GRAF: Graph Retrieval Augmented by Facts for Romanian Legal Multi-Choice Question Answering

Cristian-George Craciun[1,2], Răzvan-Alexandru Smădu[1], Dumitru-Clementin Cercel[1]*, Mihaela-Claudia Cercel[3,4]

[1]National University of Science and Technology POLITEHNICA Bucharest, Faculty of Automatic Control and Computers, Bucharest, Romania
[2]Technical University of Munich, Munich, Germany
[3]Paris 1 Panthéon-Sorbonne University, Paris, France
[4]University of Bucharest, Bucharest, Romania

cristian.craciun@tum.de, razvan.smadu@stud.acs.upb.ro, dumitru.cercel@upb.ro

Abstract

Pre-trained language models have shown remarkable performance in recent years, setting a new paradigm for natural language processing (NLP) research. The legal domain has received some attention from the NLP community, in part due to its textual nature. Question answering (QA) systems represent some of the tasks in this domain. This work explores the legal multiple-choice QA (MCQA) for Romanian. The contribution of this work is multifold. We introduce **JuRO**, the first openly available Romanian legal MCQA dataset, comprising 10,836 questions from three examinations. Along with this dataset, we introduce **CROL**, an organized corpus of laws comprising a total of 93 distinct documents with their modifications over 763 time spans, which we used for information retrieval techniques in this work. Additionally, we construct **Law-RoG**, the first graph of legal knowledge for the Romanian language, derived from the aforementioned corpus. Lastly, we propose a novel approach for MCQA, namely Graph Retrieval Augmented by Facts (**GRAF**), which achieves competitive results with generally accepted state-of-the-art methods and even exceeds them in most settings.

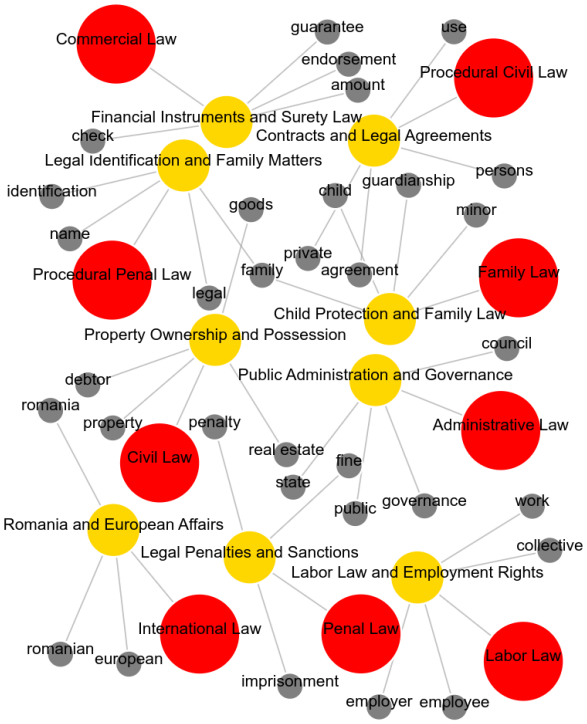

Figure 1: General topic graph for **CROL**. Red entities encompass the main legal branches, yellow entities represent topics, and grey entities are associated keywords.

1 Introduction

Question answering (QA) represents a family of downstream natural language processing (NLP) tasks explored in various settings (Zhong et al., 2020; Rogers et al., 2023; Singhal et al., 2023). Some QA tasks can be formulated as open-ended questions that require an elaborate answer, which may include the rationale behind it (Chen et al., 2017; Labrak et al., 2022). Other tasks focus on retrieving the answer from a given context. In this scenario, the answer can either be explicitly found in the provided context, in which case a model has to identify the location at which the answer occurs (Zaib et al., 2021), or infer the correct answer based on understanding and interpretation of the context at hand, a task known as machine reading comprehension (Baradaran et al., 2022).

The tasks mentioned above have been explored for various domains and languages (Hoppe et al., 2021; Louis and Spanakis, 2022; Askari et al., 2022; Sen et al., 2022; Ekram et al., 2022). Our work focuses on multiple-choice question answering (MCQA) for the Romanian legal domain. Legal

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 12708–12742

QA represents an emerging area of research due to the insights it can provide in addressing various other problems, such as the dialogue system (He et al., 2024). From a practical perspective, the average citizen of a nation can benefit from the opportunity to find answers to legal inquiries alongside experts in the field, which can lead to increased productivity.

Additionally, natural language corpora for the Romanian language are scarce. A dataset for single-choice QA was recently performed by Dima et al. (2024), which proposed a Romanian medical MCQA dataset. Although the Romanian legal domain has been tackled in the past (Masala et al., 2021, 2024), no open-source datasets are available. This motivated us to introduce new comprehensive resources for the Romanian language.

In summary, the contributions of this paper are multifold:

- We release **JuRO**[1], the first open-source legal dataset available for the Romanian MCQA.
- We release **CROL**[1], a structured corpus of Romanian law that can be utilized to query the law for answers (see Figure 1).
- We release **Law-RoG**[1], the first knowledge graph for the Romanian legal domain.
- We present a novel algorithm for the MCQA task called **GRAF**[1].
- We provide a comprehensive evaluation of our proposed MCQA dataset using state-of-the-art methods and our proposed method. We also released the code to allow future research to build on top of this work[1].

2 Related Work

2.1 Legal-Domain Question Answering

English QA. Initially, traditional legal QA systems were based on information retrieval techniques, rule-based systems, support vector machines (Kim et al., 2014, 2017), and shallow convolutional networks (CNNs) (Kim et al., 2015b) to answer legal questions from bar exams (Kim et al., 2015a). Later, deep neural networks based on CNNs (Do et al., 2017) and BiLSTM (John et al., 2017) were used with improved performance. The remarkable achievements of the Transformer models (Vaswani et al., 2017) motivated their use in this domain. Shankar et al. (2023) focused on privacy-related law, using various sources to build their benchmark dataset and evaluating several legal Bidirectional Encoder Representations from Transformers (BERT) models (Devlin et al., 2019). Similarly, Hendrycks et al. (2021b) evaluated multiple BERT-based architectures on a newly proposed dataset, demonstrating that both the volume of training data and the architecture influence performance. To support these findings, new datasets were proposed (Ravichander et al., 2019; Ahmad et al., 2020; Sovrano et al., 2021), including large multidisciplinary datasets (Hendrycks et al., 2021a; Chalkidis et al., 2022; Guha et al., 2023).

Non-English QA. Other languages also took the initiative in the early days. For example, the COLIEE shared task (Rabelo et al., 2022a), proposed in 2014, was the first legal QA task on Japanese legal documents provided in both Japanese and English. Bach et al. (2017) proposed an analysis of the Vietnamese transportation laws and evaluated a CRF-based system as a prerequisite stage for QA. An encoder-decoder architecture was proposed by Kien et al. (2020), which contained word embeddings, convolutional, and attention layers. The model was evaluated on a dataset containing 6,000 Vietnamese legal questions, outperforming existing retrieval-based methods. Later, Vuong et al. (2023) proposed an end-to-end retrieval-based system that used a pre-trained BERT model on weakly labeled data. Zhong et al. (2020) addressed the Chinese legal field, featuring an MCQA dataset containing practice exercise questions and a knowledge database. The comprehensive evaluation of various methods, including transformers, attention, and distant supervision, revealed a significant gap until human-level performance was achieved. Other works addressed languages such as Arabic (Hijazi et al., 2024), Chinese (Chen et al., 2023a; Jiang et al., 2024b), French (Louis and Spanakis, 2022; Louis et al., 2024), German (Büttner and Habernal, 2024), and Spanish (Calleja et al., 2021).

2.2 Romanian Legal Domain

Recent advances in the Romanian legal domain have focused on developing specialized models, datasets, and tools. Păiș et al. (2021) developed the LegalNERo dataset, a word embedding model, and a BiLSTM-CFR model for the legal named entity recognition (NER) task. Smădu et al. (2022) proposed a multi-task domain adaptation model to

[1] https://github.com/craciuncg/GRAF

address legal NER. Masala et al. (2021) proposed jurBERT, a BERT model adapted to the Romanian legal domain, trained on an extensive corpus of 160GB of legal text, achieving improved performance compared to RoBERT (Masala et al., 2020). Following this result, the authors attempted to build a judicial prediction system using this model (Masala et al., 2024). Other works focus on text classification (Avram et al., 2021) and anonymization of Romanian jurisprudence (Păiş et al., 2024).

2.3 Knowledge Graph-Based Question Answering

The literature presents various approaches to knowledge graph-based QA. We identify two main classes of approaches: symbolic and numeric. Since a knowledge graph (KG) is a symbolic, structured representation of factual knowledge, some works revolved around symbolic methods. Chakraborty (2024) explored multi-hop QA approaches employing large language models (LLMs) and a KG-based algorithm, which identified the correct answer. The numeric approaches attempt to employ learned numeric representations of the data to make predictions. One such work is performed by He et al. (2022), which proposes a medical dialogue dataset as well as a method for utilizing a medical graph (Wang et al., 2019) to predict the corresponding medication based on patient-doctor dialogue. Our work aims to create synergy between the two approaches and proposes new datasets and a knowledge graph for future use.

3 Novel Resources

3.1 JuRO

We introduce a new dataset for legal QA, called JuRO. It contains past examination questions from all the legal branches examined in Romania. It represents the first dataset of its kind that we release to the public (see also Table 1).

Dataset Construction. The data is extracted from various official examination portals. Some of the subjects were extracted using OCR. To avoid errors, we manually inspected all the data, and thus, the resulting samples present minimal damage. Each entry essentially consists of a body in which a theoretical question is posed regarding a legal aspect, along with three possible answer choices labeled A, B, and C, out of which at most two answers are correct.

Statistics. The dataset contains questions from three types of examinations: entrance into the judicial system (i.e., entrance), entrance into the bar (i.e., bar), and promotion exams for judicial positions (i.e., promotion). We present the distribution among legal domains and possible answer combinations in Figure 6 of the Appendix A. Promotion exams have three possible choices with a single correct answer, whereas the others have up to two possible correct answers. The distribution among correct answers is generally balanced, with a small exception for bar exams. However, it should be noted that the questions with a single correct answer are predominant and balanced. Data analysis for the JuRO dataset is presented in Appendices A and B.

JuRO vs Existing Work. We introduce a new dataset for the Romanian legal QA, which encompasses three different types of examinations: entrance, bar, and promotion. We are the first to propose a legal MCQA dataset for the Romanian language and to make it publicly available. We hope that this will open opportunities for future research in Romanian, multilingual, and low-resource language settings. In Table 1, we compare our work with other legal datasets. Although it is less than half the size of the JEC-QA dataset (Zhong et al., 2020), it is larger than other existing datasets for legal QA in other languages.

3.2 CROL

Corpus for **Ro**manian **L**aw (CROL) represents a collection of legal documents collected for law branches as follows: *civil*, *penal*, *work*, *administration*, *commercial*, *family*, and *international*.

Dataset Construction. The CROL corpus was constructed using the official Ministry of Justice department portal[2] to crawl all laws from the covered branches in the JuRO dataset. All these resources have been extracted from official sources of national state institutions. Therefore, the language is formal and presents few or no grammatical errors.

Statistics. CROL represents a collection of organized legal documents, amounting to 93 distinct laws and 768 different versions of these

[2] https://legislatie.just.ro/

Dataset	# Examples	QA Format	Language	Public
PrivacyQA (Ravichander et al., 2019)	1,750	Span	English	✓
PolicyQA (Ahmad et al., 2020)	714	Span	English	✓
JEC-QA (Zhong et al., 2020)	26,367	Multi-Choice	Chinese	✓
COLIEE-21 (Rabelo et al., 2022b)	887	Binary	Japanese	✓
BSARD (Louis and Spanakis, 2022)	1,100	Article Retrieval	French	✓
EQUALS (Chen et al., 2023a)	6,914	Long Form	Chinese	✓
LLeQA (Louis et al., 2024)	1,868	Long Form	French	✓
LegalCQA (Jiang et al., 2024b)	21,780 8,899	Long Form	Chinese English	✓
JuRO (Ours)	10,836	Multi-Choice	Romanian	✓

Table 1: Comparison of JuRO with other existing datasets.

Corpus	Size	Dataset Type	Language	Public
Marcell (Váradi et al., 2020)	317k documents, 774M tokens	Pre-training	Multilingual*	✓
RoJur (Masala et al., 2021)	11M entries, 160GB	Cases	Romanian	✗
Ostendorff et al. (2020)	200k court rulings	Rulings	German	✓
Collarana et al. (2018)	62 pages, 64 sections, 24k words	Span	English	✗
Kien et al. (2020)	8.5k documents, 118k articles	Article	Vietnamese	✗
EQUALS (Chen et al., 2023a)	3,081 articles	Article	Chinese	✓
CROL (Ours)	330k articles, 31.5M words	Article	Romanian	✓

* Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak, Slovenian

Table 2: Comparison of CROL with other datasets from various works.

	Count	Avg. Length	Max Length
Articles	330,320	95.11	24,735
Words	31,416,577	6.16	28
Vocabulary	78,355	9.59	28
Nodes	160,402	-	-
Edges	319,958	-	-

Table 3: General statistics for CROL and Law-RoG KG.

laws. It contains 330k articles totaling about 31.5M words with a vocabulary of 78.3k words. Statistics are also presented in Table 3 and Appendix A and B. See also Figure 1 for a graphical view of the topics and keywords in the corpus.

CROL vs Existing Work. Our corpus can mainly serve as a knowledge base for information retrieval (IR) techniques for Romanian legal tasks. There has been past work on creating a Romanian legal corpus, such as the Marcell project (Váradi et al., 2020), which aimed to develop a multilingual legal corpus that includes Romanian law. However, this represents an annotated text corpus useful for NER training in a legal context, as well as related, but it does not make a distinction between documents that are in effect and those that are not. We have made a clear separation between legal documents and their updates. There are also other efforts (Collarana et al., 2018; Kien et al., 2020; Masala et al., 2021); however, they are not publicly available (see Table 2).

3.3 Law-RoG

We introduce the first knowledge graph for the Romanian language, called Law-RoG. This KG is built on the CROL corpus via entity-relation extraction. In particular, following the work of Edge et al. (2024), we prompt an LLM to identify named entities and the relations between them to output entity-relation-entity triplets in our desired format using in-context learning (Brown et al., 2020) abilities that LLMs exhibit via few shot prompting (see Appendices I and J). We opted for this approach because the Romanian NLP lacks resources for building specialized pipelines for entity and relation extraction, particularly in the legal domain. To validate that the LLM produced factually and coherently correct information, we asked 5 human NLP experts to evaluate 10 randomly sampled documents and their corresponding generated relations for each legal domain. They all agreed that the outputs were coherent and did not hallucinate beyond the given document. Although not perfect, we concluded that the generated relations were appropriate for almost all related NLP applications. The resulting KG spans 160k nodes and 320k edges (see Table 3).

