic and TF–IDF keyphrase matching. The retrieved candidates are incorporated into the language model prompt, which is instructed to select one or more valid offenses exclusively from the supplied list. If the model fails to produce a valid selection, a deterministic fallback mechanism assigns the highest-ranked retrieved candidate. Restricting the output space in this manner significantly reduced terminology inconsistencies while preventing the generation of nonexistent legal offenses.

3.2.4. Prompt Engineering and Domain-Specific Constraints

A common challenge involved mitigating and aggravating circumstances that were referenced only through statutory citations, such as Sections 36 and 37 of the Slovak Criminal Code, without explicitly stating their legal descriptions. Rather than requiring the language model to infer the meaning of these references, the complete statutory definitions were embedded directly into the system prompt. Consequently, the model was able to associate abbreviated legal citations with their corresponding legal interpretations while preserving consistency across different judicial decisions.

Penalty extraction was further extended through a structured output schema capable of representing conditional attributes associated with criminal sanctions. Several sanctions, including activity bans, confiscation orders, and substitute imprisonment, contain additional legal information describing either the object of the sanction or the conditions under which it becomes enforceable. Explicitly modelling these attributes enabled the extraction framework to preserve legally meaningful relationships instead of treating each sanction as an independent entity.

The prompt, put together using the guidelines in [12] additionally distinguishes between primary and substitute criminal sanctions. This distinction prevents substitute imprisonment imposed for failure to pay a monetary fine from being interpreted as an independent sentence, thereby preserving the hierarchical structure of judicial sanctions.

Finally, evidence extraction was constrained using strict negative prompting. Certain abbreviated criminal orders contain only the operative part of the judgment without a detailed reasoning section. Preliminary experiments demonstrated that language models occasionally hallucinated plausible evidence in such cases. Explicit instructions prohibiting unsupported evidence extraction substantially reduced these hallucinations while preserving high recall for judgments containing complete legal reasoning.

3.3. Structured Output Representation

The extraction framework produces a standardized JSON representation designed to preserve legally relevant relationships while remaining directly usable for downstream statistical analysis and machine learning.

The generated structure consists of three principal categories.

First, general case metadata describe the judicial decision itself, including identifiers, referenced statutory provisions - extracted using [3], and the factual description of the criminal act.

Second, sentencing information stores every identified criminal sanction together with its legally relevant attributes. These include sanction type, duration, conditional status, substitute penalties, probation periods, confiscated objects, and other sanction-specific properties.

Third, contextual legal information includes identified criminal offenses, mitigating and aggravating circumstances, and categorized evidence extracted from the judicial reasoning.


```
{
    "decision_id": "1T/129/2013__2013-08-13__ECLI:SK:OSBA3:2013:1313010483.1",
    "analyza": {
        "oslobodenie": false,
        "tresty": [
            {
                "typ_trestu": "peňažný trest",
                "miera_trestu": "1.500,-€",
                "detail": "",
                "podmienka": "nepodmienečne",
                "kategoria_trestu": "Hlavný trest"
            },
            {
                "typ_trestu": "náhradný trest odňatia slobody",
                "miera_trestu": "6 mesiacov",
                "detail": "",
                "podmienka": "náhradný",
                "kategoria_trestu": "Vedlajší trest"
            },
            {
                "typ_trestu": "zákaz činnosti",
                "miera_trestu": "40 mesiacov",
                "detail": "viesť motorové vozidlá všetkého druhu",
                "podmienka": "nepodmienečne",
                "kategoria_trestu": "Vedlajší trest"
            }
        ],
        "okolnosti": {
            "polahcujuce": [
                "1) priznal sa k spáchaniu trestného činu a trestný čin úprimne oľutoval"
            ],
            "pritazujuce": []
        },
        "dokazy": {
            "vysluchy": [],
            "listinne": [
                "rozhodnutie",
                "zmluva",
                "výpis z registra trestov",
                "lekárske správy"
            ],
            "znalecke": [],
            "vecne_a_zaznamy": [
                "nákladné motorové vozidlo Seat Inca, červenej farby, EČV: U.-XXXAV",
                "vozidlo Honda CR-V, EČV: U.-XXXTX"
            ],
            "ine": []
        }
    }
},
```


Figure 1: Example JSON representation of extracted features

3.4. Judicial Outcome Prediction

Following information extraction, a second processing stage predicts judicial outcomes directly from the structured representation.

