LAR-ECHR: A New Legal Argument Reasoning Task and Dataset for Cases of the European Court of Human Rights

Odysseas S. Chlapanis[1,3], **Dimitrios Galanis**[2,3] and **Ion Androutsopoulos**[1,3]
[1]Department of Informatics, Athens University of Economics and Business, Greece
[2]Institute for Language and Speech Processing, Athena Research Center, Greece
[3]Archimedes Unit, Athena Research Center, Greece
odyhlapanis@aueb.gr, galanisd@athenarc.gr, ion@aueb.gr

Abstract

We present Legal Argument Reasoning (LAR), a novel task designed to evaluate the legal reasoning capabilities of Large Language Models (LLMs). The task requires selecting the correct next statement (from multiple choice options) in a chain of legal arguments from court proceedings, given the facts of the case. We constructed a dataset (LAR-ECHR) for this task using cases from the European Court of Human Rights (ECHR). We evaluated seven general-purpose LLMs on LAR-ECHR and found that (a) the ranking of the models is aligned with that of LegalBench, an established US-based legal reasoning benchmark, even though LAR-ECHR is based on EU law, (b) LAR-ECHR distinguishes top models more clearly, compared to LegalBench, (c) even the best model (`GPT-4o`) obtains 75.8% accuracy on LAR-ECHR, indicating significant potential for further model improvement. The process followed to construct LAR-ECHR can be replicated with cases from other legal systems.

Facts:
Applicant married B, had daughter C.
B abused applicant due to psychiatric disorders.
B arrested, released; applicant moved to shelter.
Marriage dissolved; B continued harassment.
Preceding arguments:
Government claimed it has established legal protections for domestic violence victims. The court noted B's repeated violence. All incidents concerned the same perpetrator and occurred in a continual manner, so the Court will examine them as a continuous situation.
Continuation Options:
(A) Psychiatric reports indicating mental disorders, tendency towards violence ... ✔
(B) Applicant's confinement in mental hospital ... ✘
(C) Police collected information about applicant ... ✘
(D) Complaint about psychiatric examination, ... ✘

Table 1: A cropped instance from LAR-ECHR.

1 Introduction

The rise of Large Language Models (LLMs) has impacted several sectors, including the legal one. In the United States, LLMs are being integrated into legal research and writing tools designed for both professionals and laypeople.[1] These advances are largely due to the effort of legal experts who contributed significantly in dataset development and manual evaluation (Guha et al., 2023; Magesh et al., 2024). Their involvement, however, is extremely costly, hence methods to construct and evaluate legal benchmarks semi-automatically are required.

LegalBench (Guha et al., 2023) is an example of a widely used legal reasoning benchmark. It consists of data for 162 tasks, hand-crafted by legal experts, that evaluate six types of legal reasoning of the US legal system, making it the most reliable dataset of this kind. MMLU-Law, a subset of MMLU (Hendrycks et al., 2021) that contains three US legal tasks only, is also commonly used. Similarly, LawBench (Fei et al., 2023) and IL-TUR (Joshi et al., 2024) were created for the Chinese and Indian regions; they comprise 20 and 8 tasks, respectively. For other legal systems, at least two other large multi-task legal benchmarks have been made available (Chalkidis et al., 2022; Niklaus et al., 2023). However, they include mostly text classification tasks that do not require understanding or generating chains of legal arguments (e.g., court arguments explaining the decisions of judges) and can be solved reasonably well using smaller BERT-based models (Chalkidis et al., 2020) or even linear classifiers (Aletras et al., 2016). Hence, it is questionable if they test legal reasoning abilities.

Motivated by the observations above, we introduce a novel task (§2), Legal Argument Reasoning (LAR), designed to evaluate the legal reasoning skills of LLMs. The task requires selecting the cor-

[1] https://www.americanbar.org/groups/law_practice/resources/tech-report/2023/2023-artificial-intelligence-ai-techreport/

Proceedings of the Natural Legal Language Processing Workshop 2024, pages 267–279
November 16, 2024

rect next statement (from multiple choice options) in a chain of legal arguments from court proceedings, given the facts of the case. We have prepared a LAR dataset for EU law, LAR-ECHR (§3), using court arguments from the European Court of Human Rights (ECHR). It contains 403 instances; Table 1 shows a cropped example. Notably, the same process could be applied to construct LAR datasets for other legal systems as well.

