Generative AI System to Support Decision-Making on Public Procurement Legislation by Fine-Tuned Large Language Models

Dragoș Datcu
Independent Research
The Netherlands
email@dragosdatcu.eu

Clara Maathuis
Open University
The Netherlands
clara.maathuis@ou.nl

ABSTRACT
Public procurement processes are often hindered by the complex nature of implementing domain-specific directives and laws, leading to human errors in decision-making during acquisition tasks. These errors can significantly impact the timely execution of national and European government projects. Given recent developments in the Generative AI (Artificial Intelligence) domain, to address this challenge, we present an innovative LLMs-based (Large Language Models) system containing a teacher-student architecture based on LLaMA 3-70B as teacher model and a series of student models like RoMistral-7B, LLaMA 3-8B, Mistral- 7B, Gemma-7B and Saul-7B specifically adapted for interpreting and reasoning on Romanian public procurement legislation.

This explorative solution is aimed to support legal experts in navigating the intricacies of procurement laws and directives, thereby enhancing decision-making accuracy and reducing errors in public acquisition tasks. The solution is developed based on brainstorming sessions conducted with experts in Romanian public procurement. To this end, a small-scale experiment is conducted to evaluate the approach by employing both quantitative and qualitative analysis methods. The results of our preliminary study demonstrate remarkable performance in understanding and reasoning on the public procurement corpus, with the fine-tuned model showing promising capabilities in interpreting complex legal texts and providing valuable insights for procurement professionals.

Keywords
LLM; Llama3 fine-tuning; Legal corpora.

DOI: 10.37789/icusi.2024.3

INTRODUCTION

Public procurement legal specialists deal with a complex landscape of regulations, directives, and procedures when managing the acquisition of goods, services, and works for public entities. This represents a challenging process due to the vast array of procurement legislation and documentation, the various ways of interpreting asset specifications, and the underlying dynamic and evolving nature of the laws involved when establishing and dealing with public contracts [9, 10]. In this process, the legal experts need to synthesize information from diverse sources, including specific laws, relevant national and international procurement laws, EU directives, and judicial decisions together with administrative guidelines, while addressing the unique requirements of each formal process (i.e., tender). This often happens under conditions of stress when the experts deal with strict deadlines, potential conflicts of interest, need of additional research processes, and potentially missing or overlooking details that could lead to errors when dealing with vast and/or complex documents [5].

The introduction of digital technologies to assist legal experts in public procurement processes and corresponding legal decisions offers significant potential for enhancing efficiency and transparency. These technologies include solutions based on AI (Artificial Intelligence), blockchain, and cloud computing, and are built in order to automate a series of repetitive tasks, provide analysis of specific articles, points, or spending patterns. Hence, these technologies support the overall decision-making processes of legal experts, they also contribute to increasing the transparency, consistency, and effectiveness of the procurement processes. Nevertheless, the implementation of such solutions also presents challenges. To name a few, the increase of complexity of the procurement processes as the legal experts also need to become familiar with these digital systems, a fact that may require dedicated training. Moreover, the impact on accountability and degree of autonomy that the system offers, which can be established since the design of the system [18, 23].

However, while AI is perceived as a promising avenue for supporting legal experts in this domain, their availability and implementation is limited. These AI systems, designed to process and analyze vast amounts of textual data, could significantly enhance the efficiency and accuracy of procurement processes by rapidly sifting through complex legal documents, identifying relevant clauses, and highlighting potential compliance issues. In particular, given the recent advancements in the Generative AI field, and in particular in the area of Large Language Models (LLMs), such a system can be particularly beneficial due to their ability to process and analyze vast volumes of text data [32], including procurement legislation, case law, tender specifications, and historical procurement records.These models can rapidly extract relevant information, identify patterns, and provide contextual insights that would be time- consuming and challenging for experts to accomplish manually under time constraints [24]. Hence, a LLMs-based solution has the potential to enhance decision-making processes by providing quick, data-driven analyses of complex procurement scenarios, helping experts to identify potential risks, evaluate tender submissions more efficiently, and ensure adherence to principles such as transparency and fairness.

To address this gap, this research aims to build a supportive explorative LLMs-based system that provides assistance on interpreting and reasoning on Romanian public procurement legislation. It does that by building a teacher-student approach where a LLaMA 3-0B teacher model generates samples to train smaller models, i.e., RoMistral-7B, Llama3- 8B, Mistral-7B, Gemma-7B and Saul-7B on specific legislation questions and answers. On this behalf, a series of brainstorming sessions with experts in Romanian public procurement were carried out between January - March 2024. This research contributes to the growing field of AI applications in legal and governmental processes, potentially improving the efficiency and accuracy of public procurement procedures. Our findings suggest that LLMs-based systems have the potential to play an important role in mitigating the challenges associated with interpreting and implementing public procurement legislation, ultimately leading to more effective and error-free procurement processes.