The prediction framework pursues two objectives:

1. determine whether the judicial decision resulted in a conviction or an acquittal;
2. predict the imposed criminal sanctions while distinguishing between primary and secondary penalties.

To ensure compatibility between extraction and prediction, the prediction module produces a JSON representation identical to that generated during information extraction. This design enables direct comparison between predicted outputs and ground-truth annotations during evaluation.

3.4.1. Hierarchical Prediction Pipeline

The prediction framework follows a hierarchical three-stage architecture designed to maximize prediction accuracy while minimizing computationally expensive language model inference.

The first stage performs deterministic filtering of judicial decisions that do not contain a final verdict. Decisions involving procedural termination, referral to another authority, or similar procedural outcomes are identified through statutory references and excluded from further prediction.

The second stage performs rule-based verdict classification using structured legal references extracted during the previous phase. References to provisions governing acquittals are interpreted as strong

indicators of a not-guilty verdict, whereas statutory provisions concerning sentencing, convictions, or dismissal of appeals indicate conviction.

Only cases containing ambiguous or conflicting legal signals proceed to the final prediction stage.

The third stage employs the Llama 3.3 language model, described in [11] to resolve these ambiguous cases. Rather than processing the complete judicial decision again, the model receives only the structured information generated during extraction, including factual descriptions, identified offenses, mitigating and aggravating circumstances, evidence, and statutory references.

The language model is constrained through carefully designed prompts specifying the required JSON schema while explicitly prohibiting common prediction errors observed during preliminary experiments.

3.4.2. Retrieval-Augmented Prediction

Preliminary experiments demonstrated that deterministic rules alone were insufficient for predicting detailed sentencing information. Consequently, penalty prediction was delegated entirely to the language model while retaining deterministic verdict classification.

To improve prediction consistency, the framework incorporates Retrieval-Augmented Generation (RAG).

For every judicial decision, semantic embeddings are generated from the factual description, extracted evidence, and contextual legal information while intentionally excluding statutory references to avoid trivial retrieval based solely on identical legal provisions.

Similarity between judicial decisions is computed using cosine similarity over embeddings generated by the `bge-m3` embedding model.

The three most semantically similar convicted cases are retrieved dynamically and inserted into the language model prompt as few-shot demonstrations.

Restricting retrieval exclusively to convicted cases prevents the model from observing acquittal examples containing empty sentencing information, thereby improving the consistency of generated penalties.

3.4.3. Post-Processing and Output Validation

Every prediction produced by the language model undergoes deterministic post-processing before being accepted as the final output.

The validation stage performs several consistency checks.

First, obligations that do not constitute criminal sanctions, such as civil compensation for damages, are removed from the generated prediction.

Second, conditional attributes are verified to ensure compatibility with the corresponding penalty type.

Third, penalties are normalized according to predefined legal categories distinguishing primary sanctions from secondary or substitute penalties.

Finally, because language model responses occasionally contain Markdown formatting or explanatory text surrounding the generated JSON object, a fault-tolerant parser extracts the first syntactically valid JSON structure using depth-aware bracket matching. Invalid outputs are regenerated automatically, with a maximum of three inference attempts.

These deterministic validation procedures substantially reduce structurally invalid predictions while preserving legally meaningful relationships between extracted entities.

3.5. Dataset and Implementation

The proposed framework was implemented in Python using the Ollama inference framework, enabling fully local deployment of large language models without transmitting judicial documents to external cloud services. Local inference is particularly important in legal applications because judicial decisions frequently contain sensitive personal information protected by privacy legislation. All experiments were

performed using the Llama 3.3 70B instruction-tuned language model. The model was selected because of its strong multilingual capabilities, competitive reasoning performance, and sufficiently large context window to process complete judicial decisions. Unlike earlier transformer-based systems requiring aggressive document segmentation, Llama 3.3 allows substantially longer contexts while maintaining coherent reasoning across distant portions of legal documents. The input dataset consists of Slovak criminal court decisions stored in JSON format. Each record contains a unique case identifier together with the complete textual content of the judicial decision. The extraction framework transforms every document into a standardized JSON representation containing structured legal information required for subsequent statistical analysis and prediction. The extracted representation includes the following categories of information:

- general case metadata,
- factual description of the criminal act,
- identified criminal offenses,
- judicial verdict,
- imposed criminal sanctions,
- mitigating circumstances,
- aggravating circumstances,
- referenced statutory provisions,
- evidence presented during the proceedings.