4 GRAF

We introduce a novel approach for retrieving information from a KG, which we applied to the legal MCQA. The same principles discussed regarding claim checking and validation can be applied to other tasks requiring factual knowledge.

4.1 Problem Formulation

At first glance, we are presented with questions that may have only one correct answer for a dataset, and in other scenarios, a question can have up to two correct choices. We will make a distinction between these two settings and formulate the goal of the problem according to the architecture of the proposed model.

The multi-choice QA problem can be formulated as follows. Consider a dataset $\mathcal{D} = \{x_i = (Q_i, C_i, T_i)\}, i = 1 : |\mathcal{D}|$, where each triplet entry x_i contains the question body Q_i in textual form, the set of $|C_i|$ possible answer choices $C_i = \{C_i^k\}, k = 1 : |C_i|$ and the set of target answers T_i with $|T_i| \leq |C_i|$ and $T_i \subseteq C_i$. Each answer choice is a tuple $C_i^j = (\sigma_i^j, \epsilon_i^j)$ where σ_i^j is the choice label (e.g., A, B, C) of the answer, and ϵ_i^j represents its textual content. In our setting, we examine questions with $|C_i| = 3$ choice answers and $|T_i| \in \{1, 2\}$ correct answers, i.e., single-choice and multi-choice QA, depending on the dataset. Moreover, we investigate two classes of models designed to address these QA variations and formulate the learning goal in Appendix D.

4.2 Algorithm Description

Our proposed algorithm is applied to every given question and each of its possible answer choices. Specifically, the inputs to GRAF are the question-choice pair and the KG, which contains entities as nodes and relationships between entities as edges. The algorithm is illustrated in Figure 2 and Algorithm 1 and will be presented in what follows.

Claim Graph. A multiple-choice question is primarily composed of choices that present different claims with various truth values. We are interested in determining claims whose entities come from (1) the question and the given choice or (2) both come from the choice. In our setting, questions generally present hypothetical scenarios or premises that do not contradict the law; therefore, we do not consider the question body itself to present any untrue claims. Therefore, to choose the most suitable answer, we decompose each pair (question, choice) into its underlying claims using an LLM, similar to Edge et al. (2024); however, any lighter solutions can be adopted with adequate resources. We call this procedure *Cross Claim Extraction*. An example is presented in Figure 3.

Algorithm 1: GRAF

Data: Q - question, C - choices, G - knowledge graph
Result: P - choice probabilities

$P \leftarrow []$
for $c_i \in split\_choices(C)$ **do**
  // Query the cross-claim extraction model and obtain the claim graph CG
  $CG \leftarrow$ cross_claim_extract$(Q||c_i)$
  // Sample a subgraph SG from G
  $SG \leftarrow$ sample_graph(G)
  // Encode SG's components
  $SGE, CGE \leftarrow \{\}, \{\}$
  for $(E_a, r_{ab}, E_b) \in SG$ **do**
    SGE.append(enc(E_a), enc(r_{ab}), enc(E_b))
  end
  // Encode CG's components
  for $(E_a, r_{ab}, E_b) \in CG$ **do**
    CGE.append(enc(E_a), enc(r_{ab}), enc(E_b))
  end
  // Compute the alignment between encoded claims and all encoded relations
  for $(h_c^i, h^j) \in GAT(SGE) \times GAT(CGE)$ **do**
    $R^{ij} \leftarrow \cos(h_c^i, h^j)$
  end
  // Compute the embedding for all neighboring claims
  $\bar{H} \leftarrow Rh$
  // Compute the final score using self-attention
  $c \leftarrow$ enc$(Q||C)$
  $c_{\text{final}}, H_{\text{final}} \leftarrow$ SelfAttention$(c||\bar{H})$
  // Save the score of the current choice
  P.append(c_{final})
end

KG Sampling. We obtain the domain-specific KG (in our setting, from Law-RoG) by using the law branch related to the question. Since it is often infeasible to consider the entire graph for inference due to its size, we resort to sampling the graph via a procedure that retrieves the most relevant nodes and edges using a heuristic. First, we preprocess the words by tokenizing and lemmatizing them using the *SpaCy*[3] package for the Romanian language to achieve better, less noisy results. Then, we use a bag-of-words (BoW) approach and incorporate the BM25 retriever (Robertson

[3] https://spacy.io

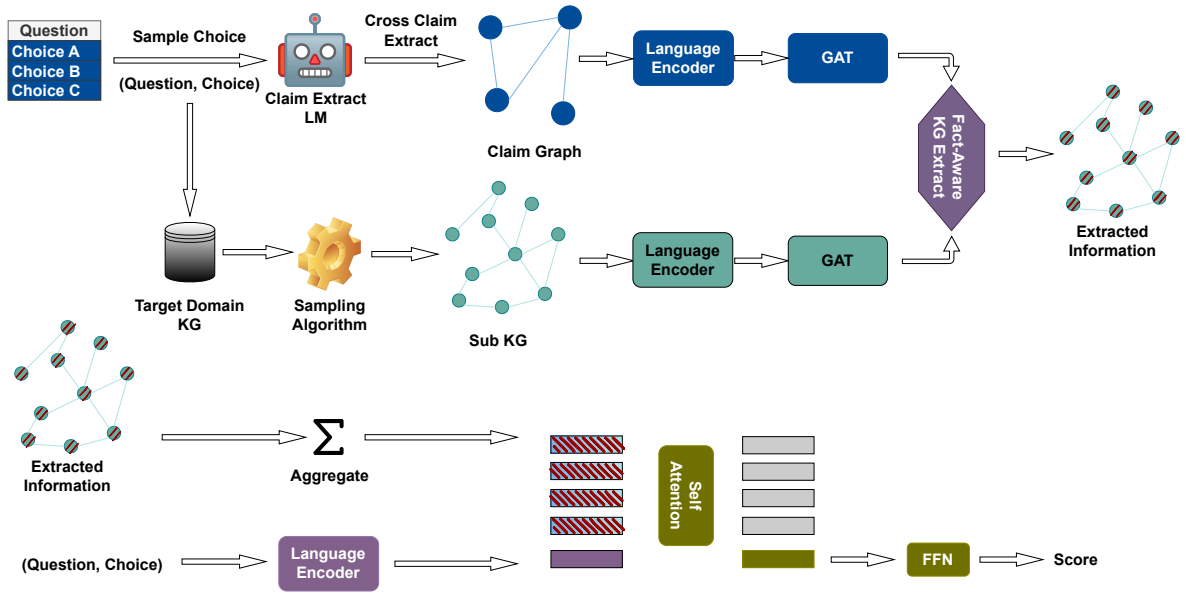

Figure 2: **GRAF** procedure overview.

and Jones, 1976) to select the top k entities in the knowledge graph. We proceed to select their vicinity with a breadth-first search for a limited depth. We also limit the maximum number of entities retrieved during this stage. In our work, we use a depth of 1, select the top 10 entities, nodes, and edges from the KG, and limit the selection to no more than 50 distinct entities.

KG Encoding. To encode the KG, we utilize a language encoder model to embed the entities and relations, resulting in sets of node and edge features $\boldsymbol{h_N}$ and $\boldsymbol{h_E}$, respectively. Then, we adapt the Graph Attention Network (GAT) (Veličković et al., 2018) to further capture the topological information for each entity. The original GAT model was developed only for graphs with no relation encoding. Therefore, we transform the set of features using two different linear transformations, parametrized by shared $\boldsymbol{W_N}$ and $\boldsymbol{W_E}$ for the nodes and edges, respectively. To capture the relational topological information, we compute the attention coefficients for the relations $\boldsymbol{e_E^{ij}}$ in which the current entity is involved between the current node i and the adjacent edges j:

$$\boldsymbol{e_E^{ij}} = \sigma_A((\boldsymbol{a_E})^T[\boldsymbol{W_N h_N^i}||\boldsymbol{W_E h_E^j}]) \quad (1)$$

and calculate the attention coefficients for the nodes e_N^{ij} to capture inter-entity relations between the current node i and neighboring nodes j:

$$\boldsymbol{e_N^{ij}} = \sigma_A((\boldsymbol{a_N})^T[\boldsymbol{W_N h_N^i}||\boldsymbol{W_N h_N^j}]) \quad (2)$$

where $\boldsymbol{a_N}$ and $\boldsymbol{a_E}$ represent two distinct attention vectors for nodes and edges, respectively. We also use a nonlinearity σ_A as Veličković et al. (2018), which in our case is the LeakyReLU activation function. The $\cdot^T$ operator represents transposition, while $||$ is the concatenation operator.

We obtain the final nodes and edges representations by aggregating the information from the adjacent nodes for each node in $\boldsymbol{h_N^{'}}$ and the information from the adjacent edges for each node into $\boldsymbol{h_E^{'}}$:

$$\boldsymbol{h_N^{'}} = \text{softmax}(\boldsymbol{e_N})\boldsymbol{W_N h_N} \quad (3)$$

$$\boldsymbol{h_E^{'}} = \text{softmax}(\boldsymbol{e_E})\boldsymbol{W_E h_E} \quad (4)$$

In the end, we combine this information into a single representation for each node into $\boldsymbol{h^{'}}$ as follows:

$$\boldsymbol{h^{'}} = \boldsymbol{h_N^{'}} + \boldsymbol{h_E^{'}} \quad (5)$$

Final Score. After encoding the graphs, we select the relevant information from the provided knowledge, given the encoded claims. For this, we compute a relevance matrix by calculating the alignment between each encoded claim $\boldsymbol{h_c^i}$ and all the encoded relations $\boldsymbol{h^j}$ from the sampled KG

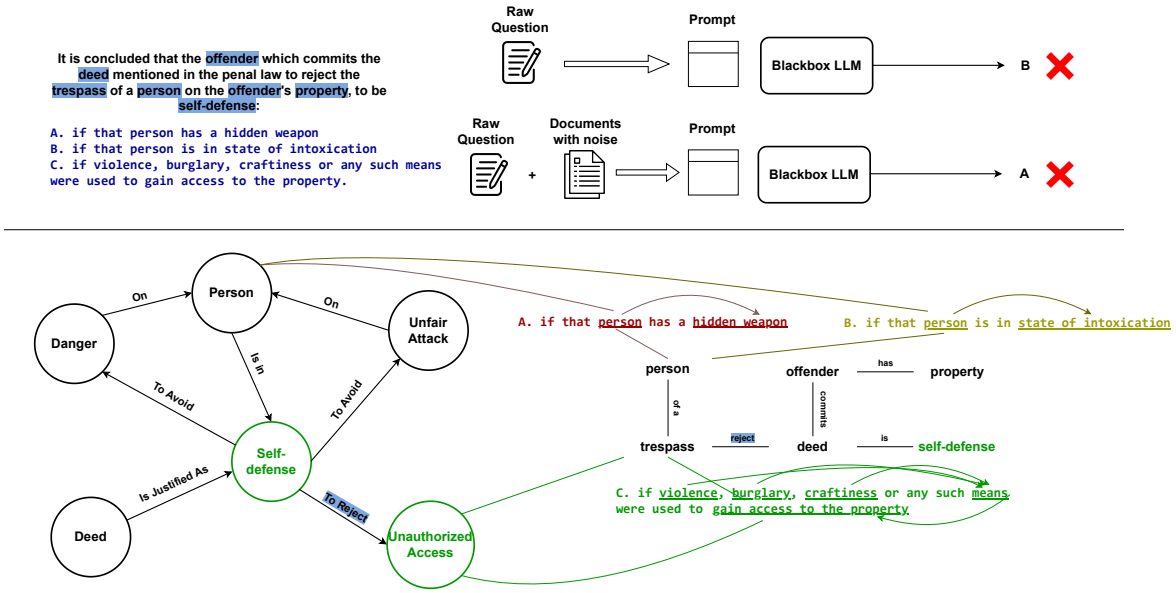

Figure 3: A showcase of how **GRAF** works on a sample question from the **JuRO** dataset translated to English. First, we construct the claim graph using an LLM that extracts the entities (nodes) and relations between them (edges). Based on the sub-KG extracted from Law-RoG and the claim graph, we determine that the entities "Self-defense" and "Unauthorized Access" have the best alignment with the entities in choice "C", thus, it is most likely to be the correct answer.

(i.e., sub-KG). We calculate the alignment using the cosine similarity:

$$\boldsymbol{R}^{ij} = \cos(\boldsymbol{h}_{\boldsymbol{c}}^{\boldsymbol{i}}, \boldsymbol{h}^{\boldsymbol{j}}) \tag{6}$$

We then use this matrix to finally aggregate all the relevant information from the sub-KG into a matrix containing as many vectors as nodes in the original claim graph, each vector being a numeric encoding representation for every encoded claim node, which in turn represents an embedding for all the neighboring claims:

$$\bar{\boldsymbol{H}} = \boldsymbol{R}\boldsymbol{h} \tag{7}$$

We separately encode the (question, choice) pair into $\bar{c}$ and decide what information is better suited for the final decision for the current choice. We employ a self-attention mechanism for this task and provide a score based on the gathered information and the given choice:

$$[\boldsymbol{c}_{\text{final}}||\boldsymbol{H}_{\text{final}}] = \text{SelfAttention}([\bar{\boldsymbol{c}}||\bar{\boldsymbol{H}}]) \tag{8}$$

Finally, we compute the score:

$$\text{score} = \sigma(\boldsymbol{W}_{\text{final}}\boldsymbol{c}_{\text{final}}) \tag{9}$$

where σ is the sigmoid logistic activation used to provide a probability score, and W_{final} is a learnable parameter.

5 Experiments and Results

In this section, we present the results of our extensive experimentation and discuss the findings obtained.

5.1 Baselines

For encoder models, we adopt approaches similar to those in information retrieval (IR) and retrieval augmented generation (RAG) (Wang et al., 2024). In Appendix E, more experimental details are discussed. As baselines, we employ QBERT, ColQBERT (Manotumruksa et al., 2020), ColBERT (Khattab and Zaharia, 2020), Large Language Models (LLMs) such as FLAN-T5 (Raffel et al., 2020), Mistral (Jiang et al., 2023) and Llama 3.1 8B (Dubey et al., 2024), LLM with RAG (Lewis et al., 2020), and LLM fine-tuned with Low-Rank Adaptation method (LoRA) (Hu et al., 2021). More details regarding these baselines can be found in Appendix F. For information on the language models employed, see Appendix G.

5.2 Evaluation Metric

We evaluated the models using the score that a model would receive on the given test, equivalent to the model's accuracy on the task. No extra points are given or deducted if the model mispredicts correct answers or fails to include all correct answers.

Model	Civil	Penal	Civil Proc.	Penal Proc.	Administrative	Work	Family	International	Commercial	Average
QBERT	35.48	38.29	35.82	40.10	36.63	40.40	38.24	39.29	39.39	38.18
CrossQBERT	41.94	36.04	37.31	41.15	41.58	35.35	34.31	42.86	36.36	38.54
ColBERT	44.09	38.29	47.76	37.50	41.58	42.42	41.18	48.81	40.40	42.45
LLM	48.94	40.18	41.00	42.27	46.53	48.48	49.02	52.38	39.39	45.35
LLM RAG	**53.19**	43.75	38.81	42.78	57.43	56.57	63.73	52.38	**57.58**	51.81
LLM LFT	45.74	51.79	74.63	48.97	52.48	49.49	54.90	**70.24**	53.53	55.75
GRAF(Ours)	49.46	**52.70**	**78.46**	**51.05**	**59.18**	**57.29**	**68.69**	67.12	56.84	**60.09**

Table 4: Accuracy results on promotion exams.