The remainder of this research is structured as follows. Section Related Work discusses relevant studies in this domain. Section Method presents the architecture of the system developed. Section Data Preparation explains the preparations made to use the legislation data in the system developed. Section Experiments discusses the experiments conducted and the results obtained in the LLMs-based teacher-student modeling approach. At the end, concluding remarks and future research perspectives are provided in the Conclusions section.

RELATED WORK

AI presents a significant potential for advancing legal analytics across diverse domains. These technologies facilitate the efficient processing and analysis of extensive legal datasets composed by various frameworks, packages, and reports that contain case law, statutes, contracts, and regulatory documents, with high speed and precision [28]. In particular, ML algorithms demonstrate proficiency in identifying patterns and trends within legal decisions, predicting case outcomes, and augmenting legal research capabilities [2]. DL models excel in complex tasks such as natural language processing and document classification while LLMs exhibit remarkable aptitude in understanding and generating human-like text, proving especially valuable for contract analysis, legal drafting, and preliminary legal advice [15, 19]. The integration of these AI technologies in law promises to enhance operational efficiency, reduce costs, and potentially improve access to justice by increasing the accessibility and interpretability of legal information [17].

Homoki and Ződi [11] and Yang et al. [32] stress the fact that LLMs such as ChatGPT are emerging as powerful technologies in the legal domain as they present a significant potential to transform various aspects of legal practice and access to justice. The authors argue that these models already excel in essential legal tasks such as text retrieval, generation, labeling, and classification, demonstrating their capability to process and understand large volumes of legal text data. At the same time, the adaptability of LLMs is evident in their successful application to specific legal settings, such as small law firms. Beyond direct legal operations, LLMs serve as enablers by enhancing knowledge management systems, reducing the need for human intervention in knowledge capture, and acting as middleware between various ICT systems and AI solutions. This versatility positions LLMs to potentially democratize access to legal information and services, making them more efficient and widely available.

At the same time, LLMs face in the legal domains a series of challenges that have the potential to impact their effectiveness and integration into legal practice. These challenges include issues such as the dataset quality, which can lead to inaccuracies in learned information, and algorithmic shortcomings that may hinder the models' understanding of complex legal concepts. Furthermore, safety, security, and privacy attacks such as data poisoning and backdoor attacks (examples of adversarial attacks) [33]. At the same time, addressing ethical concerns like hallucination and false information are paramount, as ensuring data protection and ethical AI use is crucial for maintaining trust in the legal system [16]. At the same time, the widespread adoption of these AI systems in courts may disrupt the traditional hierarchical system of trials, potentially affecting the fairness and consistency of legal judgments. This disruption could challenge established practices of lower court supervision by higher courts, raising concerns about the overall integrity of the judicial process. Addressing these challenges is essential for the successful integration and ethical use of LLMs in legal applications [14, 27, 34].

In the context of criminal law, Jimma [13] stresses that LLMs demonstrate a high potential as advisors particularly in the realm of cybercrime. These models can assist in understanding and interpreting complex legal regulations related to cybercrime and other criminal activities, while also helping to draft suitable laws and policies crucial for effectively combating such crimes. A direct use case is represented by support to developing legal strategies to fight against cybercrime and protect the rights of victims. Herein, LLMs can analyze the legal aspects of cybercrime laws and ensure that the rights of individuals are upheld during legal proceedings. Furthermore, LLMs can collaborate with law enforcement agencies, intelligence, and security organizations to effectively enforce laws related to cybercrime, facilitate collaboration between different entities, and aid in establishing partnerships between law enforcement agencies and private sectors to enhance research, training, and capacity building in forensic investigation technologies. At the same time, LLMs could also have an important role in promoting international cooperation in enforcing cybercrime laws, navigating legal processes such as mutual legal assistance (MLA) requests, extradition, and data preservation across borders, while addressing challenges such as the lack of uniformity in cybercrime laws among states and the slow response to international cooperation requests.

In the context of tax law, Nay et al. [24] argue that LLMs show a real potential as tax law advisors of attorneys by assisting with

various tasks such as contract analysis, case prediction, and tax law inquiry analysis. Specifically, these models can be used alone or integrated with legal texts to provide accurate and accessible legal advice, reducing costs and complexity for individuals navigating the tax law system. Experimental setups show that LLMs' legal understanding capabilities improve with each model release, and reveal the fact that LLMs may predict the impacts of new tax laws or policies by scanning vast legal texts, potentially influencing future lawmaking and necessitating changes in legal services and AI governance regimes.