The resulting representation preserves the logical relationships between legal entities while remaining directly convertible into relational tables suitable for statistical analysis and machine learning.

4. Evaluation

4.1. Experimental Setup

The proposed prediction framework was evaluated by comparing its outputs against the structured annotations produced by the extraction pipeline, which served as the ground truth. Although the extraction dataset contained 78 judicial decisions and the prediction system processed 50 cases, only decisions successfully processed by both pipelines were included in the evaluation. After excluding skipped and incomplete records, the final evaluation consisted of **46 matched criminal cases**. These were then manually annotated by legal experts and compared to the results of our models.

Three complementary tasks were evaluated:

- Verdict prediction (conviction vs. acquittal),
- Penalty type prediction, and
- Penalty magnitude prediction (e.g., imprisonment duration).

Binary verdict prediction was evaluated using Accuracy and F1-score. Penalty type prediction was assessed using Exact Match accuracy and the Jaccard similarity coefficient between the predicted and ground-truth penalty sets. Quantitative sentence prediction was evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

4.2. Verdict Prediction

The proposed hierarchical architecture achieved perfect performance on binary verdict classification. All **46 judicial decisions** were classified correctly, yielding an **Accuracy** and **F1-score** of **1.00**. The confusion matrix (Figure 2) shows that the model correctly identified all **43 convictions** and **3 acquittals**, without producing any false positives or false negatives.

These results demonstrate that statutory references constitute highly reliable indicators of judicial outcomes. The rule-based component was therefore sufficient for this task, eliminating the need for LLM inference in the majority of cases.

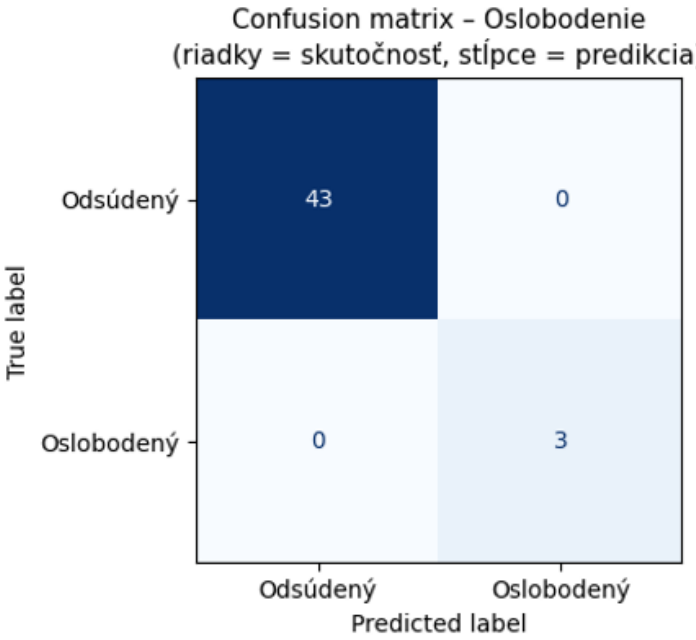


Figure 2: Confusion matrix of binary verdict prediction - conviction (odsúdený) vs acquittal (oslobodený)

4.3. Penalty Type Prediction

Predicting the imposed penalties proved substantially more challenging than verdict classification. The system achieved an **Exact Match** score of **10.9%**, indicating that the complete set of predicted penalties exactly matched the ground truth in approximately one out of every ten evaluated cases. The average **Jaccard similarity** reached **0.346**, suggesting moderate overlap between predicted and actual penalty sets.

Qualitative analysis revealed a systematic tendency to over-predict common sanctions, particularly imprisonment and probation periods, while under-predicting less frequent secondary penalties such as substitute imprisonment, probation supervision, or compensation-related measures. These findings indicate that the model successfully captures the dominant sentencing patterns but struggles with legally complex combinations of multiple sanctions.