We evaluate two closed-weight and 5 open-weight LLMs on LAR-ECHR (§4), comparing their performance and rankings against two legal benchmarks: LegalBench and MMLU-Law. For completeness, we also report scores on two general reasoning benchmarks: the entire MMLU (MMLU-Full) and HellaSwag (Zellers et al., 2019). We find that: (a) the ranking of the models on LAR-ECHR is aligned with that of LegalBench, even though the two benchmarks are from different legal systems (US, EU); (b) LAR-ECHR provides clearer distinctions among top models, compared to LegalBench; (c) even the best model (`GPT-4o`) obtains only 75.8% accuracy on LAR-ECHR, similar to the top accuracy on LegalBench (73.3%), indicating significant potential for further model improvement.

Our contributions are three-fold: (i) we introduce a novel task, Legal Argument Reasoning (LAR), to evaluate the legal reasoning abilities of state-of-the-art LLMs; (ii) we create and publicly release LAR-ECHR, a challenging EU-specific LAR dataset; (iii) we present a method to create LAR datasets for any other legal system using court proceedings with annotated arguments.

2 The LAR task

We introduce Legal Argument Reasoning (LAR), a novel task to evaluate the legal reasoning abilities of LLMs. The LLM is fed with the *facts* of the case (a list of sentences summarizing the events considered, see Table 1), a sequence of *preceding arguments* (statements) from the court proceedings, and *continuation options* (the correct next statement and distractors). The LLM has to select the correct next statement. (In court proceedings, 'arguments' are numbered statements documenting the legal reasoning of the court.)

LAR requires various types of legal and commonsense reasoning that extend beyond simple pattern recognition or memorization. As demonstrated by HellaSwag (Zellers et al., 2019), even predicting the next sentence in a generic corpus can be challenging and requires reasoning skills. In the legal domain, the complex terminology and inferences make the task of predicting a legal professional's next argument significantly harder (§4). The arguments embody the court's legal reasoning for its decisions. As Medvedeva and Mcbride (2023) state: "[J]udges usually offer explanations which serve to justify their decisions with reference to the facts found to be established and the relevant law. [These explanations] involve the exercise of legal reasoning". Merely understanding the legal terminology is insufficient, as the the facts and the relevant law must also be considered.

3 The LAR-ECHR dataset

The LAR-ECHR dataset contains arguments from the proceedings of the European Court of Human Rights (ECHR). An ECHR court decision typically begins with the facts of the case, followed by the 'Law' section, where the arguments of the parties and the court are presented, followed by the court's conclusion (e.g., verdict, fines). To create the dataset, we used statements from 'Law'.

To ensure that LAR-ECHR is challenging and effectively evaluates legal reasoning, we select appropriate arguments based on criteria derived from the annotations of the Legal Argument Mining ECHR (LAM:ECHR) dataset (Habernal et al., 2024) (§3.1). The criteria we use are described in §3.2 below. Instead of generating the distractors using a language model, as in HellaSwag (Zellers et al., 2019) and LegalLens (Bernsohn et al., 2024), we draw them from similar ECHR cases using an algorithm we developed (§3.3), to avoid introducing language model biases and hallucinations.

LAR-ECHR is based on 191 ECHR court cases. From the 191 cases, we derive 403 samples (like the example of Table 1), which we split randomly into three sets: 5 samples for few-shot prompts, 98 samples for development, 300 samples for testing. In our experiments, we use only the test set, but we release the full dataset for future research.[2] Below we describe in detail how the dataset was created.

3.1 The previous LAM:ECHR dataset

In the aforementioned LAM:ECHR dataset, the arguments of 373 ECHR court decisions were annotated for legal argument mining purposes. The

[2] https://huggingface.co/datasets/AUEB-NLP/lar-echr

cases pertain to alleged violations of Article 8 ('Respect for private and family life') and, to a lesser extent, Article 7 ('No punishment without law') of the European Convention on Human Rights. The arguments were annotated for the actors stating them ('ECHR', 'Applicant', 'State', 'Third Parties', 'Commission/Chamber') and the type of argument (16 types). The argument types are: 'Procedural', 'Interpretation' (five variations), 'Principle of proportionality' (four variations), 'Institutional' (three variations), 'Precedents', 'Decision', 'Application to the concrete case'. The latter type is the most common (57%), and we use only arguments of this type in the new LAR-ECHR dataset (§3.2).

There are 9,950 arguments (65%) labeled with the 'ECHR' actor in LAM:ECHR, 2,471 (16%) arguments labeled with 'Applicant', 2,399 (16%) labeled with 'State'. Only the remaining 385 (3%) arguments are associated with the last two actors; for simplicity, we discard these 385 statements.