In the context of public procurement, de Paiva et al. [7] develop a LLMs-based system to identify products in textual descriptions related to the Brazilian government's education programs, addressing the challenges posed by the lack of standardized formats in submitted invoices for financial transfers to municipalities, which hinders the analysis and comparison of purchased items due to diverse product specifications, by enhancing the models' effectiveness in handling and accurately identifying referred products to facilitate expense analysis and accountability of received funds.

Tufis et al. [30] present a Romanian legislative corpus for the development of machine translation systems, containing more than 144k documents representing the legislative body of Romania. Masala et al. [20] present considerable resources obtained by collecting and translating a large corpora of texts, instructions, and benchmarks and train, evaluate, and release open-source LLMs tailored for Romanian language. Masala et al. [21] focus on the legal domain and introduce a Romanian BERT model pre-trained on a large specialized corpus. In a different paper, Masala et al. [20] propose models for legal judgment prediction by specialized and general models for predicting the final ruling of a legal case. The experiments run on four datasets highlighted superior performance for the specialized models and long texts handling.

As the literature review conducted in this research shows, in the field of public procurement law a series of AI applications exist and are developed using classical machine learning and deep learning techniques. Nevertheless, a LLMs-based approach in this direction is lacking. This represents the knowledge gap that this research aims to tackle.

METHOD

Figure 1 displays the training strategy we are following for the dataset preparation and for the fine-tuning of LLMs.
Due to the fact our initial dataset is very small, first we made use of knowledge distillation to generate supplementary synthetic data for training and validation.

In the knowledge distillation stage, we used a version of Llama-3-70B to play the role of a teacher LLM for analyzing and extracting QA pairs. At a subsequent stage, the merged original and synthetic datasets were used to fine-tune ten different LLM models by using recent publicly available LLM bases.

The base models we have used for fine-tuning are intended for commercial and research use. They can be adapted for a variety of code synthesis and understanding tasks.
Except Saul [6] and RoMistral [22], the models we use are quantized in four bits. All LLM models we have used for knowledge distillation and fine-tuning were released in 2024.

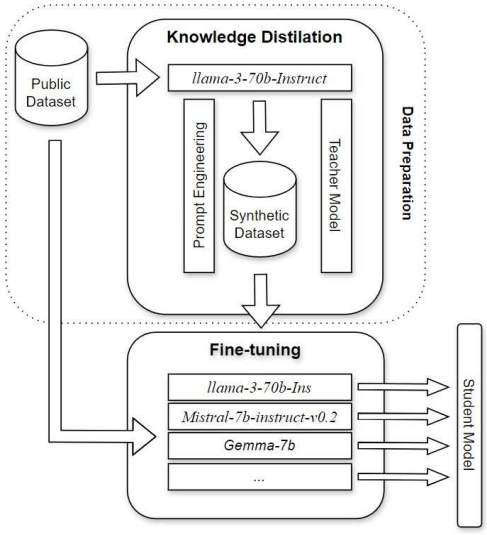

Figure 1. LLM fine-tuning for procurement legislation by knowledge distillation diagram

LLM base models

LLama-3-8B is an LLM model released by Meta in April 2024. We have used Meta-Llama-3-8b-instruct, an instruct fine-tuned version of the base 8B model. Regarding the licensing terms, Llama 3 comes with a permissive license that allows redistribution, fine-tuning, and derivative works. LLaMA-2-13B is an older model version by Meta which was released in January 2024. For our research, we have used Llama 2 13B Bnb 4bit, a 13B model of type Llama with 4K context length and 32K vocabulary size [29]. According to the LLaMA team, LLaMA-13B outperforms GPT-3 (175B) on most benchmarks. This model is quantized on 4 bits and uses 7.2 GB VRAM. The model is available under Apache-2.0 license.

Saul-7B model [6] targets text generation tasks for legal use cases. For our research, we have used Saul 7B Instruct V1, a 7B model of type Mistral released in January 2024 and which has 32K context length and requires 28.9 GB VRAM. The model is available under MIT license.

Mistral-7B [12]*:* For our research, we have used Mistral 7B Instruct V0.2 Bnb 4bit, a 7B model of type Mistral and has 32K context length. This model version is quantized on 4 bits and uses 4.1 GB VRAM. According to the Mistral team, the model outperforms Llama 2 13B on all benchmarks. The model is available under Apache-2.0 license.

Gemma-7B [8]: For our research, we have used Gemma 7B Bnb 4bit, a 7B model of type Gemma and has 8K context length. This model version is quantized on 4 bits and uses 5.6 GB VRAM. The model is available under Apache-2.0 license.