4.4. Penalty Magnitude Prediction

The weakest performance was observed when predicting quantitative sentencing attributes. For imprisonment durations expressed in months, the system achieved a **Mean Absolute Error (MAE)** of **8.68 months** and a **Root Mean Squared Error (RMSE)** of **15.02 months**. Considering imprisonment sentences alone, the average prediction error was **7.88 months**.

The error distribution (Figure 3) demonstrates that while many predictions remain relatively close to the ground truth, several substantial outliers are present. These large deviations confirm that, although the model is capable of inferring appropriate categories of punishment, it cannot reliably estimate exact sentencing lengths without explicit access to the sentencing section of the judicial decision.

4.5. Discussion

The evaluation demonstrates a clear distinction between qualitative and quantitative legal prediction. Judicial outcomes can be predicted with extremely high reliability using deterministic legal signals, whereas predicting the complete structure of criminal sanctions remains considerably more difficult. The primary limitation concerns secondary penalties and exact sentencing magnitudes, both of which

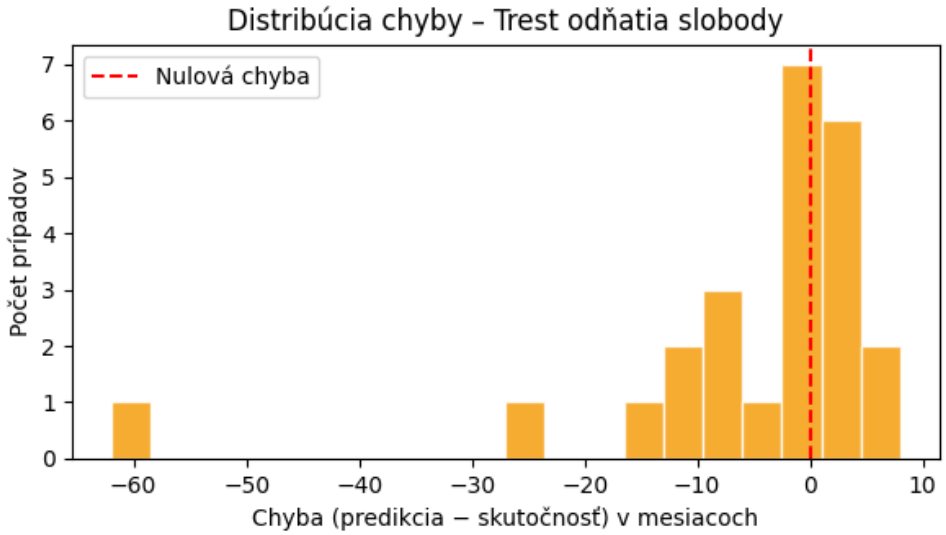


Figure 3: Distribution of error in months (reality vs. prediction) - imprisonment duration

depend on detailed legal reasoning and information contained within the sentencing section rather than the factual description of the offense alone.

These findings also validate the design choices adopted in the proposed hybrid architecture. Deterministic rule-based reasoning is sufficient for verdict classification, while Retrieval-Augmented Generation provides measurable improvements for predicting penalty types. However, accurate prediction of sentence duration will likely require enriching the retrieval context with explicit sentencing passages or incorporating additional structured judicial metadata.

5. Conclusion

The aim of our research was to design and implement a system for the automated extraction of structured information from Slovak criminal court decisions and subsequently utilize the extracted data for verdict prediction. This objective was successfully achieved through the development of a functional hybrid solution combining natural language processing techniques, large language models, deterministic rule-based methods, and a Retrieval-Augmented Generation (RAG) approach.

Experimental evaluation demonstrated that the proposed system achieves excellent performance in binary verdict classification. For the task of distinguishing between conviction and acquittal, the model achieved 100% accuracy on the evaluated dataset. Predicting the specific types of imposed penalties proved to be considerably more challenging. While the model successfully identified the primary sanctions in a substantial proportion of cases, accurately predicting secondary penalties remains an open research problem. The greatest limitation was observed in predicting the exact sentence length or monetary penalty. The experiments confirmed that, without explicit access to the sentencing section of the court judgment, large language models tend to generate statistically plausible numerical values rather than the actual imposed sentence.