We use this metric to emphasize the actual test performances of the models. Moreover, we argue that the dataset is balanced and suitable for comparative analysis, and thus, we consider this metric sufficient to avoid overwhelming the results section with excessive numbers.

5.3 Analysis

Our analysis evaluates the performance of our proposed model through quantitative and qualitative evaluations against baseline approaches.

From a quantitative **perspective**, we systematically compare the performance of the model in different legal examinations. All models are evaluated directly for the promotion exams with a single correct answer, as shown in Table 4. However, encoder-based models and LLMs are evaluated separately for exams with multiple correct answers to ensure a fair comparison (§4.1). Our model outperforms baselines in 6 out of 9 legal branches, despite slight inconsistencies due to training on the entire dataset. Tables 5 and 6 present detailed performance breakdowns for encoder-based and LLM-based models across different examination types, with additional granular results available in Appendix H.

From a **qualitative perspective**, we provide a comparative study for the promotion exams to improve our approach over the baselines shown in Figure 4. We analyze the improvements introduced by our model, particularly in terms of backbone model scaling and domain-specific fine-tuning. As illustrated in Figure 5, pre-trained models for the legal domain significantly outperform their general-purpose counterparts. Our model surpasses baseline encoder models in all evaluation settings, while evidence suggests that poorer performance can be solved by increasing the size of the backbone model. Our framework learns to extract relevant information while answering

Model	Civil	Penal	Civil Proc.	Penal Proc.	Average
QBERT	41.18	49.10	37.25	39.22	41.69
CrossQBERT	41.18	49.02	43.14	37.25	42.65
ColBERT	45.10	43.14	50.98	41.18	45.10
LLM	39.21	31.37	15.69	27.45	28.43
LLM RAG	43.14	27.45	21.57	27.45	29.90
LLM LFT	37.25	45.10	41.18	41.18	41.18
GRAF(Ours)	**60.78**	**62.75**	**54.90**	**56.86**	**58.82**

Table 5: Accuracy results on entrance exams.

Model	Civil	Penal	Civil Proc.	Penal Proc.	Average
QBERT	40.32	32.26	40.32	36.29	37.30
CrossQBERT	41.94	38.52	39.52	34.68	38.67
ColBERT	38.71	33.87	36.29	39.52	37.10
LLM	21.77	25.00	23.39	18.54	22.18
LLM RAG	26.61	29.03	33.87	20.97	27.62
LLM LFT	31.45	32.26	33.06	36.29	33.27
GRAF(Ours)	**51.61**	**61.29**	**56.10**	**59.02**	**57.01**

Table 6: Accuracy results on bar exams.

questions by combining retrieval, fine-tuning, KGs, and inter-entity relationships within legal texts. This structured approach improves performance, allowing fine-tuning while maintaining competitive results in areas where RAG excels. Conversely, though effective in specific legal branches, fine-tuned LLMs struggle to generalize across domains and are susceptible to hallucinations when provided with external context.

We also measure the agreement among LLMs on various categories and topics. We assess the average pairwise percentage agreement (APPA) by computing the percentage of samples in every pair of LLMs responses that produced the same result and then averaging the scores. Table 7 presents the APPA for every exam type and category. The values range between 40% and 50%, meaning low agreement. However, the Fleiss' κ is slightly negative, resulting in no agreement. Therefore, there are questions on which LLMs perform poorly.

Additionally, we provide an in-depth analysis of LLM performance based on question difficulty in Appendix C.

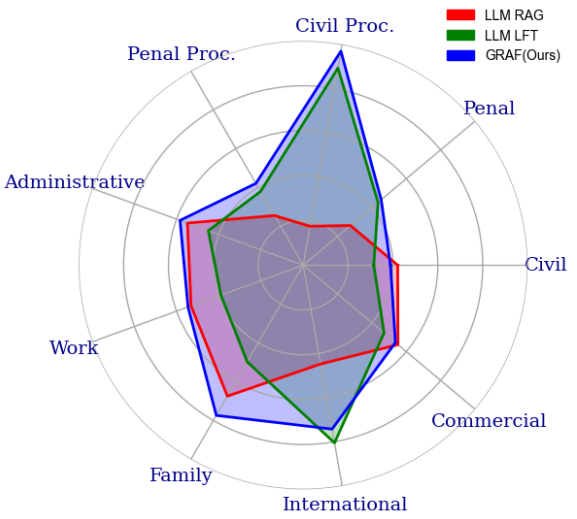

Figure 4: Comparative performances of best baselines and our approach on different law branches.

Exam Type	Category	APPA (%)	κ
Entrance	Civil	47.65	-0.0200
	Penal	44.51	-0.0200
	Civil Procedure	47.25	-0.0200
	Penal Procedure	45.88	-0.0200
Bar	Civil	45.00	-0.0081
	Penal	48.55	-0.0081
	Civil Procedure	49.11	-0.0081
	Penal Procedure	47.50	-0.0081
Promotion	Administrative	40.99	-0.0100
	Civil	42.34	-0.0108
	Commercial	41.92	-0.0102
	International	43.10	-0.0120
	Penal	40.18	-0.0045
	Civil Procedure	40.25	-0.0017
	Penal Procedure	43.30	-0.0052
	Family	41.35	-0.0100
	Work	41.01	-0.0102

Table 7: Summary of average pairwise percentage agreement (APPA) and Fleiss' κ scores for LLM models.

5.4 Ablation Study

In Table 8, we present the effect of removing different parts of our algorithm in an ablation study conducted on the promotion exams. We remove the claim graph and the KG, and we experiment with collapsing the KG embeddings via summation (denoted *emb. sum*). In this sense, we experiment with various ablated components and found that missing KG or claim graphs underperform the LLM baselines. Even when the model uses only the KG, its performance is poorer because irrelevant information may serve as a distractor in the decision-making process. We show that our claim-aware information extraction mechanism is

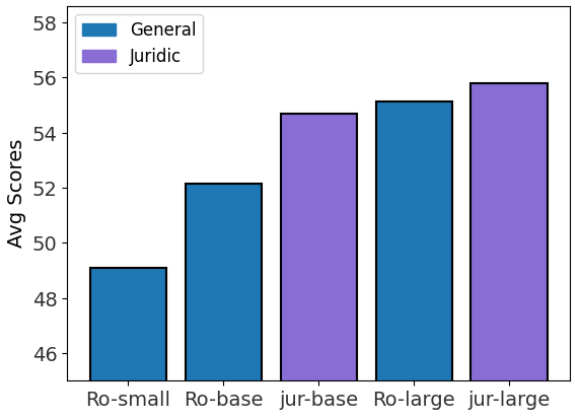

Figure 5: Comparative performances of different backbone encoder models (i.e., BERT-based models) of different sizes and pre-trained on different domains.

Model	Avg. Accuracy
GRAF	55.61
w/o claim graph	$53.14_{\downarrow 2.47}$
w/o KG	$53.61_{\downarrow 2.00}$
w/o claim graph + emb. sum	$49.51_{\downarrow 6.10}$

Table 8: Results for different ablated components of our framework on the Promotion exams.

effective in enhancing retrieval capabilities and, consequently, overall performance. Merely collapsing the extracted information via summation would still not compensate for the missing claim-aware IR, but would significantly damage the model's performance.

6 Conclusions

In this paper, we have introduced new resources for the Romanian legal domain, being the first to construct and release a Romanian legal KG. We also proposed a novel approach for KG IR, namely GRAF, which surpassed the baselines proposed for MCQA.

Future Work. We expect solutions that will improve upon our proposed method in every possible aspect. Additionally, there may be solutions that could potentially explore dataset augmentation using LLMs. Studies could be conducted on target domain IR, which may include multiple languages, and JuRO, CROL, and even the KG Law-RoG could represent good foundational resources. We hope that our work will motivate further exploration of underrepresented languages and, in turn, inspire the development of solutions that work in low-resource settings.

Limitations

We have released a legal MCQA dataset by gathering questions from all available law examinations nationwide, providing sufficient samples for training. However, it may not be enough for training in a single law branch, which is why we opted for training on the entire dataset.

The goal of our work is to enhance resources and develop a methodological approach to answering legal questions. Since such systems are meant to help users understand the law, they are not yet entirely accurate. The best average score achieved by our approach is only 60%. Therefore, further research is required in this direction, as the legal domain is a sensitive topic when considering the application of machine learning systems in the assessment of laws. We believe that deploying such systems would require human validation by legal experts to minimize the risk of providing unlawful responses.

Ethical Considerations

We have collected our dataset from various official public portals. To protect this dataset from improper use, we have decided to license its use solely for research purposes. It should not be used in commercial settings under any circumstances. Our work was performed in a manner that did not rely on external human crowd-workers and did not raise any ethical concerns. The data do not contain sensitive personal information that could identify any real person. Anonymized abbreviations are used in all of the hypothetical presented scenarios rather than any person's name. Since the data was collected from the public domain and made available by applicable law by the administrative institutions in question, we release these resources under the CC BY-NC-SA 4.0 license[4], allowed by the current European regulations[5].

Acknowledgements

This work was supported by the National University of Science and Technology POLITEHNICA Bucharest through the PubArt program.

[4] https://creativecommons.org/licenses/by-nc-sa/4.0/

[5] https://eur-lex.europa.eu/eli/dir/2019/790/oj

References

Wasi Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. PolicyQA: A reading comprehension dataset for privacy policies. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 743–749, Online. Association for Computational Linguistics.

Arian Askari, Suzan Verberne, and Gabriella Pasi. 2022. Expert finding in legal community question answering. In *European Conference on Information Retrieval*, pages 22–30. Springer.

Andrei-Marius Avram, Vasile Păiș, and Dan Ioan Tufis. 2021. Pyeurovoc: A tool for multilingual legal document classification with eurovoc descriptors. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 92–101.

Ngo Xuan Bach, Le Thi Ngoc Cham, Tran Ha Ngoc Thien, and Tu Minh Phuong. 2017. Question analysis for vietnamese legal question answering. In *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, pages 154–159.

Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. 2022. A survey on machine reading comprehension systems. *Natural Language Engineering*, 28(6):683–732.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Marius Büttner and Ivan Habernal. 2024. Answering legal questions from laymen in German civil law system. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2015–2027, St. Julian's, Malta. Association for Computational Linguistics.

Pablo Calleja, Patricia Martín Chozas, Elena Montiel-Ponsoda, Víctor Rodríguez-Doncel, Elsa Gómez, and Pascual Boil. 2021. Bilingual dataset for information retrieval and question answering over the spanish workers statute. In *XIX Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA)*.

Abir Chakraborty. 2024. Multi-hop question answering over knowledge graphs using large language models. *arXiv preprint arXiv:2404.19234*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English.

In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Andong Chen, Feng Yao, Xinyan Zhao, Yating Zhang, Changlong Sun, Yun Liu, and Weixing Shen. 2023a. Equals: A real-world dataset for legal question answering via reading chinese laws. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 71–80.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023b. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *arXiv preprint arXiv:2310.14735*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Diego Collarana, Timm Heuss, Jens Lehmann, Ioanna Lytra, Gaurav Maheshwari, Rostislav Nedelchev, Thorsten Schmidt, and Priyansh Trivedi. 2018. A question answering system on regulatory documents. In *Legal knowledge and information systems*, pages 41–50. IOS Press.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George-Andrei Dima, Andrei-Marius Avram, Cristian-George Craciun, and Dumitru-Clementin Cercel. 2024. RoQLlama: A lightweight Romanian adapted language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4531–4541, Miami, Florida, USA. Association for Computational Linguistics.

Phong-Khac Do, Huy-Tien Nguyen, Chien-Xuan Tran, Minh-Tien Nguyen, and Minh-Le Nguyen. 2017. Legal question answering using ranking svm and deep convolutional neural network. *Preprint*, arXiv:1703.05320.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Syed Mohammed Sartaj Ekram, Adham Arik Rahman, Md Sajid Altaf, Mohammed Saidul Islam, Mehrab Mustafy Rahman, Md Mezbaur Rahman, Md Azam Hossain, and Abu Raihan Mostofa Kamal. 2022. Banglarqa: A benchmark dataset for under-resourced bangla language reading comprehension-based question answering with diverse question-answer types. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2518–2532.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 44123–44279. Curran Associates, Inc.

Zhenfeng He, Yuqiang Han, Zhenqiu Ouyang, Wei Gao, Hongxu Chen, Guandong Xu, and Jian Wu. 2022. DialMed: A dataset for dialogue-based medication recommendation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 721–733, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Kang Liu, and Jun Zhao. 2024. Agentscourt: Building judicial decision-making agents with court debate simulation and legal knowledge augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9399–9416.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021b. CUAD: an expert-annotated NLP dataset for legal contract review. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Faris Hijazi, Somayah Alharbi, Abdulaziz AlHussein, Harethah Shairah, Reem Alzahrani, Hebah Alshamlan, George Turkiyyah, and Omar Knio. 2024. ArabLegalEval: A multitask benchmark for assessing Arabic legal knowledge in large language models. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 225–249, Bangkok, Thailand. Association for Computational Linguistics.

Geoffrey Hinton. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Christoph Hoppe, David Pelkmann, Nico Migenda, Daniel Hötte, and Wolfram Schenck. 2021. Towards intelligent legal advisors for document retrieval and question-answering in german legal documents. In *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 29–32. IEEE.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024a. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Yue Jiang, Ziyu Guan, Jie Zhao, Wei Zhao, and Jiaqi Yang. 2024b. H-legalki: A hierarchical legal knowledge integration framework for legal community question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14614–14625.

Adebayo Kolawole John, Luigi Di Caro, and Guido Boella. 2017. Solving bar exam questions with deep neural networks. In *Proceedings of the Second Workshop on Automated Semantic Analysis of Information in Legal Texts: co-located with the 16th International Conference on Artificial Intelligence and Law*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Phi Manh Kien, Ha-Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen, and Tu Minh Phuong. 2020. Answering legal questions by learning neural attentive text representation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 988–998, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mi-Young Kim, Randy Goebel, and S Ken. 2015a. Coliee-2015: evaluation of legal question answering. In *Ninth International Workshop on Juris-informatics (JURISIN 2015)*.

Mi-Young Kim, Ying Xu, and Randy Goebel. 2015b. A convolutional neural network in legal question answering. In *JURISIN Workshop*.

Mi-Young Kim, Ying Xu, Randy Goebel, and Ken Satoh. 2014. Answering yes/no questions in legal bar exams. In *New Frontiers in Artificial Intelligence*, pages 199–213, Cham. Springer International Publishing.