Zephyr [31]*:* For our research, we have used Zephyr Sft Bnb 4bit, a 7B model of type Mistral and has 32K context length. The model is quantized on 4 bits and requires 4.1 GB VRAM. The model is available under Apache-2.0 license.

Qwen2 [3]*:* For our research, we have used three models of type Qwen2. The first is Qwen2 7B Instruct Bnb 4bit, a 7B model with 32K context length and 152K vocabulary size. The model is quantized on 4 bits and requires 5.5 GB VRAM. The model is available under Apache-2.0 license. Next, we also used two lighter versions of type Qwen2, both released in June 2024. The first is Qwen2 1.5B Bnb 4bit. This model has 4 bits quantization, 152KB vocabulary size, has 131KB context length and requires less memory (1.1 GB VRAM). The second is Qwen2 0.5B Bnb 4bit, a similar model with 4 bits quantization, 152KB vocabulary size, 131KB context length, requiring even less memory (0.5 GB VRAM).

RoMistral-7B-Instruct [22] is a model fine-tuned from Mistral-7B-v0.1 by OpenLLM-Ro, the first open-source initiative aiming to build a LLM specialized for Romanian language. The model is available under cc-by-nc-4.0 license.

DATA PREPARATION

To prepare the training and evaluation corpus in Romanian language, we used several sources of procurement-related legislation corpus. First, we targeted the Romanian Government's official corpus of openly discussed cases (Romanian National Agency for Public Procurement - ANAP) [1]. Next, we set up a small dataset of multiple- choice questions from online public learning resources. In addition, we generated a dataset of synthetic QA samples based on the public procurement legislation (Romanian Law 98/2016). All the collected and synthetic collections were curated and assembled to make up training and validation datasets.

Corpus of ANAP openly discussed cases

We collected a small corpus from the procurement text library of the Governmental Agency ANAP covering openly discussed cases. The library is publicly available on the Internet (ANAP).

We filtered out the formerly-discussed cases so that only the valid cases under the current legislation were kept (ANAP corpus: 325 samples). Figure 2 shows a sample of ANAP corpus (ANAP case 828).

In the context of large language models (LLMs), a token is the smallest unit of text that the model processes. It can be a word, part of a word, punctuation mark, or special character. Tokens are used to break down input text and generate output, and they form the basis for the model's understanding and generation of language. The ANAP corpus we have collected includes 320.506 tokens.

Multiple choice tests

As additional data, we made use of online public learning resources to create a very small dataset of 112 multiple- choice questions like the sample in Figure 3.

Law 98/2016

We applied knowledge distillation to generate new QA samples. For that, we used *llama-3-70B-Instruct* as a teacher model. The model was released by Meta has 128K vocabulary size, 8KB context and requires 39.5 GB VRAM.

Prompt engineering for multiple-choice question generation

The teacher model was instructed to generate synthetic multiple-choice corpus. The synthetic multiple-choice corpus we generated contains 871 samples and has 140.858 tokens.

	Romanian	English
Question	*"În cazul în care autoritatea contractantă intenționează să depună două proiecte pentru obținerea finanțării nerambursabile cum se calculează valoarea estimată a contractului de servicii de consultanță pentru întocmirea și pregătirea dosarului cererii de finanțare și depunerea on-line a cererii de finanțare?"*	*"If the contracting authority intends to submit two projects to obtain non-reimbursable financing, how is the estimated value of the consulting services contract for the preparation and preparation of the financing application file and the online submission of the financing application calculated?"*
Answer	*"Alegerea modalității de achiziție a serviciilor de consultanțăpentru întocmirea și pregătirea unui dosar de cerere de finanțare pentru un proiect se realizează în conformitate cu prevederile art. 17 alin. (1) și alin. (2) din Anexa la H.G. nr. 395/2016, prin raportare la valoarea estimată cumulată a serviciilor considerate similare, pe care autoritatea contractantă le-a atribuit sau intenționează să le atribuie pe întregul an bugetar, indiferent dacă celelalte servicii similare vizează proiecte distincte."*	*"The choice of the method of acquisition of consulting services for the drawing up and preparation of a funding application file for a project is carried out in accordance with the provisions of art. 17 para. (1) and para. (2) from the Annex to H.G. no. 395/2016, by referring to the cumulative estimated value of the services considered similar, which the contracting authority assigned or intends to assign for the entire budget year, regardless of whether the other similar services refer to distinct projects."*

Figure 2. Sample from publicly open ANAP corpus on procurement cases (ANAP case 828).

	Romanian	English
Question	*"ANAP este în subordinea: a) Președintelui Curții de Conturi b) Ministrului finanțelor publice c) Guvernului"*	*"ANAP is subordinate to: a) the President of the Court of Accounts b) the Minister of Public Finance c) the Government"*
Answer	*"Răspunsul corect este c"*	*"The correct answer is c"*

Figure 3. Sample of multiple-choice questions which are publicly available online.