The facts of each case are not included in LAM:ECHR, but they are included in the ECtHR B dataset (Chalkidis et al., 2021), which does not provide arguments. We unified the two datasets using regular expressions. A further complication is that LAM:ECHR was published after ECtHR B. The 373 cases of LAM:ECHR include 94 cases that are not covered by ECtHR B. Consequently, we used only the 279 cases covered by both datasets. Recently, a new dataset, ECtHR-PCR (T.y.s.s. et al., 2024), which contains both facts and arguments of ECHR cases (even the most recent ones), was released. Using this dataset, the missing cases will be included in an update of LAR-ECHR in the future.

From the remaining 279 cases, we selected the most appropriate *target arguments* (correct next statements) according to criteria described in §3.2 below. Some cases included multiple arguments that satisfied the criteria, while others none (88 cases). Consequently, we selected target arguments from the remaining 191 cases. The distractor arguments (incorrect next statements) were also selected from the 191 cases (§3.3). This process led to 403 instances, like the one of Table 1.

3.2 Selection of target arguments

Here we describe the process used to select the target arguments (correct next statements) of the new dataset from the 191 cases of §3.1.

As already noted, the 'Law' section of each case contains the arguments of the parties and the court. Actually, a case usually examines multiple *issues* and the 'Law' section contains the arguments of the parties, followed by the arguments of the court, separately per issue. For each issue, the arguments of the parties (in the court proceedings) are actually also written by the judges, in a way that supports the reasoning of the judges. Therefore, for each issue, the arguments (statements) of both the parties and the court actually form a reasoning chain. From that chain, we wish to focus on the arguments of the judges, especially those annotated as 'Application to the concrete case' in LAM:ECHR, which are the most demanding in terms of reasoning, as they consider and combine the arguments of the parties, the law, and the facts of the particular case. Those arguments are "concerned with determining the relation between the concrete case and the abstract legal norm by the subsumption of the facts of a case under a legal norm" (Habernal et al., 2024). In other words, they are parts of the reasoning that the judges follow to connect the law to the facts by 'subsumption', i.e., checking if the facts meet the conditions specified by the law.

Furthermore, in our experience, among the arguments of the judges, the first one (per issue) is the most difficult to predict; we leave an experimental validation of this claim for future work (§7). Therefore, we select as target arguments those that satisfy the following criteria: (i) the argument must be annotated as 'ECHR' (argument of the judges), (ii) the argument must be annotated as 'Application to the concrete case', and (iii) it must be the first one (per issue) after the arguments of the parties.

Due to the limited context length of LLMs, in LAR-ECHR the facts of each case are summarized (using `GPT-4o`) and only the last three of the arguments preceding the target one are retained.

3.3 Selection of distractors

Distractors are incorrect next statements, as opposed to the target argument, which is the correct one. Some studies use synthetic distractors generated by LLMs, e.g., HellaSwag (Zellers et al., 2019) and Legalens (Bernsohn et al., 2024). We opt to use arguments from the same dataset as distractors, following the approach in EntailmentBank (Dalvi et al., 2021). This approach avoids the introduction of biases and hallucinations of LLM generators, as reported in the work of HellaSwag.

The most suitable distractors are algorithmically selected. The algorithm adheres to the following desiderata. (a) The distractors must be *similar* to the target argument, i.e., they must have roughly

Text	Score
Target argument	
The Court notes that this complaint is not manifestly ill-founded within the meaning of Article 35 §3 ...	1.00
Candidate distractors	
The Court notes that *the application* is not manifestly ill-founded within the meaning of Article 35 §3 ...	0.95
The Court, having examined those complaints under Articles 5 §1 and 6 §1 of the Convention ...	0.85
The Court considers that this part of the application raises questions of law which are important ...	**0.79**
The Court notes that the Government put forward reasons for this complaint to be declared inadmissible ...	0.73

Table 2: Exploring the effect of the cosine similarity threshold τ on the candidate distractors.

the same style, length, and vocabulary. As seen in the example in Table 1, the target (correct) argument refers to an event about 'psychiatric reports', 'mental disorder', and 'tendency towards violence'; each one of the distractors mentions relevant terms ('mental hospital', 'police', 'psychiatric examination'). However, (b) the distractors should not be near-duplicates or paraphrases of the target argument or another distractor. With these desiderata, we developed the following algorithm.