Mi-Young Kim, Ying Xu, Yao Lu, and Randy Goebel. 2017. Question answering of bar exams by paraphrasing and legal text analysis. In *New Frontiers in Artificial Intelligence*, pages 299–313, Cham. Springer International Publishing.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Béatrice Daille, Pierre-Antoine Gourraud, Emmanuel Morin, and Mickaël Rouvier. 2022. Frenchmedmcqa: A french multiple-choice question answering dataset for medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 41–46.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Antoine Louis and Gerasimos Spanakis. 2022. A statutory article retrieval dataset in French. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6789–6803, Dublin, Ireland. Association for Computational Linguistics.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22266–22275.

Jarana Manotumruksa, Jeff Dalton, Edgar Meij, and Emine Yilmaz. 2020. Crossbert: a triplet neural architecture for ranking entity properties. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2049–2052.

Mihai Masala, Radu Cristian Alexandru Iacob, Ana Sabina Uban, Marina Cidota, Horia Velicu, Traian Rebedea, and Marius Popescu. 2021. jurbert: A romanian bert model for legal judgement prediction. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 86–94.

Mihai Masala, Traian Rebedea, and Horia Velicu. 2024. Improving legal judgement prediction in Romanian with long text encoders. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Underresourced Languages @ LREC-COLING 2024*, pages 126–132, Torino, Italia. ELRA and ICCL.

Mihai Masala, Stefan Ruseti, and Mihai Dascalu. 2020. RoBERT – a Romanian BERT model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6626–6637, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Malte Ostendorff, Till Blume, and Saskia Ostendorff. 2020. Towards an open platform for legal information. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 385–388, New York, NY, USA. Association for Computing Machinery.

Vasile Păiş, Radu Ion, Elena Irimia, Verginica Barbu Mititelu, Valentin Badea, and Dan Tufiș. 2024. System for the anonymization of romanian jurisprudence. *Artificial Intelligence and Law*, pages 1–23.

Vasile Păiș, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named entity recognition in the romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18.

Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. 2022a. Overview and discussion of the competition on legal information extraction/entailment (COLIEE) 2021. *Rev. Socionetwork Strateg.*, 16(1):111–133.

Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. 2022b. Overview and discussion of the competition on legal information extraction/entailment (coliee) 2021. *The Review of Socionetwork Strategies*, 16(1):111–133.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4947–4958, Hong Kong, China. Association for Computational Linguistics.

Stephen E Robertson and K Sparck Jones. 1976. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3):129–146.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *International Conference on Learning Representations (ICLR)*.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619.

Atreya Shankar, Andreas Waldis, Christof Bless, Maria Andueza Rodriguez, and Luca Mazzola. 2023. Privacyglue: A benchmark dataset for general language understanding in privacy policies. *Applied Sciences*, 13(6):3701.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, and 1 others. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Răzvan-Alexandru Smădu, Ion-Robert Dinică, Andrei-Marius Avram, Dumitru-Clementin Cercel, Florin Pop, and Mihaela-Claudia Cercel. 2022. Legal named entity recognition with multi-task domain adaptation. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 305–321.

Francesco Sovrano, Monica Palmirani, Biagio Distefano, Salvatore Sapienza, and Fabio Vitali. 2021. A dataset for evaluating legal question answering on private international law. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 230–234.

Tamás Váradi, Svetla Koeva, Martin Yamalov, Marko Tadić, Bálint Sass, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pęzik, Verginica Barbu Mititelu, Radu Ion, and 1 others. 2020. The marcell legislative corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3761–3768.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations (ICLR)*.

Thi-Hai-Yen Vuong, Ha-Thanh Nguyen, Quang-Huy Nguyen, Le-Minh Nguyen, and Xuan-Hieu Phan. 2023. Improving vietnamese legal question–answering system based on automatic data enrichment. In *JSAI International Symposium on Artificial Intelligence*, pages 49–65. Springer.

Jiajia Wang, Jimmy Xiangji Huang, Xinhui Tu, Junmei Wang, Angela Jennifer Huang, Md Tahmid Rahman Laskar, and Amran Bhuiyan. 2024. Utilizing bert for information retrieval: Survey, applications, resources, and challenges. *ACM Computing Surveys*, 56(7):1–33.

Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 950–958.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*.

Munazza Zaib, Dai Hoang Tran, Subhash Sagar, Adnan Mahmood, Wei E Zhang, and Quan Z Sheng. 2021. Bert-coqac: Bert-based conversational question answering in context. In *Parallel Architectures, Algorithms and Programming: 11th International Symposium, PAAP 2020, Shenzhen, China, December 28–30, 2020, Proceedings 11*, pages 47–57. Springer.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jec-qa: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9701–9708.

A Dataset Analysis

The domain distribution of the JuRO, along with the distribution of the answers, is presented in Figure 6. Because of the examinations' format, at most two answers are correct. However, in the case of promotion exams, only one answer is correct. The domains of the questions are *civil procedure*, *penal procedure*, *penal*, *civil*, *work*, *administration*, *commercial*, *family*, and *international*.

Table 9 presents the TF-IDF scores (Salton et al., 1975) for JuRO dataset, calculated using the following formula:

$$\text{score}(t, C) = \frac{f(t,C)}{|\{w|d \in C, w \in d\}|} \log \frac{|C|}{|\{d|d \in C \text{ and } t \in d\}|} \quad (10)$$

where:

- the current term for which we compute the score is denoted by t;
- C is the corpus of documents, each document containing multiple words;
- $f(t, C)$ is the frequency of the term t relative to the corpus C.

We notice a high score for the word "penal" compared to other words, indicating a possible prevalence of penal-related content in the dataset. Moreover, the terms such as "case", "term", "appeal", "judgement", "court", and "request" indicate procedures.

Word	Translation	TF-IDF Score
cerere	request	0.03507
instanţă	court	0.03352
caz	case	0.02769
lege	law	0.02627
penal	penal	0.02622
hotărâre	decision	0.02614
termen	term	0.02602
apel	appeal	0.02565
sine	self	0.02553
judecată	judgement	0.02486

Table 9: TF-IDF scores for the top ten words in the JuRO dataset.

In Table 10, we report the TF-IDF scores for the CROL corpus. Generally, there are words commonly found in articles, but no word indicates a significant bias towards some specific legal area.

In Figure 9, we show the token distribution for both the FLAN-T5 (Raffel et al., 2020) and Mistral (Jiang et al., 2023) models using their tokenizers.

Word	Translation	TF-IDF Score
articol	article	0.03262
caz	case	0.02550
persoană	person	0.02510
lege	law	0.02471
bun	good	0.02429
alineat	paragraph	0.02314
prevedea	stipulate	0.02260
drept	just/correct/law	0.02213
număr	number	0.01965
public	public	0.01961

Table 10: TF-IDF Scores for the top ten words from the CROL corpus.

The distributions behave approximately the same; the difference is that the Mistral tokenizer tends to use more tokens to represent the text than the FLAN-T5 one. Table 11 shows a detailed distribution of questions for each examination.

B Topic Analysis

To better understand the performance of the LLMs used in our work, we extracted the main topics from the CROL and JuRO datasets and present the performance relative to these. For both datasets, the topic extraction procedure is similar. First, we preprocess the JuRO dataset by merging each question with the set of answer choices. In the case of CROL, we perform a minimal data cleaning procedure to remove frequent words and structures that do not represent topics such as law numbers (e.g., Arabic numerals, Roman numerals, references to paragraphs or other laws like "lit. (a)", months of the year, and dates), separators, and the repealed laws, since they have a very similar formulation and any line shorter than five characters. Then, we employ BERTopic (Grootendorst, 2022), which generates transformer-based embeddings and class-based TF-IDF to create dense clusters of semantically similar documents. We set the language to Romanian to output 100 topics and a fixed random seed for reproducibility. The other parameters were left to their default values. We remove outliers and topics that contain only stopwords from the resulting output.

The extracted topics for the JuRO dataset are presented in Table 15 in the original Romanian language and Table 16 for the topics and keywords translated into English. Similarly, Tables 17 and 18 present the top 30 topics of the CROL corpus in Romanian and English, respectively. Both datasets cover a wide variety of topics in the legal domain,

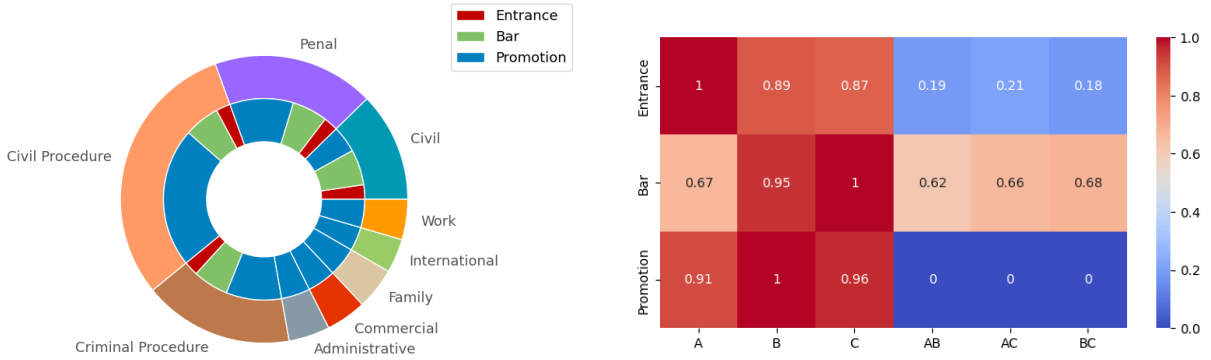

Figure 6: On the left, the number of samples from each law category in the JuRO dataset. On the right, the class equilibrium is depicted via color variations in the heatmap. The heatmap scores are normalized by dividing each value by the maximum of the respective row.

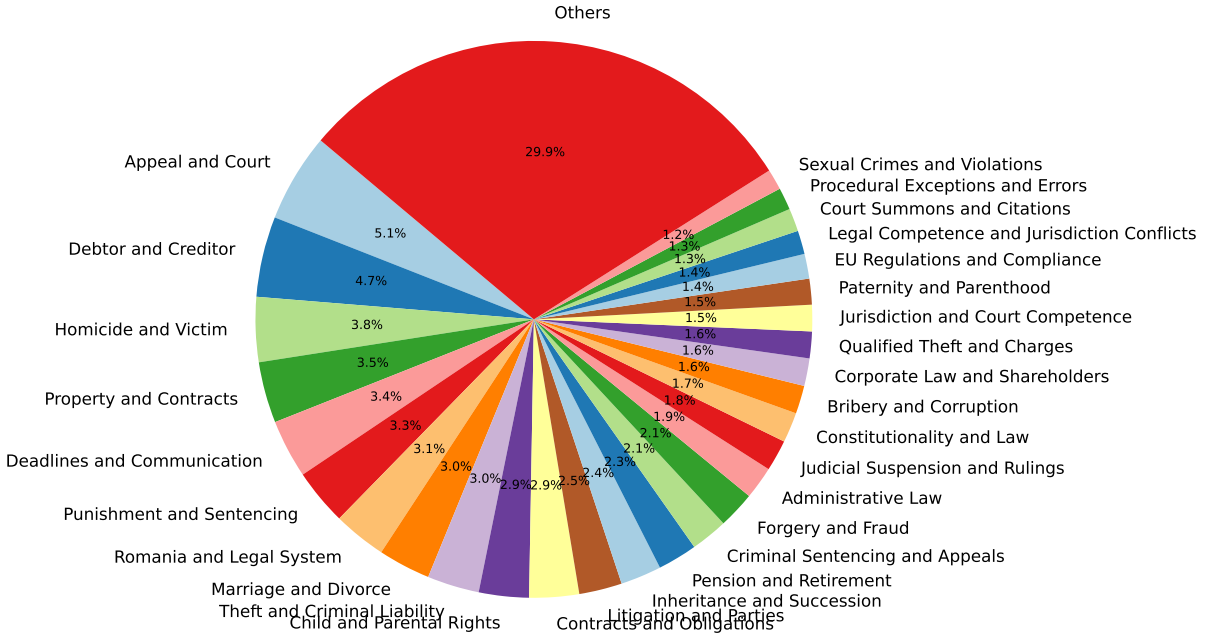

Figure 7: The distribution of top 30 topics in the JuRO dataset.

ranging from *Appeal and Court*, *Punishment and Sentencing*, *Romanian Legal System*, to *Public Administration and Governance*, *Child Protection and Family Law*, *Labor Law and Employment Rights*, and many other legal subjects. We provide the distribution of those main topics in Figures 7 and 8 for JuRO and CROL, respectively. Most of the topics represent a small percentage of the datasets, emphasizing the large diversity of topics addressed in our proposed resources.

C Dataset Difficulty

Inspired by other works (Zheng et al., 2023; Muennighoff et al., 2025), we estimate the question difficulty from the JuRO dataset by analyzing the LLMs' performance (i.e., using the LLMs to judge the difficulty of the questions).

We base our approach on the topics identified in Appendix B. Breaking down the model-level results in Figure 11, we notice that FLAN-T5 XXL performs the best on *procedural exceptions*, outperforming other models by 20-40%. However, there are topics where some models did not answer any question correctly, such as *marriage and divorce*, *bribery and corruption*, and *fraud*.

We also decompose the APPA score for every topic in Figure 10 to identify situations where the models perform poorly. We observe that the models yield better results on questions related to *EU Regulations* and *theft*, while performing poorly on

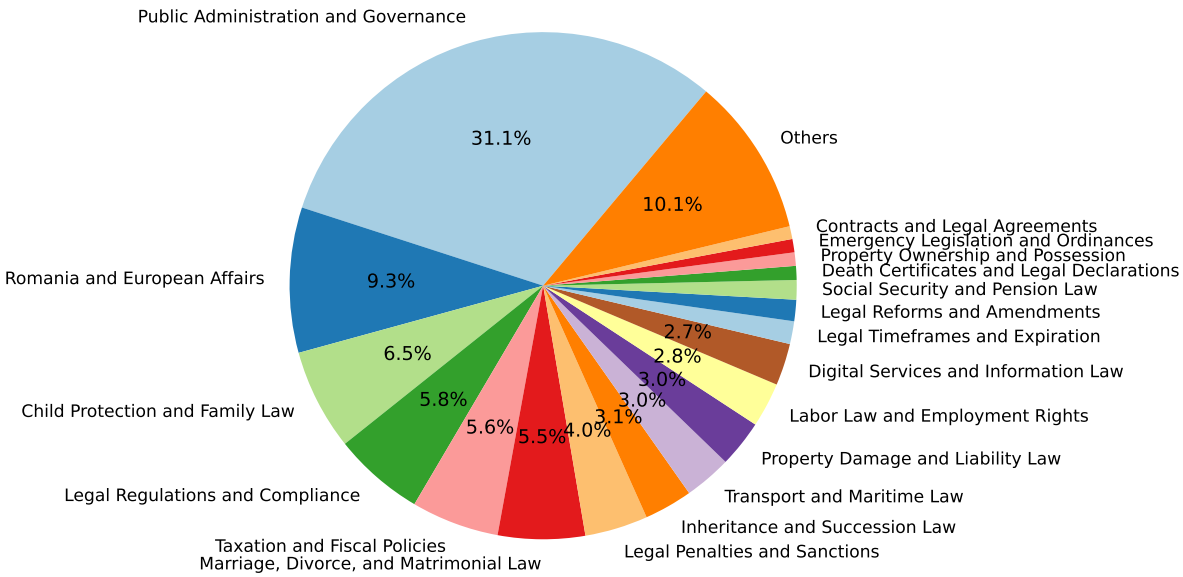

Figure 8: The distribution of top 20 topics in the CROL dataset.

subjects such as *sexual crimes* and *jurisdiction conflicts*. However, the agreement is below 50% for most topics.