Prompt engineering for question-answer generation

Next to the synthetic multiple-choice corpus, we instructed the teacher model (using the prompt in Figure 4) to generate a corpus of 863 synthetic QA samples (95.995 tokens), based on Romanian procurement law 98/2016.

Final Dataset

Based on all the datasets compiled and synthesized at the previous stage, we have made the final training corpus. The training corpus includes 2.171 samples (568.251 tokens). In addition, using the same teacher LLM we generated a separate validation dataset of 400 synthetic samples. Figure 5 shows a synthetic sample of question-answer pairs we obtained.

Romanian	*English*
"Răspunsul tău va fi strict în limba română. Generează între trei și zece întrebări și răspunsuri pentru articolul " + **<id_articol>** + *" din legea 98/2016, în funcție de dimensiunea articolului. Dacă articolul este scurt, generează trei întrebări. Dacă articolul este lung, generează zece întrebări. Pentru fiecare întrebare generată, mai furnizează și câte o variantă alternativă de întrebare. Generează doar întrebări pentru care știi răspunsul. Întrebările și răspunsurile generate trebuie să fie în limba română. Nu prescurta răspunsul absolut deloc."* + <**text_articol**>	*"Your answer will be strictly in Romanian. Generate between three and ten questions and answers for the article "* + ***<article_id>*** + *" from Law 98/2016, depending on the size of the article. If the article is short, generate three questions. If the article is long, it generates ten questions* *It also provides one alternative question. The questions and answers generated must be in Romanian. Do not abbreviate the answer at all."* + <***article_text***> .

Figure 4. Prompt sample used for generating synthetic QA pairs based on articles from law 98/2016.

	Romanian	English
Question	*"Cum se calculează valoarea estimată a achiziției în cazul contractelor de achiziție publică de produse sau de servicii cu caracter de regularitate, referitor la articolul 20 din legea 98/2016?"*	*"How is the estimated value of the purchase calculated in the case of contracts for the public purchase of products or services of a regular nature, referring to Article 20 of Law 98/2016?"*
Answer	*"Valoarea estimată a achiziției se calculează pe baza valorii reale globale a tuturor contractelor similare succesive atribuite în ultimele 12 luni sau în anul bugetar anterior, ajustată pentru a lua în considerare modificările cantitative și valorice care pot surveni într-un interval de 12 luni de la data contractului inițial."*	*"The estimated purchase value is calculated on the basis of the global real value of all successive similar contracts awarded in the last 12 months or in the previous fiscal year, adjusted to take into account the quantitative and value changes that may occur in a 12-month interval from the date of the original contract."*

Figure 5. Example of synthetic question and synthetic answer generated by the teacher model (llama-3-70b-Instruct), given art.20, law 98/2016.

RESULTS AND DISCUSSIONS

For fine-tuning, we have made use of models with 4 bits quantization (with exception of Saul-7B). The training hyperparameters set training steps at 2.000 steps, 8 epochs, with 5 warm-up steps, learning_rate at 2e-4, weights decay at 0.01, linear learning rate scheduler, and 4 gradient accumulation steps. For each training session, the seed parameter was set to the same value.

The validation set included 400 synthetic samples. Also, the validation set did not contain any multiple-choice questions. The evaluation step implied the computation of the value of the loss function only.

The experiments including the training sessions were all carried out in Google Colab[1], using the T4 GPU hardware accelerator.

Figure 6 shows the training loss and the validation loss for RoMistral-7b-Instruct, the best performing fine-tuned LLMs in our study.

Table 1 shows the evaluation loss as well as the training steps of the best fine-tuned LLM models for procurement QA tasks under investigation. The models in Table 1 are presented in the descending order of their evaluation loss, meaning the first model in the table is the best performing and the last model represents the worst performing fine- tuned LLM.

The first four entries in Table 1 relate to the best performing fine-tuned LLM models and are displayed in bold style and green color.

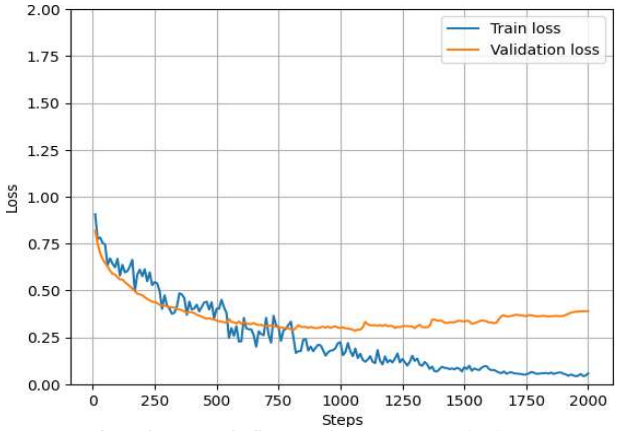

Figure 6. Loss function graph fine-tuning *RoMistral-7b- Instruct*.