The research also demonstrated the considerable potential of large language models for the analysis of legal documents. The combination of deterministic algorithms, domain-specific rules, and modern generative AI models resulted in a practical system capable of processing large collections of criminal court decisions. At the same time, the findings highlight that legal-domain applications require robust validation mechanisms, post-processing procedures, and domain knowledge to minimize hallucinations and ensure consistent, legally accurate outputs.

Several directions for future work have been identified. These include expanding the dataset to a substantially larger collection of court decisions, integrating a dedicated vector database to efficiently

scale the Retrieval-Augmented Generation architecture, and enriching the model input with the sentencing section of court judgments. Another promising direction is domain-specific fine-tuning of large language models using Slovak legal corpora. Finally, we also intend to explore traditional machine learning methods that work with structured tabular data and compare them to LLM-based prediction.

Acknowledgments

This article was supported by the Slovak Research and Development Agency under Contract No. VV-MVP-24-0038 – project ALIT: Analysis of liability for offenses committed on the Internet using machine learning methods.

Declaration on Generative AI

During the preparation of this manuscript, the authors used ChatGPT-4o exclusively for language editing, grammatical corrections, stylistic refinement, and improvements in the clarity, structure, and presentation of the manuscript.

The AI tool was not used to generate scientific findings, develop algorithms, process or analyze experimental data, or draw scientific conclusions. All research activities, including system design, implementation, experimental evaluation, statistical analysis, and interpretation of results, were performed independently by the authors.

The authors carefully reviewed and verified all AI-assisted revisions and take full responsibility for the accuracy, originality, and scientific integrity of the published work.

References

[1] X. Wang, X. Zhang, V. Hoo, Z. Shao, and X. Zhang, "LegalReasoner: A Multi-Stage Framework for Legal Judgment Prediction via Large Language Models and Knowledge Integration," *IEEE Access*, vol. 12, pp. 166843–166854, 2024. doi: 10.1109/ACCESS.2024.3496666.

[2] D. Varga, Z. Szoplák, S. Krajči, P. Sokol, and P. Gurský, "Analysis and Prediction of Legal Judgements in the Slovak Criminal Proceedings," 2021.

[3] D. Varga, M. Gojdič, Z. Szoplák, P. Gurský, Š. Horvát, S. Krajči and Ľ. Antoni " Extraction of Legal References from Court Decisions.," 2023

[4] C. Jiang and X. Yang, "Legal Syllogism Prompting: Teaching Large Language Models for Legal Judgment Prediction," in *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023)*, pp. 1–5, 2023. doi: 10.1145/3594536.3595170.

[5] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras and I. Androutsopoulos, LEGAL-BERT: The Muppets straight out of Law School. 2898-2904., 2020. doi: 10.18653/v1/2020.findings-emnlp.261.

[6] Odysseas S. Chlapanis, Dimitrios Galanis, and Ion Androutsopoulos. 2024. LAR-ECHR: A New Legal Argument Reasoning Task and Dataset for Cases of the European Court of Human Rights. In Proceedings of the Natural Legal Language Processing Workshop 2024, pages 267–279, Miami, FL, USA. Association for Computational Linguistics.

[7] Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, Jianfeng Xu. CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. arXiv preprint arXiv:1807.02478, 2018.

[8] H. Dai, W. Zhao, and L. Li, "Enhancing Legal Judgment Prediction in LLMs via Legal Norms Integration," in *Knowledge Science, Engineering and Management (KSEM 2025)*, pp. 202–214, 2025. doi: 10.1007/978-981-95-3055-7_16.

[9] R. Shui, Y. Cao, X. Wang, and T.-S. Chua, "A Comprehensive Evaluation of Large Language Models on Legal Judgment Prediction," pp. 7337–7348, 2023. doi: 10.18653/v1/2023.findings-emnlp.490.

[10] Y. Liu, "Efficient Prediction of Judicial Case Decisions Based on State Space Modeling," *International Journal of Computational Intelligence Systems*, vol. 17, p. 280, 2024. doi: 10.1007/s44196-024-00695-2.

[11] Meta AI, *Llama 3.3 Model Card and Documentation*, https://www.llama.com/docs/how-to-guides/prompting/
[12] Meta AI, *Llama Prompting Guide*, https://www.llama.com/docs/how-to-guides/prompting/