Additionally, we estimate the difficulty of each question per topic based on the model's performance. We normalize performance per model to account for the fact that some models perform better than others. If a better model fails on a question that weaker models also fail on, the question is likely to be more difficult. Formally, for every i sample, we first compute the performance score $score_{i,m}$ assigning 1 for every correct prediction with ground truth and 0 otherwise for every model experiment m. Then we calculate the overall per-model performance $\mu_m = \text{mean}_i(score_{i,m})$ and standard deviation $\sigma_m = \text{std}_i(score_{i,m})$ for every model $m \in M$ across all topics. For every prediction i associated with a model m, the z-score is defined as:

$$z\text{-}score_{i,m} = \frac{score_{i,m} - \mu_m}{\sigma_m} \quad (11)$$

Then, to compute the topic-based z-score, we average the z-scores within a given topic t for all models $m \in M$:

$$z\text{-}score_t = \frac{1}{|t| \cdot |M|} \sum_{i \in t} \sum_{m \in M} z\text{-}score_{i,m} \quad (12)$$

The final z-scores are shown for the most frequent topics in Figure 12. The most straightforward topics from the LLM perspective are *procedural exceptions and errors*, *jurisdiction and court competence*, *corporate law*, and *court summons and citations*. On the other end of the spectrum, the most challenging questions were related to *constitutional* and *administrative* laws.

We present a multi-dimensional analysis in Figure 13 considering model accuracy, z-score-based question difficulty, and topic size. We demonstrate that model performance is influenced by the pre-training dataset (i.e., whether it includes the Romanian language), the number of parameters, and the difficulty of the questions.

D Model Architectures

Autoregressive Models. These models are known to exhibit impressive capabilities in generative tasks. They can also be adapted to classification tasks by teaching them the correlation between the class concept and the chosen class symbol. Their goal is to minimize the negative log-likelihood of the class symbol given the input. Specifically:

$$\mathcal{L} = -\sum \log P(t_i^j | \mathcal{Z}(Q_i, C_i, t_i^k); \boldsymbol{\theta}) \quad (13)$$

where $j > k$ and t_i^0 represents the empty sequence. The $\mathcal{Z}$ function maps a given triplet to a sequence that can be processed by the given probabilistic model, known as the model prompt. The prompt serves to facilitate and guide the model towards a lower on-average negative log-likelihood, and

consequently, the correct answer. Although our work also explores sequence-to-sequence models, the ones we chose in particular generate output in an autoregressive manner via the decoder module; thus, our previous discussion still holds.

Encoder Models. These models showed excellent performance on classification tasks despite their relatively smaller size in practice. They feature a good semantic understanding of a given sequence via their pre-training objectives. For instance, BERT (Devlin et al., 2019) featured word- and sentence-level pre-training, which allowed it to gain a semantic understanding of language. However, they do not exhibit symbol-level correlation (Robinson et al., 2023), unlike LLMs, and thus, we resort to using their semantic understanding of textual sequences to output a number that represents the degree to which a given choice is correct given a question. We consider two learning goals for these models, the binary cross-entropy minimization for models outputting probabilities:

$$\mathcal{L}_1 = -(o_i^j \log(y_i^j) + (1 - o_i^j) \log(1 - y_i^j)) \quad (14)$$

where $o_i^j = \mathbb{1}_{T_i}(C_i^j)$ is the ground truth and y_i^j is the model output probability. We also use the cosine similarity embedding loss to align a given question with the correct answer choice:

$$\mathcal{L}_2 = (1 + o_i^j)(1 - y_i^j) + (1 - o_i^j) y_i^j \quad (15)$$

where o_i^j has the same meaning as above except the negative class becomes -1 instead of 0, whereas $y_i^j = \cos(\bar{Q}_i, \bar{C}_i^j)$ and we refer to the bar notation as the embeddings of the question and choice respectively.

During inference, we consider the question along with the set of choices and select the top $|T_i|$ scores in the following way:

$$Y_i^* = \text{TopK}(Y_i, |T_i|) \quad (16)$$

where $Y_i = \{y_i^j | y_i^j = \text{Score}(Q_i, C_i^j)\}$. TopK is a generalized argmax function that selects the best K candidates from a given list. In the end, the chosen options by the model are C_i^k with $k \in Y_i^*$.

E Experimental Setup

Training was performed on the entire JuRO dataset for each model and, for testing, we considered the checkpoint with the best evaluation results obtained during the training phase. For encoders, we used BERT-based models that were trained for 50 epochs, even though in almost all cases, the best model was found around epoch 10. We used a learning rate of 10^{-7} and the AdamW (Kingma and Ba, 2015) optimizer via vanilla *PyTorch*. All BERT models were fine-tuned on all parameters. LLMs were fine-tuned for 50 epochs using the *Trainer API* provided by the *transformers* library using a learning rate of 10^{-7}, AdamW optimizer, LoRA (Hu et al., 2022) alpha of 32, LoRA rank 64, and 2 warm-up steps. All of our experiments were performed on a single NVIDIA A100 80GB to which we had limited and restricted access. We report the results of a single run.

F Baselines

QBERT. We consider the BERT model (Devlin et al., 2019) and construct the input to the LM by appending the given question and the choice in the following way: [CLS] + QUESTION + [SEP] + CHOICE + [SEP]. We then use the classification token to attach a fast forward network (FFN) on top with a sigmoid activation function, which will report a score between 0 and 1 for the correctness of the answer choice.

CrossQBERT. As proposed by Manotumruksa et al. (2020), we proceed by taking the question and the entire set of possible choices and concatenating them in the same fashion as for QBERT. We consider the first three separator tokens and a single FFN, with a sigmoid activation function, which outputs three scores for the same question corresponding to each answer choice. In this way, we provide BERT with more context to gather additional information about neighboring choices, allowing a better and more informed decision.

ColBERT. Initially, an architecture used for information retrieval tasks(Khattab and Zaharia, 2020), we use it for our task because of its underlying philosophy: aligning textual representations. Thus, we use a model to encode the question and a model to encode the individual choice, and we use the resulting embeddings to perform cosine similarity.

LLMs. We use the generalization capabilities of the LLMs (Zhao et al., 2023), having decent performances on tasks in no-data settings and no further training. We perform prompt engineering (Chen et al., 2023b) and extensively experiment with mul-

tiple prompts, ultimately providing the results for the prompts that obtain the best performance. For prompts, see Appendix I and the translations in Appendix J.

LLM RAG. We use Retrieval Augmented Generation (RAG) (Gao et al., 2023) to provide LLMs with contextual information that would answer the question or guide the model towards the answer. We employ the BM25 retriever (Robertson and Jones, 1976) along with the *SpaCy* package for text normalization (tokenization and lemmatization) to extract the top 10 most relevant documents from the corpus. We take the articles from the CROL corpus and split them into 50-word documents. We allow consecutive chunks to overlap by 25 words to maintain context and avoid abrupt disruptions to the flow of information.

LLM LFT. Finally, for our experiments, we fine-tune these LLMs using the LoRA (Hu et al., 2021) adaptation method, which was experimentally shown to match the performance of classic full parameter fine-tuning. This, together with the previous baseline, achieves the best results among the baselines. We opt for the LoRA strategy, since our computational resources would not allow a full arameter fine-tuning of all our proposed LLMs.

G Language Models

The BERT models that we used in our work are the RoBERT (Masala et al., 2020), a Romanian BERT model trained on the general domain, and jurBERT (Masala et al., 2021) – a Romanian BERT model trained on the legal domain. After attempting multiple LLM models and comparing their preliminary performances on this task, we identified the best-performing LLMs, which were of reasonable size for our resources, for the conducted experiments. We used Flan-T5 (Raffel et al., 2020), XL 3B and XXL 11B variants, Mistral 7B Instruct (Jiang et al., 2023), and Llama 3.1 8B (Dubey et al., 2024). These LLMs are instruction-tuned, and we opt for this type of LLM for its better performance on instruction-following and target tasks (Wei et al., 2022). We could leverage their initial performance for further fine-tuning. For the GAT model, we used six attention heads, as we experimentally observed that this value represents an equilibrium between the average number of non-zero entries and computational demands when the GAT is initialized and tested with given inputs, aiming to potentially mitigate the dying gradient phenomenon caused by null entries. For KG construction and claim extraction, we used the Mixtral-8x7B-Instruct LLM (Jiang et al., 2024a) quantized to 8 bits using the int8 algorithm (Dettmers et al., 2022) implemented in the *bitsandbytes* library. Although this model is relatively large, we utilized it in the KG construction and claim extraction process as a trustworthy means, which is more likely to correctly extract entities and relations. More lightweight solutions can be built by training a smaller language model for this task or by distilling (Hinton, 2015; Gu et al., 2023) Mixtral to a small language model (SLM), for example. Mixtral did not contribute in any way to helping the rest of our framework make the right choice. It had a very clear and definite role in our algorithm, which can be easily replaced with any other lightweight solution that we could not implement due to the lack of available data for the Romanian language in this sense.

H Detailed Evaluation

Tables 12, 13, 14 show all our extensive evaluations conducted on different backbone models and different prompts. P1 and P2 refer to the best prompts for the Mistral and FLAN-T5 models, respectively. Our approach surpasses all the baseline combinations using jurBERT-large (Masala et al., 2021) as our backbone encoder model. Moreover, it outperforms encoder models in all settings.

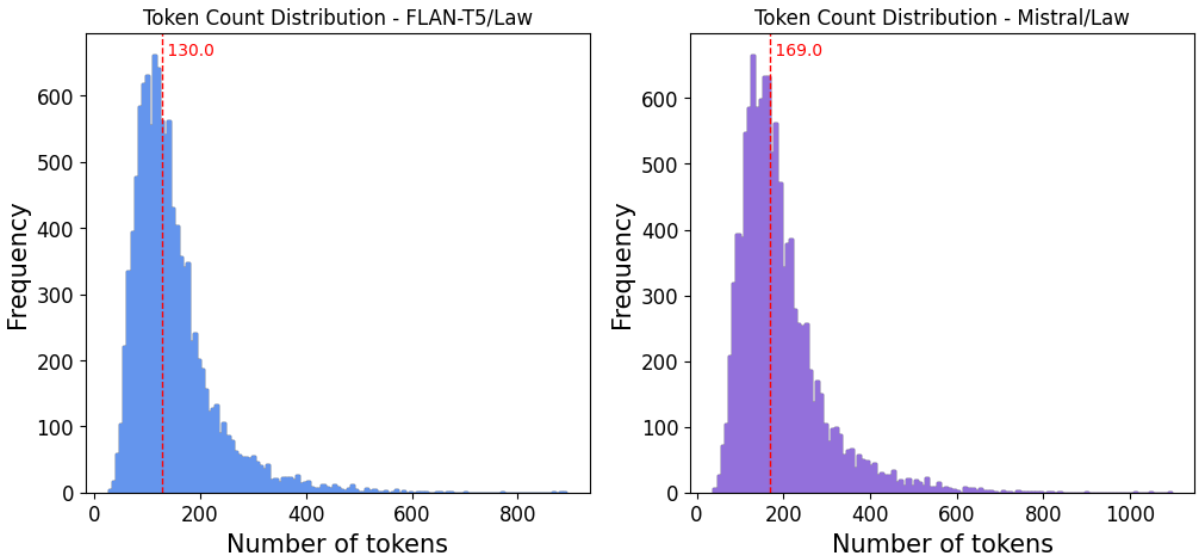

Figure 9: The token distribution on the JuRO dataset for Flan-T5 and Mistral tokenizers.

Task	**Training/Test/Validation**	**# Classes**
Civil	933/266/136	3/6
Entrance	175/50/26	6
Bar	432/123/63	6
Promotion	326/93/47	3
Penal	1380/393/200	3/6
Entrance	175/50/26	6
Bar	430/123/63	6
Promotion	775/220/111	3
Civil Procedure	2070/293/91	3/6
Entrance	174/50/26	6
Bar	431/123/63	6
Promotion	207/67/29	3
Penal Procedure	1282/296/186	3/6
Entrance	174/50/26	6
Bar	432/123/63	6
Promotion	676/123/97	3
Other Promotion Exams	1339/376/195	3
Administrative	351/99/51	3
Commercial	344/98/50	3
Family	355/96/52	3
International	289/83/42	3
Work	343/95/50	3

Table 11: A detailed view of the JuRO dataset regarding the sample distribution among Romanian legal exams and split into training/test/validation sets.