From the Llama type, Llama2-13B reveals its superiority over the Llama3-8B, highlighting that in this case a bigger model of an old LLM generation can perform better than a smaller yet newer one.

RoMistral-7B is followed by Zephyr and by the pioneer LLM tailored for the legal domain Saul-7B. The least performant model out of the Mistral class is the base Mistral- 7B LLM.

Considering the models of LLM type Qwen2, the results emphasize the bigger the models the more performant they are. That means Qwen2-7B provides the best results from the three models of Qwen2 type investigated, followed by Qwen2-1.5B and Qwen2-0.5B. The lighter LLM models of type Qwen2 (1.5B and 0.5B) show good support for running such LLMs for procurement on mobile devices such as smartphones. Furtheron, Gemma-7B and Qwen2-7B show comparable performance.

[1] https://colab.research.google.com/

Base model (type)	Performance Evaluation Loss (tr.steps)
RoMistral-7b-Instruct (Mistral)	**0.284 (1060)**
Zephyr-sft (Mistral)	**0.291 (1020)**
Saul-7B-Instruct-v1 (Mistral)	**0.292 (1020)**
Mistral-7b-instruct-v0.2 (Mistral)	**0.294 (1000)**
Llama2-13b (Llama)	0.308 (1080)
Gemma-7b (Gemma)	0.367 (1080)
Qwen2-7B (Qwen2)	0.378 (1080)
Qwen2-1.5B (Qwen2)	0.449 (1350)
Llama3-8b (Llama)	0.452 (1080)
Qwen2-0.5B (Qwen2)	0.531 (1890)

Table 1. Best performing models on Romanian procurement law (ordered by the evaluation loss).

Surprisingly, the much lighter model Qwen2-1.5B shows slightly higher performance than Llama3-8B.

Qualitative Evaluation

Furthermore, a qualitative human evaluation is conducted with two professionals in order to assess the system's efficacy, accuracy, and practical utility in real-world procurement scenarios. This is done to examine the system's ability to properly interpret complex legal language, identify relevant clauses, and provide actionable insights for procurement professionals. In this sense, coherence, consistency, and relevance evaluation criteria are considered and measured as existing studies in this field do, on a scale between 0 and 1 [4, 25, 26].

The human evaluation implied two professionals ranking the text responses generated by four LLMs based on ten different prompts. Figure 7 displays some of the ten questions used in preparing the evaluation prompts.

Romanian	English
Existăposibilitateaînlocuirii unui membru al unei asocieri temporare de operatori economici, căreia i-a fost atribuit un contract/acord– cadru, cu un alt operator economic care îndeplinește criteriile de calificare și selecție stabilite inițial, ca urmare a unei succesiuni cu titlu universal în cadrul unui proces de divizare?	There is the possibility of replacing a member of a temporary association of economic operators, to whom a contract/framework agreement has been assigned, with another economic operator who meets the initially established qualification and selection criteria, as a result of a universal title succession within a division process?
Care ar fi exemple de produse/ servicii/lucrări ce se pot achizitona în cadrul procedurii parteneriatului pentru inovare? Care sunt diferențele de concept între negociere competitiva, dialog competitiv, parteneriat pentru inovare?	What would be examples of products/services/works that can be purchased under the innovation partnership procedure? What are the differences in concept between competitive negotiation, competitive dialogue, and partnership for innovation?
Care este diferența între noțiunea „specificații tehnice", folosită la nivelul legislației primare, și noțiunea „caiet de sarcini", folosită la nivelul legislației secundare?	What is the difference between the notion of "technical specifications", used at the level of primary legislation, and the notion of "specifications", used at the level of secondary legislation?
Este posibilă acceptarea înlocuiriiunui subcontractant (ulterior semnării contractului), cu un alt subcontractant?	Is it possible to accept the replacement of a subcontractor (after signing the contract) with another subcontractor?

Figure 7. Examples of questions used in preparing the prompts for evaluating the first four best fine-tuned LLMs.

The selected LLMs for the human evaluation were the first four best performing models (see Table 1). Each evaluator had to assess the ten responses on each of the three metrics. Figure 8 illustrates the results of the qualitative human evaluation run on the first best four models from Table 1.