Model	Civil	Penal	Civil Proc.	Penal Proc.	Administrative	Work	Family	International	Commercial	Average
QBERT										
RoBERT-small	29.03	38.29	34.33	29.17	30.69	31.31	31.37	27.38	39.39	32.33
RoBERT-base	25.81	37.39	34.33	32.29	36.63	37.37	37.25	27.38	36.36	33.87
RoBERT-large	33.33	35.14	29.85	35.94	27.72	31.31	38.24	30.95	36.36	32.20
jurBERT-base	35.48	28.38	35.82	37.50	25.74	40.40	29.41	28.57	30.30	32.40
jurBERT-large	32.26	34.68	31.34	40.10	25.74	27.27	29.41	39.29	39.39	29.94
CrossQBERT										
RoBERT-small	35.48	33.78	32.84	38.54	41.58	26.26	28.43	42.86	35.35	35.01
RoBERT-base	41.94	28.83	28.36	28.65	36.63	33.33	28.43	26.19	30.30	31.41
RoBERT-large	41.94	29.28	34.33	29.69	31.68	29.29	24.51	35.71	35.35	32.42
jurBERT-base	41.94	31.08	37.31	35.42	30.69	35.35	29.41	30.95	36.36	34.28
jurBERT-large	29.03	36.04	35.82	41.15	30.69	35.35	34.31	26.19	34.34	33.66
ColBERT										
RoBERT-small	33.33	31.98	32.83	36.46	32.67	33.33	37.25	44.05	34.34	35.14
RoBERT-base	40.86	29.28	31.34	33.33	41.58	31.31	41.18	48.81	35.35	37.00
RoBERT-large	33.33	38.29	37.31	37.50	27.72	29.29	36.27	32.14	40.40	36.69
jurBERT-base	36.56	33.78	37.31	35.42	22.77	42.42	32.35	35.71	33.33	34.41
jurBERT-large	44.09	32.43	47.76	32.81	35.64	36.36	36.27	36.90	39.39	37.96
LLM ZS										
FLAN-T5 XL	39.36	39.29	41.00	36.08	40.59	39.39	40.20	45.24	39.39	39.44
FLAN-T5 XXL	48.94	40.18	38.00	37.63	46.53	37.37	48.04	52.38	37.37	42.94
Mistral-Instruct v0.1	47.87	35.27	34.96	37.11	44.55	43.43	46.07	41.67	37.37	40.92
Mistral-Instruct v0.2	41.49	39.73	35.63	42.27	44.55	48.48	49.02	47.62	37.37	42.80
Llama-3.1 8b Instruct	46.53	33.04	32.84	32.47	46.53	38.38	29.41	46.43	41.41	38.56
LLM RAG										
FLAN-T5 XL	51.06	39.29	32.84	40.21	47.52	47.47	58.82	46.43	51.52	46.13
FLAN-T5 XXL	47.87	43.75	31.34	42.27	55.45	56.57	60.78	47.62	**57.58**	49.25
Mistral-Instruct v0.1	42.55	34.82	34.33	35.05	51.49	41.41	55.88	47.62	50.51	43.74
Mistral-Instruct v0.2	51.06	39.29	38.81	42.78	57.43	53.54	63.73	45.24	56.57	49.83
Llama-3.1 8b Instruct	**53.19**	43.30	40.30	33.51	54.46	58.59	62.75	64.28	**57.58**	52.00
LLM LFT										
FLAN-T5 XL	45.74	50.00	71.64	43.30	48.51	45.45	54.90	55.95	45.45	51.22
FLAN-T5 XXL	45.74	51.79	74.63	48.97	50.50	49.49	52.94	**70.24**	53.53	55.31
Mistral-Instruct v0.1	38.30	46.43	71.64	42.78	52.48	49.49	54.90	65.48	47.47	52.11
Mistral-Instruct v0.2	42.55	51.79	67.16	45.88	48.51	47.47	53.92	**70.24**	49.49	53.00
Llama-3.1 8b Instruct	45.74	50.00	68.66	47.94	54.46	48.48	51.97	66.67	55.56	54.39
GRAF(Ours)										
RoBERT-small	35.48	49.10	67.69	42.11	50.00	48.96	48.48	54.79	45.26	49.10
RoBERT-base	46.24	50.45	69.23	42.63	45.91	**57.29**	57.58	58.90	41.05	52.14
RoBERT-large	47.31	**52.70**	66.15	**51.05**	**59.18**	51.04	56.57	61.64	50.53	55.13
jurBERT-base	49.46	49.55	**78.46**	46.84	53.06	50.00	57.58	58.90	48.42	54.70
jurBERT-large	45.16	47.30	69.23	45.79	47.96	54.17	**68.69**	67.12	56.84	**55.81**

Table 12: Detailed results for promotion exams.

Model	Civil	Penal	Civil Proc.	Penal Proc.	Average
QBERT					
RoBERT-small	29.41	49.10	37.25	33.33	37.27
RoBERT-base	27.45	45.09	27.45	25.49	31.37
RoBERT-large	29.41	37.25	37.25	35.29	34.80
jurBERT-base	41.18	33.33	33.33	31.37	34.80
jurBERT-large	31.37	21.57	33.33	39.22	31.37
CrossQBERT					
RoBERT-small	21.57	49.02	43.14	37.25	37.75
RoBERT-base	31.37	27.45	27.45	25.49	27.94
RoBERT-large	41.18	29.41	29.41	31.37	32.84
jurBERT-base	25.49	35.29	35.29	42.14	34.55
jurBERT-large	35.29	41.18	43.14	27.45	36.77
ColBERT					
RoBERT-small	23.53	33.33	35.29	27.45	29.90
RoBERT-base	33.33	29.41	27.45	21.57	27.94
RoBERT-large	21.57	43.14	50.98	31.37	36.77
jurBERT-base	31.37	33.33	43.14	41.18	37.26
jurBERT-large	45.10	41.18	29.41	31.37	36.77
LLM ZS					
FLAN-T5 XL	29.41	21.57	13.72	27.45	23.04
FLAN-T5 XXL	31.37	25.49	15.69	25.49	24.51
Mistral-Instruct v0.1	33.33	21.57	9.80	25.49	22.55
Mistral-Instruct v0.2	39.21	19.61	11.76	21.57	23.04
Llama-3.1 8b Instruct	33.33	31.37	17.65	27.46	27.45
LLM RAG					
FLAN-T5 XL	31.37	17.65	21.57	23.53	23.53
FLAN-T5 XXL	39.22	27.45	21.57	27.45	28.92
Mistral-Instruct v0.1	11.76	7.84	5.88	15.69	10.29
Mistral-Instruct v0.2	43.14	23.53	21.57	27.45	28.92
Llama-3.1 8b Instruct	35.29	11.76	23.53	13.73	21.08
LLM LFT					
FLAN-T5 XL	35.29	45.10	41.18	41.18	40.69
FLAN-T5 XXL	37.25	43.14	35.29	29.41	36.27
Mistral-Instruct v0.1	19.61	29.41	29.41	27.45	26.47
Mistral-Instruct v0.2	17.65	17.65	17.65	23.53	19.12
Llama-3.1 8b Instruct	31.76	33.33	19.61	35.29	30.00
GRAF(Ours)					
RoBERT-small	43.14	52.94	45.10	45.10	46.57
RoBERT-base	**60.78**	58.82	41.18	50.98	52.94
RoBERT-large	52.94	**62.75**	**54.90**	**56.86**	**56.86**
jurBERT-base	47.06	52.94	50.98	45.10	49.02
jurBERT-large	54.90	54.90	52.94	49.01	52.94

Table 13: Detailed results for entrance exams.

Model	Civil	Penal	Civil Proc.	Penal Proc.	Average
QBERT					
RoBERT-small	37.90	32.26	33.06	32.26	33.87
RoBERT-base	26.61	27.42	29.84	36.29	30.04
RoBERT-large	34.68	32.26	40.32	33.87	35.28
jurBERT-base	33.87	23.58	33.06	33.06	30.89
jurBERT-large	40.32	31.71	32.26	30.65	33.74
CrossQBERT					
RoBERT-small	34.68	31.97	39.52	34.68	35.21
RoBERT-base	30.65	38.52	29.84	30.65	32.42
RoBERT-large	41.94	36.07	33.87	34.68	36.64
jurBERT-base	22.58	34.43	29.84	34.68	30.28
jurBERT-large	36.29	27.05	31.45	30.65	31.36
ColBERT					
RoBERT-small	29.03	33.06	28.23	27.42	29.44
RoBERT-base	38.71	27.42	34.68	29.84	32.66
RoBERT-large	36.29	32.26	36.29	39.52	36.09
jurBERT-base	23.39	29.03	31.45	35.48	29.84
jurBERT-large	29.03	33.87	35.48	33.87	33.06
LLM ZS					
FLAN-T5 XL	19.35	25.00	17.74	12.90	18.75
FLAN-T5 XXL	21.77	24.19	23.38	18.54	21.97
Mistral-Instruct v0.1	14.52	12.10	16.13	13.70	14.11
Mistral-Instruct v0.2	19.35	22.58	15.32	14.52	17.94
Llama-3.1 8b Instruct	16.13	24.19	20.16	19.35	19.96
LLM RAG					
FLAN-T5 XL	21.77	22.58	20.16	16.94	20.36
FLAN-T5 XXL	20.97	26.61	33.87	20.97	25.61
Mistral-Instruct v0.1	5.65	11.29	12.90	5.65	8.87
Mistral-Instruct v0.2	26.61	29.03	20.16	19.35	23.79
Llama-3.1 8b Instruct	25.00	26.61	16.12	23.38	22.78
LLM LFT					
FLAN-T5 XL	31.45	32.26	29.03	36.29	32.26
FLAN-T5 XXL	27.42	28.23	33.06	22.58	27.82
Mistral-Instruct v0.1	27.42	31.45	21.77	29.03	27.42
Mistral-Instruct v0.2	13.71	14.52	24.19	23.39	18.95
Llama-3.1 8b Instruct	25.80	24.19	25.00	23.39	24.60
GRAF(Ours)					
RoBERT-small	46.77	39.52	43.90	43.44	43.41
RoBERT-base	48.39	49.19	52.85	46.72	49.29
RoBERT-large	50.00	47.58	49.59	55.73	50.73
jurBERT-base	**51.61**	55.65	51.22	**59.02**	54.38
jurBERT-large	**51.61**	**61.29**	**56.10**	57.38	**56.60**

Table 14: Detailed results for bar exams.

I Romanian Prompts

Prompt 1 – Entrance and Bar Exams
Răspunde la următoarea întrebare de legalitate din {tip drept}. Cel mult 2 răspunsuri sunt corecte. Dacă un singur răspuns este corect, vei răspunde doar cu litera răspunsului corect. Dacă 2 răspunsuri sunt corecte, vei răspunde doar cu literele răspunsurilor corecte: {tip drept} {întrebare} {variante de răspuns}

Prompt 1 – Promotion Exam
Răspunde la următoarea întrebare de legalitate din {tip drept}. Un singur răspuns este corect. Tu vei răspunde doar cu litera răspunsului corect: {tip drept} {întrebare} {variante de răspuns}

Prompt 2 – Entrance and Bar Exams
Răspunde la următoarea întrebare de legalitate din {tip drept}. Cel mult 2 răspunsuri sunt corecte. Dacă un singur răspuns este corect, vei răspunde doar cu litera răspunsului corect. Dacă 2 răspunsuri sunt corecte, vei răspunde doar cu literele răspunsurilor corecte Răspunde cu doar unul dintre simbolurile din lista [A, B, C, AB, AC, BC]: {tip drept} {întrebare} {variante de răspuns}

Prompt 2 – Promotion Exam
Răspunde la următoarea întrebare de legalitate din {tip drept} cu doar una dintre literele din lista [A, B, C]. Un singur răspuns este corect: {tip drept} {întrebare} {variante de răspuns}

FLAN-T5 RAG – Entrance and Bar Exams
Răspunde la următoarea întrebare de legalitate din {tip drept} din context. Cel mult 2 răspunsuri sunt corecte. Dacă un singur răspuns este corect, vei răspunde doar cu litera răspunsului corect. Dacă 2 răspunsuri sunt corecte, vei răspunde doar cu literele răspunsurilor corecte Dacă informația din context nu este în întrebare atunci ignoră contextul și doar răspunde la întrebare. Răspunde cu doar unul dintre simbolurile din lista [A, B, C, AB, AC, BC]. Context: {documente} Întrebare: {întrebare} {variante de răspuns}

FLAN-T5 RAG – Promotion Exams
Răspunde la următoarea întrebare de legalitate din {tip drept} din context cu doar una dintre literele din lista [A, B, C]. Dacă informația din context nu este în întrebare atunci ignoră contextul și doar răspunde la întrebare. Un singur răspuns este corect. Context: {documente} Întrebare: {întrebare} {variante de răspuns}

Mistral RAG – Entrance and Bar Exams
Răspunde la următoarea întrebare de legalitate din {tip drept} din context. Cel mult 2 răspunsuri sunt corecte. Dacă un singur răspuns este corect, vei răspunde doar cu litera răspunsului corect. Dacă 2 răspunsuri sunt corecte, vei răspunde doar cu literele răspunsurilor corecte. Dacă informația din context nu este în întrebare atunci ignoră contextul și doar răspunde la întrebare. Context: {documente} Întrebare: {întrebare} {variante de răspuns}

Mistral RAG – Promotion Exam
Răspunde la următoarea întrebare de legalitate din {tip drept} din context. Un singur răspuns este corect. Tu vei răspunde doar cu litera răspunsului corect. Dacă informația din context nu este în întrebare atunci ignoră contextul și doar răspunde la întrebare. Context: {documente} Întrebare: {întrebare} {variante de răspuns}

LLM Prompt for Claim Graph Extraction

Extrage toate entitățile și toate relațiile dintre entități din textul legal pe baza exemplului. La final adaugă STOP. Tu vei răspunde cu triplete de forma: (entitate;relație;entitate). Tripletele sunt separate pe linii. Fiecare relație triplet se va trece separat. Entitățile pot fi instituții, organizații, persoane, funcții, documente, instanțe și altele.

Text:
(1) Pe lângă fiecare curte de apel va funcţiona o comisie de cercetare a averilor, denumită în continuare comisie de cercetare, formată din:
a) 2 judecători de la curtea de apel, desemnaţi de preşedintele acesteia, dintre care unul în calitate de preşedinte,
b) un procuror de la parchetul care funcţionează pe lângă curtea de apel, desemnat de prim-procurorul acestui parchet.
(2) Preşedintele şi membrii comisiei de cercetare sunt desemnaţi pe o perioadă de 3 ani. Pe aceeaşi perioadă şi de către aceleaşi persoane vor fi desemnaţi şi 3 supleanţi, care îi vor înlocui pe titulari în cazul în care aceştia, din motive legale, nu vor putea lua parte la lucrările comisiei de cercetare.
(3) Comisia de cercetare are un secretar, desemnat de preşedintele curţii de apel dintre grefierii acestei instanţe.

Entitate;Relație;Entitate:
(curte de apel;funcționează pe lângă;comisie de cercetare a averilor)
(comisie de cercetare a averilor;denumită;comisie de cercetare)
(comisie de cercetare;formată din;2 judecători)
(2 judecători;desemnați de;președinte curte de apel)
(comisie de cercetare;formată din;procuror)
(procuror;de la;parchetul care funcționează pe lângă curtea de apel)
(procuror;desemnat de;prim-procuror)
(președinte comisie de cercetare;desemnat pe o perioadă de;3 ani)
(membrii comisiei de cercetare;desemnat pe o perioadă de;3 ani)
(3 supleanți;desemnați de;președinte curte de apel)
(3 supleanți;desemnați de;prim-procuror)
(3 supleanți;desemnați pe o perioadă de;3 ani)
(3 supleanți;îi vor înlocui dacă nu vor putea lua parte la lucrările comisiei de cercetare pe;titulari)
(comisie de cercetare;are;un secretar)
(un secretar;desemnat dintre grefieri de;președinte curte de apel)
STOP

Text:
{text}

Entitate;Relație;Entitate:

J Translated Prompts

Prompt 1 – Entrance and Bar Exams
Answer the following legal question from {law type}. At most 2 answers are correct. If a single answer is correct, you will only answer with the letter of the correct answer. If 2 answers are correct, you will answer only with the letters of the correct answers: {law type} {question} {answer choices}

Prompt 1 – Promotion Exam
Answer the following legal question from {law type}. A single answer is correct. You will only answer with the letter of the correct answer: {law type} {question} {answer choices}

Prompt 2 – Entrance and Bar Exams
Answer the following question from {law type}. At most 2 answers are correct. If a single answer is correct, you will only have an answer with the letter of the correct answer. If 2 answers are correct, you will answer with the letters of the correct letters. Answer with only one of the symbols from the list [A, B, C, AB, AC, BC]: {law type} {question} {answer choices}

Prompt 2 – Promotion Exam
Answer the following legal question from {law type} with only one of the letters from the list [A, B, C]. A single answer is correct: {law type} {question} {answer choices}

FLAN-T5 RAG – Entrance and Bar Exams
Answer the following legal question from {law type}. At most 2 answers are correct. If a single answer is correct, you will only answer with the letter of the correct answer. If 2 answers are correct, you will answer with only the letters of the correct answers. If the information from the context is not in the question, then ignore the context and answer the question. Answer with only one of the symbols from the list [A, B, C, AB, AC, BC]. Context: {documets} Question: {question} {answer choices}

FLAN-T5 RAG – Promotion Exam
Answer the following legal question from {law type} with only one of the letters from the list [A, B, C]. If the information from the context is not in the question, then ignore the context and answer the question. A single answer is correct. Context: {documents} Question: {question} {answer choices}

Mistral RAG – Entrance and Bar Exams
Answer the following legal question from {law type}. At most 2 answers are correct. If a single answer is correct, you will only answer with the letter of the correct answer. If 2 answers are correct, you will answer only with the letters of the correct answers. If the information from the context is not in the question, then ignore the context and answer the question. Context: {documents} Question: {question} {answer choices}

Mistral RAG – Promotion Exam
Answer the following legal question from {law type}. You will only answer with the letter of the correct answer. If the information from the context is not in the question, then ignore the context and answer the question. A single answer is correct. Context: {documents} Question: {question} {answer choices}

LLM Prompt for Claim Graph Extraction

Extract all entities and relationships between entities from the legal text based on the example. In the end, add STOP. You will answer with triplets of the form: (entity;relation;entity). The triplets are separated on lines. Each triplet relationship will be entered separately. Entities can be institutions, organizations, persons, functions, documents, courts and others.

Text:
(1) An assets investigation commission, hereinafter referred to as the investigation commission, shall operate in addition to each court of appeal, consisting of:
a) 2 judges from the court of appeal, designated by its president, one of whom shall act as president,
b) a prosecutor from the prosecutor's office operating under the court of appeal, designated by the chief prosecutor of this prosecutor's office.
(2) The president and members of the investigation commission shall be designated for a period of 3 years. During the same period and by the same persons, 3 alternates will also be appointed, who will replace the holders in the event that they, for legal reasons, are unable to participate in the work of the investigation commission.
(3) The investigation commission has a secretary, appointed by the president of the court of appeal from among the clerks of this court.

Entity;Relationship;Entity:
(court of appeal;shall operated in addition to;assets investigation commission)
(assets investigation commission;referred to as;investigation commission)
(investigation commission;consisting of;2 judges)
(2 judges;designated by;president of the court of appeal)
(investigation commission;consisting of;prosecutor)
(prosecutor;from;prosecutor's office operating under the court of appeal)
(prosecutor;designated by;chief prosecutor)
(president of the investigation commission;designated for a period of;3 years)
(members of the investigation commission;designated for a period of;3 years)
(3 alternates;appointed by;the president of the court of appeal)
(3 alternates;appointed by;the chief prosecutor)
(3 alternates;designated for a period of;3 years)
(3 alternates;will replace the holders if they cannot take part in the work of the investigation commission on;the heads)
(investigation commission;has;a secretary)
(a secretary;appointed by among the clerks of;the president of the court of appeal)
STOP

Text:
{text}

Entity;Relationship;Entity:

K Figures and Tables for Topic Analysis

Topic	Keywords
Apel și Instanță	apel, în, poate, de, care, nu, judecată, instanţa, fi, se
Debitor și Creditor	debitorului, debitorul, debitor, creditor, creditorul, creditorilor, creditorului, insolvenţă, procedurii, insolvenţei
Omor și Victimă	omor, faptei, rezultatul, infracţiunea, victimei, fapta, autorul, făptuitorul, culpă, victima
Proprietate și Contracte	bunului, dreptul, vânzătorul, bunul, vânzării, proprietate, vânzare, cumpărătorului, contractului, cumpărătorul
Termene și Comunicare	zile, termen, data, comunicare, la, termenul, de, comunicării, comunicarea, 30
Pedeapsă și Condamnare	pedepsei, pedeapsa, vigoare, supraveghere, pedeapsă, amenzii, închisorii, condamnare, legea, executării
România și Sistemul Legal	României, românia, român, română, arestare, române, european, decizia, teritoriul, monitorul
Căsătorie și Divorț	căsătoriei, divorţ, soţi, căsătoria, căsătorie, divorţul, desfacerea, culpa, casarea, legea
Furt și Răspundere Penală	infracţiunea, furt, libertate, tentativa, infracțiunea, lipsire, posibilă, art, infracțiunile, infracţiunii
Copil și Drepturi Părintești	copilului, părinteşti, copilul, minorului, părinţilor, vârsta, copil, părintele, drepturile, protecţie
Contracte și Obligații	contractului, contractul, încheiat, contract, mandatarul, mandatului, mandantului, produce, secret, este
Litigii și Părți	bb, aa, reclamantul, judecată, pârâtul, chemare, lei, cerere, acest, contradictoriu
Moștenire și Succesiune	defunctului, moștenire, moştenire, succesorală, moștenirea, moştenirii, lui, moștenirii, culege, moştenirea
Pensie și Retragere din Activitate	cotizare, stagiul, pensiei, invaliditate, pensii, pensie, standard, realizat, pensionare, pensia
Sentință Penală și Apeluri	pedeapsa, ani, închisoare, inculpatul, închisorii, pedepsei, reabilitare, apel, apelul, condamnat
Fals și Fraudă	fals, înscrisul, privată, semnătură, înscrisuri, înscrisului, sub, înscris, 250, oficiale
Drept Administrativ	administrativ, administrative, contencios, publice, actul, act, actului, actele, publică, nelegalitate
Suspendare Judiciară și Hotărâri	suspendarea, suspendare, justiţie, apel, hotărâri, judecăţii, se, rolul, primă, dispusă
Constituționalitate și Lege	art, alin, neconstituţionalitate, curtea, prevederile, constituţională, din, constituţionale, că, excepţia
Mită și Corupție	serviciu, mită, luare, bani, public, abuz, funcţionar, foloase, infracţiunii, sumă
Drept Comercial și Acționari	societăţii, adunarea, social, adunării, acţiuni, generală, societate, capitalul, generale, vot
Furt Calificat și Acuzații	lui, infracțiunea, furt, și, calificat, pe, sarcina, că, inculpatul, un
Competență și Jurisdicție a Instanțelor	grad, conexitate, instanţe, cereri, litispendenţa, divergență, litispendenţă, cealaltă, invocată, două
Paternitate și Parentalitate	copilului, mamei, paternităţii, paternitate, copilul, tată, acţiunea, căsătoriei, născut, timpul
Reglementări UE și Conformitate	membru, stat, statul, regulamentului, european, executoriu, 44, nr, 2001, materie
Competență Legală și Conflicte de Jurisdicție	competenţa, competenţă, apel, instanţei, conflictul, instanţe, competente, rediscutată, îşi, dintre
Citație și Chemare în Instanță	prezentă, partea, termenul, citare, citată, termen, amânarea, studierii, legal, fost
Excepții Procedurale și Erori	procesuale, capacităţii, excepţia, lipsei, folosinţă, excepţiile, invocate, active, erori, necompetenţa
Infracțiuni Sexuale și Viol	viol, sexual, incest, agravată, violare, infracţiunea, sexuală, infracțiunea, domiciliu, prevăzută
Instanțele din București și Jurisdicție	bucureşti, ab, judecătoria, in, procurorul, sector, disp, lângă, pen, suspectul

Table 15: List of top 30 topics and associated keywords from the JuRO dataset, in Romanian.

Topic	Keywords
Appeal and Court	appeal, in, may, of, which, not, trial, court, be, is
Debtor and Creditor	debtor's, debtor, debtor, creditor, creditor's, creditors, creditor's, insolvency, procedure, insolvency
Homicide and Victim	homicide, act, result, crime, victim's, act, author, perpetrator, guilt, victim
Property and Contracts	asset's, right, seller, asset, sale, property, sale, buyer's, contract's, buyer
Deadlines and Communication	days, deadline, date, communication, at, term, of, communication's, communication, 30
Punishment and Sentencing	penalty's, penalty, enforcement, supervision, punishment, fine, imprisonment, conviction, law, execution
Romania and Legal System	Romania's, Romania, Romanian, Romanian, arrest, Romanian, European, decision, territory, official journal
Marriage and Divorce	marriage's, divorce, spouses, marriage, marriage, divorce, dissolution, fault, annulment, law
Theft and Criminal Liability	crime, theft, freedom, attempt, crime, deprivation, possible, article, crimes, crime
Child and Parental Rights	child's, parental, child, minor's, parents', age, child, parent, rights, protection
Contracts and Obligations	contract's, contract, concluded, contract, agent, mandate's, principal's, produce, secret, is
Litigation and Parties	bb, aa, claimant, trial, defendant, summons, lei, request, this, adversarial
Inheritance and Succession	deceased's, inheritance, inheritance, succession, inheritance, inheritance's, his, inheritance's, collects, inheritance
Pension and Retirement	contribution, period, pension's, disability, pensions, pension, standard, achieved, retirement, pension
Criminal Sentencing and Appeals	penalty, years, imprisonment, defendant, imprisonment, penalty's, rehabilitation, appeal, appeal, convicted
Forgery and Fraud	forgery, document, private, signature, documents, document's, under, document, 250, official
Administrative Law	administrative, administrative, litigation, public, act, act, act's, acts, public, illegality
Judicial Suspension and Rulings	suspension, suspension, justice, appeal, rulings, court, is, role, first, ordered
Constitutionality and Law	article, paragraph, unconstitutionality, court, provisions, constitutional, from, constitutional, that, exception
Bribery and Corruption	service, bribe, taking, money, public, abuse, official, benefits, crime, sum
Corporate Law and Shareholders	company's, assembly, social, assembly's, shares, general, company, capital, general, vote
Qualified Theft and Charges	his, crime, theft, and, qualified, on, charge, that, defendant, a
Jurisdiction and Court Competence	level, connection, courts, requests, lis pendens, divergence, lis pendens, other, invoked, two
Paternity and Parenthood	child's, mother's, paternity's, paternity, child, father, action, marriage's, born, time
EU Regulations and Compliance	member, state, state's, regulation's, European, enforceable, 44, no, 2001, matter
Legal Competence and Jurisdiction Conflicts	competence, competence, appeal, court's, conflict, courts, competent, reconsidered, its, among
Court Summons and Citations	present, party, deadline, citation, cited, term, postponement, study, legal, was
Procedural Exceptions and Errors	procedural, capacity's, exception, lack, use, exceptions, invoked, active, errors, incompetence
Sexual Crimes and Violations	rape, sexual, incest, aggravated, violation, crime, sexual, crime, domicile, provided
Bucharest Courts and Jurisdiction	Bucharest, ab, court, in, prosecutor, sector, ordered, near, criminal, suspect

Table 16: List of top 30 topics and associated keywords from the JuRO dataset, translated to English.

Topic	Keywords
Administrație Publică și Guvernare	publice, publici, de, al, şi, sau, în, consiliului, și, care
România și Afaceri Europene	românia, românieі, în, de, din, europene, române, și, la, al
Protecția Copilului și Dreptul Familiei	copilului, copilul, tutelă, minorului, familie, copil, vârsta, minorul, de, tutore
Reglementări Legale și Conformitate	articolul, următorul, cuprins, avea, modifică, alineatul, va, şi, la, se
Impozitare și Politici Fiscale	fiscal, fiscale, fiscală, din, şi, de, pentru, organul, al, în
Căsătorie, Divorț și Drept Matrimonial	căsătoriei, căsătoria, soţi, căsătorie, soți, matrimonial, divorț, dintre, comune, soții
Pedepse Legale și Sancțiuni	ani, amendă, închisoare, închisoarea, 000, pedepseşte, lei, pedeapsa, la, cu
Moștenire și Dreptul Succesiunii	moştenire, moștenire, moştenirii, moștenirii, defunctului, moştenirea, moștenirea, privilegiaţi, succesorală, privilegiați
Transport și Drept Maritim	transport, transportatorul, vasului, transportului, transportatorului, vasul, expeditorul, capitanul, sau, pentru
Deteriorarea Proprietății și Răspundere Legală	bunului, prejudiciul, locatarul, prejudiciului, cauzat, dacă, bunul, este, repararea, să
Dreptul Muncii și Drepturile Angajaților	muncă, colective, colectiv, angajatorul, sindicale, individual, salariatul, salariatului, muncii, de
Servicii Digitale și Dreptul Informației	publice, servicii, electronic, electronice, și, ministerul, furnizorul, electronică, şi, informației
Termene Legale și Expirare	termenul, zi, termen, ziua, ani, prescrie, zile, ultima, data, termenului
Reforme Legale și Amendamente	2019, publicată, monitorul, oficial, 597, ix, 07, 05, ordonanța, urgență
Securitate Socială și Dreptul Pensiilor	investiţii, şi, cnpp, snpct, pensii, 01, de, consiliul, din, financiare
Certificate de Deces și Declarații Legale	decedate, moartea, morţii, morții, decesul, mort, cel, declarat, viaţă, viață
Proprietate și Posesiune	mobile, bunurile, mobil, bunurilor, bun, bunuri, bunului, imobile, debitorului, sunt
Legislație de Urgență și Ordonanțe	ordonanţa, urgenţă, din, nan, ordonanța, urgență, persoanele, astăzi, stat, de
Contracte și Acorduri Legale	persoanei, private, unei, fără, utilizarea, sau, difuzarea, acordul, persoane, precum
Fructe și Drept Agricol	fructele, fructelor, industriale, naturale, cuvin, bunului, recoltelor, civile, fructe, dreptul
Proprietate Funciară și Cadastru	zidului, comune, zidul, linia, hotar, clădirii, apartamente, clădire, necomunitate, proprietari
Drepturi de Proprietate și Liberalități	liberalitatea, liberalității, liberalităţii, dispunătorul, instituitului, impută, primi, dispunător, rezervatari, liberalitate
Drepturile Omului și Identitate Civilă	umane, fizică, civile, imagine, portret, psihică, dreptul, ființei, orice, civilă
Identificare Legală și Aspecte Familiale	identificare, numele, numelui, nume, persoanei, prenumele, atribute, juridice, familie, identificarea
Transfer Juridic și Divizare	juridice, persoanei, organic, divizarea, patrimoniul, multe, legea, juridică, persoane, transferă
Dreptul Ospitalității și Răspundere	hotelierul, hotel, hotelierului, clientului, aduse, cazare, hotelului, bunurilor, cazării, răspunderea
Capacitate Juridică și Drepturile Minorilor	exerciţiu, exercițiu, capacitate, capacitatea, restrânsă, deplină, minorul, actele, lipsit, tutelă
Solidaritate și Răspundere în Grupuri	solidar, răspund, nelimitat, primii, fondatorii, grupului, prejudiciul, fondatori, cauzat, sînt
Datorii și Drepturile Creditorilor	cedată, creanța, creanţa, creanței, cesiune, cesiunea, creanţei, cesionarul, cesionarului, constatator
Instrumente Financiare și Dreptul Garanției	girul, alb, gir, cec, litere, suma, carat, sa, avalul, este

Table 17: List of top 30 topics and associated keywords from the CROL dataset, in Romanian.