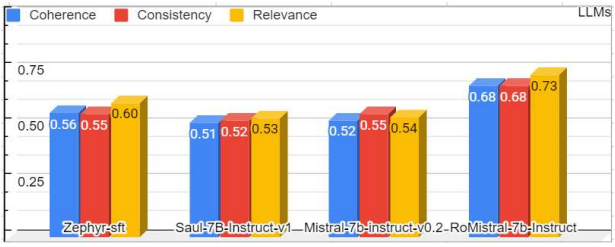

Figure 8. Qualitative evaluation considering coherence, consistency and relevance for the first four best models.

Discussions

One observation is that the first four best performing LLMs judged on the evaluation loss, are all part of the Mistral type of LLMs. Out of the Mistral class, the LLM model specialized for Romanian (RoMistral-7B) shows the best performance. RoMistral-7B represents the best performing model not only by judging on the evaluation loss, but also in the qualitative human evaluation. The two professionals involved in the qualitative evaluation ranked the answers generated by RoMistral-7B as superior on each of the three metrics used. Subsequently, Zephyr shows it is superior on relevance and on coherence, when compared to Saul-7B and Mistral-7B. Based on the qualitative human evaluation, the pioneer LLM tailored for the legal domain Saul-7B ranks the lowest on coherence, consistency and relevance.

These are the preliminary findings of our explorative study. More in-depth investigation is required to fully understand the limitations and performance of the fine-tuned LLMs applied for Romanian procurement subdomain.

CONCLUSIONS

In this paper, we showed and discussed the preliminary results from our research aiming at creating legislature specialized supportive LLMs for question-answering with immediate application to the procurement subdomain. We have detailed the necessary steps including the data collection and preparation, prompt engineering part, modeling and testing.

As part of the data collection, we applied knowledge distillation by using Llama-70b as a teacher model to generate samples for training much smaller models under exploration. Further, as the system proposed represents an explorative single-model solution, we provided a comparative approach

benchmarking student models using various base models such as RoMistral-7B, Llama3-8b, Mistral-7b, Gemma-7b and Saul-7b.

The main drawback of our research consists in the very limited size of the training dataset used for fine-tuning the procurement legislature models. This implies potential risks in relation to aspects such as capability of generalization, incomplete coverage, and robustness. To address these, s future work, we aim at scaling up the text collection, synthesis and curation of the training corpus. To increase the training dataset, we plan to extend the application of knowledge distillation to get many more samples related to Romanian law 98/2016; in addition, we plan to add samples related to Romanian hg. 395/2016. Further on, we will fine-tune larger LLM models and check their performance on question-answering tasks targeting the Romanian procurement legislature. At the same time, ethical considerations such as bias analysis and interpretability enhancement of the models used need to be further developed in relation to the underlying legal perspective and implications.

Eventually, our intention is to make available the best performing models as a public service (SaaS architecture) readily accessible for all the users from public and private sectors from Romania as well as from other European countries. We consider such a service would significantly facilitate easy access to the Romanian procurement legislation for all the parties concerned on applying and following the Romanian law for procurement operations.

REFERENCES

1. ANAP Governmental Agency's procurement corpus, https://achizitiipublice.gov.ro/questions/view.

2. Ashley, K. D. Prospects for legal analytics: some approaches to extracting more meaning from legal texts. *University of Cincinnati Law Review*, *90*(4), 5, (2022).

3. Bai, J. et. al, "*Qwen Technical Report"*, arXiv preprint arXiv:2309.16609, doi: 10.48550/arXiv.2309.16609, (2023).

4. Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., ... and Testoni, A. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. *arXiv preprint arXiv:2406.18403,* (2024).

5. Bosio, E., Djankov, S., Glaeser, E., and Shleifer, A. Public procurement in law and practice. *American Economic Review*, *112*(4), 1091-1117, (2022).

6. Colombo, P. et al., *SaulLM-7B: A pioneering Large Language Model for Law*, doi: 10.48550/arXiv. 2403.03883, (2024).

7. de Paiva, E. S. et al. *Continued pre-training of LLMs for Portuguese and Government domain: A proposal for product identification in textual purchase descriptions.* In *AAAI-2024 Workshop on Public Sector LLMs: Algorithmic and Sociotechnical Design*, (2024).

8. Gemma Team et. al, "*Gemma: Open Models Based on Gemini Research and Technology*", doi: 10.48550/arXiv.2403.08295, (2024).

9. Graells, A. S. Public procurement and competition: some challenges arising from recent developments in EU public procurement law. *Research Handbook on EU Public Procurement Law*, 423-451, (2016).

10. Harland, C., Telgen, J., Knight, L., and Callender, G. Challenges facing public procurement. In *Public Procurement*, Routledge, (2012), 379-386.

11. Homoki, P. and Ződi, Z. Large Language Models and Their Possible Uses in Law. 10.13140/RG.2.2.14315 .41764, (2023).

12. Jiang, A. Q. et al., "*Mistral 7B"*, doi: 10.48550/arXiv.2310. 06825, (2023).

13. Jimma, E. *College of Law and Governance School of Law LLM in Human Rights and Criminal Law* (Doctoral dissertation, Jimma University), (2022).

14. Lai, J., Gan, W., Wu, J., Qi, Z., and Yu, P. S. Large language models in law: A survey. *arXiv preprint arXiv:2312.03718*, (2023).

15. Limsopatham, N. Effectively leveraging BERT for legal document classification. In *Proceedings of the Natural Legal Language Processing Workshop 2021,* (2021), 210-216.

16. Lin, Z., Guan, S., Zhang, W., Zhang, H., Li, Y., and Zhang, H. Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review*, *57*(9), (2024), 1-50.

17. Linna, D. W. Evaluating Artificial Intelligence for Legal Services: Can "Soft Law" Lead to Enforceable Standards for Effectiveness?. *IEEE Technology and Society Magazine*, *40*(4), (2021), 37-51.

18. Mackey, T. K., and Cuomo, R. E. An interdisciplinary review of digital technologies to facilitate anti-corruption, transparency and accountability in medicines procurement. *Global health action*, *13*(sup1), 1695241, (2020).

19. Mamakas, D., Tsotsi, P., Androutsopoulos, I., and Chalkidis, I. Processing long legal documents with pre- trained transformers: Modding legalbert and longformer. *arXiv preprint arXiv:2211.00974*, (2022).

20. Masala, M., Rebedea, T., and Velicu, H. Improving Legal Judgment Prediction in Romanian with Long Text Encoders, 2402.19170, (2024).

21. Masala, M., Iacob, R., Uban, A., S., Cidota, M.,

Velicu, H., Rebedea, T., and Popescu, M. jurBERT: A Romanian BERT Model for Legal Judgement Prediction, Natural Legal Language Processing Workshop 2021, Association for Computational Linguistics, (2021), 86–94.

22. Masala, M., Ilie-Ablachim, D. C., Dima, A., Corlatescu, D., Zavelca, M., Olaru, O., Terian, S., Terian, A., Leordeanu, M., Velicu, H., Popescu, M., Dascalu, M., and Rebedea, T. "Vorbeşti Românește?" A Recipe to Train Powerful Romanian LLMs with English, doi: 10.48550/arXiv.2406.18266, (2024).

23. Nagitta, P. O., Mugurusi, G., Obicci, P. A., and Awuor,
E. Human-centered artificial intelligence for the public sector: The gate keeping role of the public procurement professional. *Procedia Computer Science*, *200*, (2022), 1084-1092.

24. Nay, J. J., et al. Large language models as tax attorneys: a case study in legal capabilities emergence. *Philosophical Transactions of the Royal Society A*, *382*(2270), 20230159, (2024).

25. Pan, Q., Ashktorab, Z., Desmond, M., Cooper, M. S., Johnson, J., Nair, R., ... and Geyer, W. Human-Centered Design Recommendations for LLM-as-a-Judge. *arXiv preprint arXiv:2407.03479*, (2024).

26. Shankar, S., Zamfirescu-Pereira, J. D., Hartmann, B., Parameswaran, A. G., and Arawjo, I. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. *arXiv preprint arXiv:2404.12272*, (2024).

27. Sun, Z. A short survey of viewing large language models in legal aspects. *arXiv preprint arXiv:2303.09136*, (2023).

28. Surden, H. Artificial intelligence and law: An overview.
Georgia State University Law Review, *35*, (2019), 19-22.

29. Touvron, H. et al., "*LLaMA: Open and Efficient Foundation Language Models*", doi: 10.48550/arXiv.2302.13971, (2023).

30. Tufiș, D., Mitrofan, M., Păiș, V., Ion, R., and Coman, A. Collection and Annotation of the Romanian Legal Corpus, *Proceedings of the 12th Conference on Language Resources and Evaluation* (LREC 2020), (2020), 2773–2777.

31. Tunstall, L. et. al, *Zephyr: Direct Distillation of LM Alignment*, doi: 10.48550/arXiv.2310.16944, (2023).

32. Yang, X., Wang, Z., Wang, Q., Wei, K., Zhang, K., & Shi, J. Large language models for automated q&a involving legal documents: a survey on algorithms, frameworks and applications. *International Journal of Web Information Systems*, (2024).

33. Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., and Zhang, Y. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 100211, (2024).

34. Zhang, R., Li, H., Wu, Y., Ai, Q., Liu, Y., Zhang, M., and Ma, S. Evaluation Ethics of LLMs in Legal Domain. arXiv preprint arXiv:2403.11152, (2024)