Topic	Keywords
Public Administration and Governance	public, public, of, the, and, or, in, council, and, which
Romania and European Affairs	romania, romania's, in, of, from, european, romanian, and, at, the
Child Protection and Family Law	child, the child, guardianship, minor, family, child, age, the minor, of, tutor
Legal Regulations and Compliance	article, following, content, have, modifies, paragraph, will, and, at, is
Taxation and Fiscal Policies	fiscal, fiscal, fiscal, from, and, of, for, body, the, in
Marriage, Divorce, and Matrimonial Law	marriage, the marriage, spouses, marriage, spouses, matrimonial, divorce, between, common, spouses
Legal Penalties and Sanctions	years, fine, imprisonment, the imprisonment, 000, punishes, lei, penalty, at, with
Inheritance and Succession Law	inheritance, inheritance, inheritance, inheritance, of the deceased, the inheritance, inheritance, privileged, succession, privileged
Transport and Maritime Law	transport, transporter, vessel, transport, the transporter, vessel, sender, captain, or, for
Property Damage and Liability Law	property, damage, tenant, the damage, caused, if, the property, is, repair, to
Labor Law and Employment Rights	work, collective, collective, employer, union, individual, employee, the employee, work, of
Digital Services and Information Law	public, services, electronic, electronic, and, ministry, provider, electronic, and, information
Legal Timeframes and Expiration	term, day, term, the day, years, prescribes, days, last, date, term
Legal Reforms and Amendments	2019, published, monitor, official, 597, ix, 07, 05, ordinance, emergency
Social Security and Pension Law	investments, and, cnpp, snpct, pensions, 01, of, council, from, financial
Death Certificates and Legal Declarations	deceased, death, of death, of death, death, dead, the, declared, life, life
Property Ownership and Possession	movable, properties, mobile, properties, property, goods, property, real estate, debtor, are
Emergency Legislation and Ordinances	ordinance, emergency, from, nan, ordinance, emergency, people, today, state, of
Contracts and Legal Agreements	person, private, one, without, use, or, dissemination, agreement, persons, such as
Fruits and Agricultural Law	fruits, of the fruits, industrial, natural, words, property, harvests, civil, fruit, right
Land Ownership and Cadastre	wall, common, the wall, line, border, building, apartments, building, non-community, owners
Property Rights and Liberal Ownership	liberality, liberality, liberality, disposer, instituted, imputes, receive, disposer, reserved, liberality
Human Rights and Civil Identity	human, physical, civil, image, portrait, mental, right, being, any, civil
Legal Identification and Family Matters	identification, name, the name, name, person, first name, attributes, legal, family, identification
Juridical Transfer and Division	legal, person, organic, division, patrimony, many, law, legal, persons, transfers
Hospitality Law and Liability	hotelier, hotel, the hotelier, client, damages, accommodation, hotel, goods, accommodation, liability
Legal Capacity and Minor Rights	exercise, exercise, capacity, the capacity, restricted, full, the minor, acts, deprived, guardianship
Solidarity and Liability in Groups	solidarity, respond, unlimited, first, founders, group, damage, founders, caused, are
Debt and Creditors' Rights	assigned, claim, claim, claim, assignment, assignment, claim, assignee, assignee, certifier
Financial Instruments and Surety Law	endorsement, blank, endorsement, check, letters, amount, carat, his, guarantee, is

Table 18: List of top 30 topics and associated keywords from the CROL dataset, translated to English.

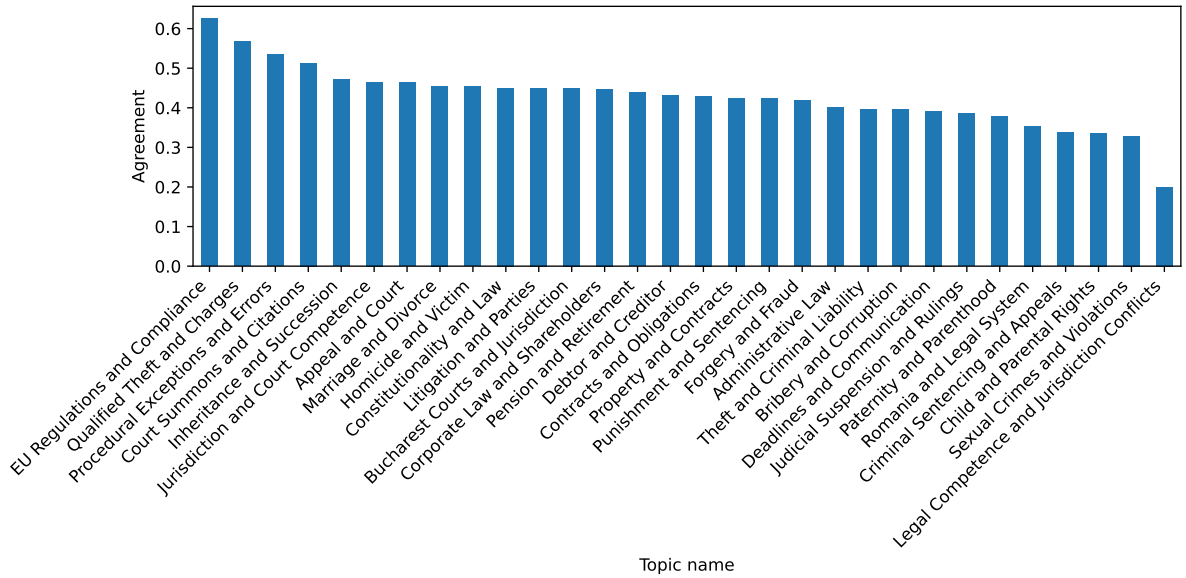

Figure 10: Per topic average pairwise percentage agreement scores when employing Llama-3.1 8B Instruct, FLAN-T5 XL, FLAN-T5 XXL, Mistral 7B Instruct v0.1, and Mistral 7B Instruct v0.2 LLMs. Higher is better.

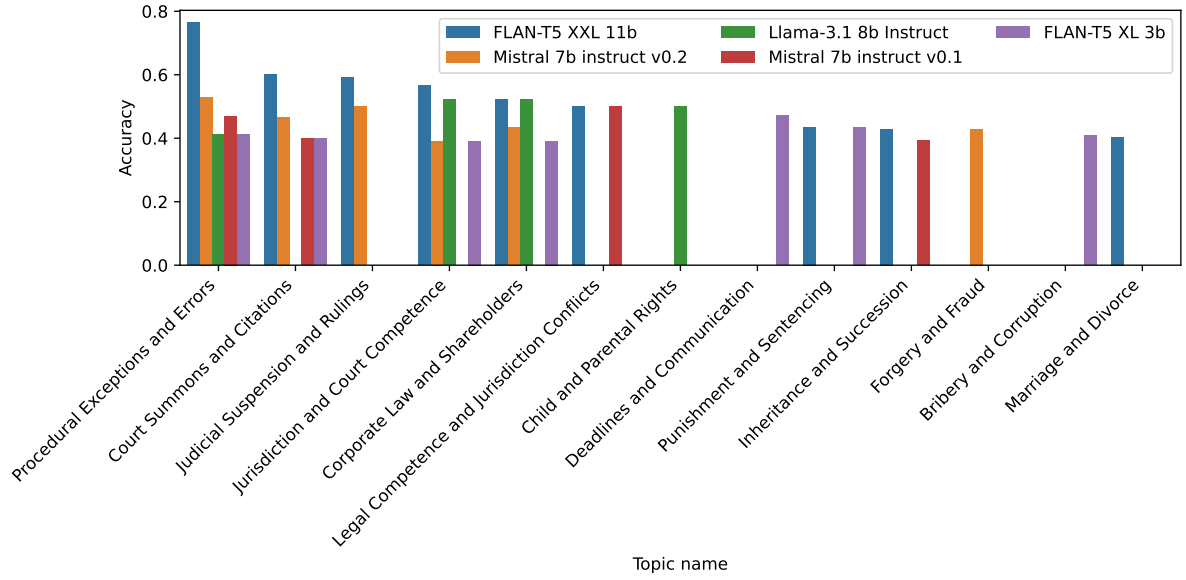

Figure 11: Accuracy computed for samples in the top 13 topics from the JuRO dataset, for every LLM. Higher is better.

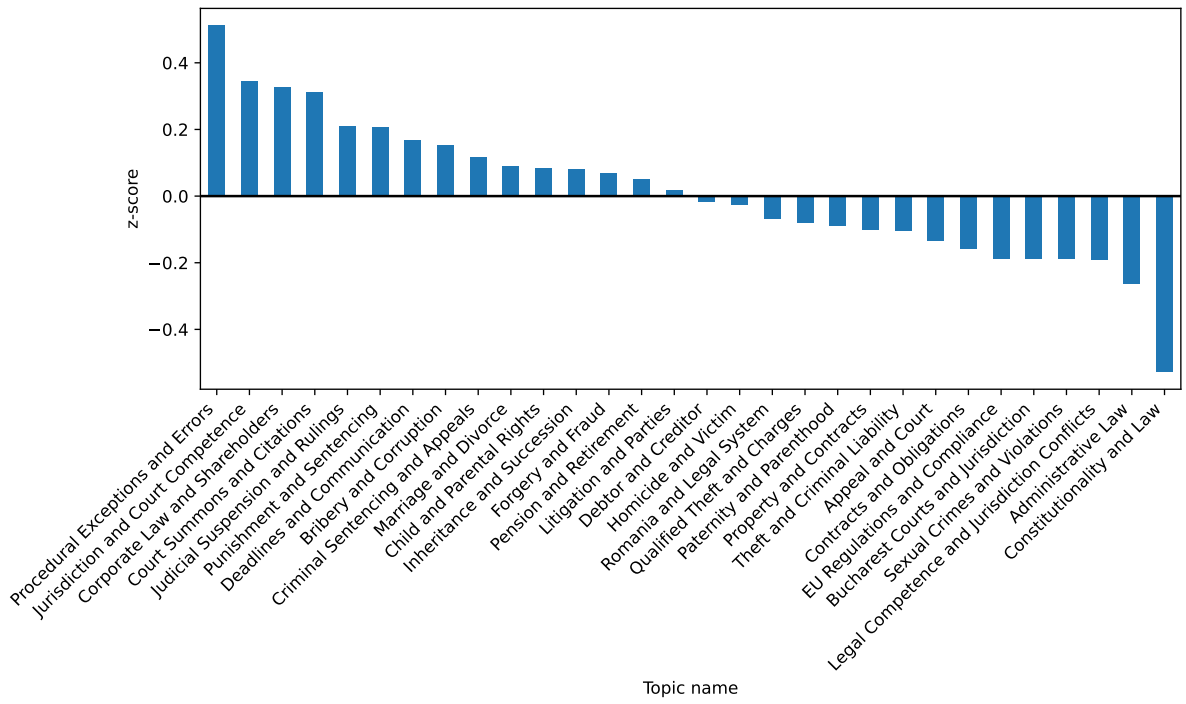

Figure 12: Per-topic question difficulty on the JuRO dataset relative to LLM performance using the z-score normalization. High positive values indicate that the questions from the given topic are easier, while lower negative values indicate that the questions from a given topic are more difficult.

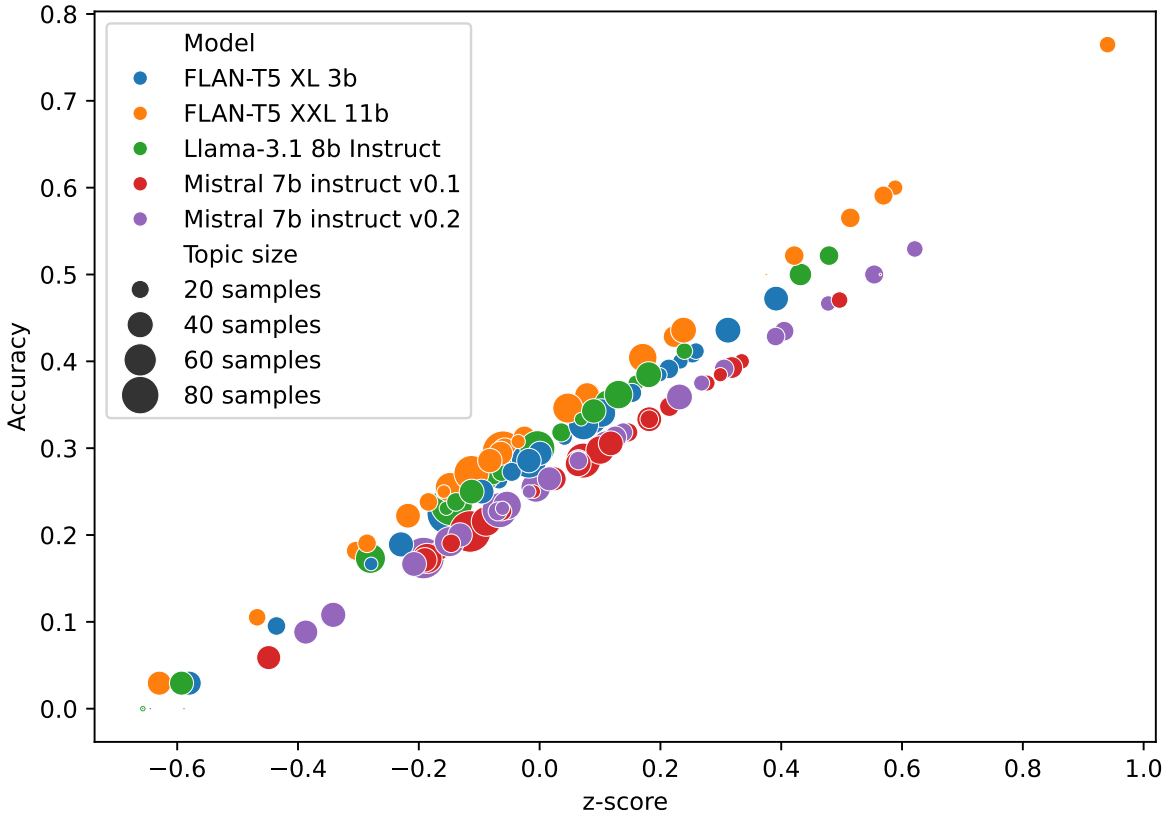

Figure 13: The dependency between accuracy, question difficulty (as z-score), model, and topic size. Larger language models having the Romanian language in the training set (i.e., FLAN-T5) perform better than smaller models trained on English-only data (i.e., Mistral 7B). Most topics reside in the medium to higher difficulty levels from the LLM performance perspective (i.e., z-score less than 0), achieving lower accuracy on those topics (i.e., under 40%). There is a single exception for FLAN-T5 XXL on *Procedural Exceptions and Errors*, achieving 76% with a z-score of 0.94. At the bottom of the scale, the models perform worse at *Constitutionality and Law* and *Legal Competence and Jurisdiction Conflicts* topics.