Distractor selection algorithm: For each target argument, the candidate distractors are the target arguments of the other cases (of all the issues of the other 190 cases, §3.1). For each candidate distractor, its embedding is computed using an LLM.[3] The candidate distractors are then ranked based on their cosine similarity to the embedding of the target argument (desideratum (a)), from highest to lowest. While the top-3 ranked candidates could present the greatest challenge, they may also be paraphrases of the target argument or another distractor (desideratum (b)). Hence, before selecting the top-3 ranked candidate distractors, we discard candidate distractors whose similarity to the target argument or a more highly ranked candidate distractor exceeds a threshold τ.

Cosine similarity threshold: To select the τ threshold, we conducted the following experiment: for each one of a few target arguments of the development subset (§4.1), we ranked the candidate distractors as above, and manually inspected the texts of the target and the distractors and their similarity scores (see Table 2 for an example). We observed that for similarity scores above 0.9, the two texts were almost identical. For scores between 0.9 and 0.85, they shared several words. For lower similarity scores, no such issues were visible, so we set $\tau = 0.8$.

4 Experiments

4.1 Experimental setup

We evaluate the reasoning skills of seven general-purpose LLMs using the respective web APIs and three random seeds. We employed closed-weight OpenAI models (GPT family), namely `gpt-4o` (L), `gpt-4o-mini` (S) (OpenAI et al., 2024)[4]; open-weight models by Mistral (Mistral family), namely `open-mixtral-8x22b` (L), `open-mixtral-8x7b` (M), `open-mistral-7b` (S) (Jiang et al., 2024)[5]; and open-weight models by Meta (Llama family), namely `llama-3.1-70b` (L), `llama-3.1-8b` (S) (Dubey et al., 2024)[6] .

We report the average classification accuracy (over the three random seeds) and the standard deviation for each LLM on the test subset of LAR-ECHR. We also show results on two previous legal benchmarks (LegalBench, MMLU-Law) and two general benchmarks (MMLU-full, HellaSwag), as previously reported (Liang et al., 2023).

MMLU (Hendrycks et al., 2021) is the most widely used benchmark for evaluating the knowledge and reasoning abilities of instruction following LLMs (Liang et al., 2023). MMLU-Law is a subset of MMLU that contains three legal tasks ('International Law', 'Jurisprudence', 'Professional Law'). LegalBench is the largest (in terms of tasks) benchmark for the evaluation of legal reasoning (Magesh et al., 2024). It includes 162 tasks that assess 6 different reasoning types. HellaSwag (Zellers et al., 2019) is a dataset created automatically that only contains the next statement prediction task, similar to LAR-ECHR. However, in HellaSwag the texts are collected from online articles and not chains of legal arguments, as in LAR-ECHR. In the three previous benchmarks that have multiple tasks (MMLU, MMLU-Law, LegalBench), we report macro-average over their tasks.

[3]We use `openai-embed-small` (https://openai.com/index/introducing-text-and-code-embeddings/) (Neelakantan et al., 2022).

[4]https://openai.com/index/hello-gpt-4o/

[5]https://mistral.ai/news/mixtral-8x22b/

[6]https://ai.meta.com/blog/meta-llama-3-1/

Models	LAR-ECHR (Ours)	Legal Bench*	MMLU Law*	MMLU Full*	Hella Swag
GPT-4o (L)	**75.8** ± 1.8 [**1**]	**73.3** [**1**]	**85.2** [**1**]	**74.8** [**1**]	**89.1** [**1**]
GPT-4o-mini (S)	61.6 ± 2.2 [4]	65.3 [4]	79.6 [2]	66.8 [4]	83.4 [3]
Mistral-8x22B (L)	69.8 ± 1.3 [2]	70.8 [2]	79.1 [3]	70.1 [3]	79.6 [4]
Mistral-8x7B (M)	57.2 ± 1.6 [5]	63.0 [5]	74.3 [4]	64.9 [5]	70.5 [5]
Mistral-7B (S)	49.6 ± 1.9 [7]	33.1 [7]	63.2 [6]	58.4 [6]	60.7 [7]
Llama-3.1-70B (L)	67.2 ± 2.6 [3]	68.7 [3]	67.4 [5]	70.9 [2]	86.2 [2]
Llama-3.1-8B (S)	54.1 ± 1.6 [6]	34.2 [6]	57.3 [7]	50.0 [7]	68.0 [6]

Table 3: Comparison of LLMs from three families on LAR-ECHR, LegalBench, MMLU-Law, MMLU-Full. L, M, S denote the largest, medium, smallest models per family, respectively. Accuracy in %. Rankings in square brackets. LAR-ECHR results averaged over three random seeds. HellaSwag results on a subset of 1000 samples and one random seed. *Results from Liang et al. (2023).

4.2 Zero-shot prompting

All experiments are performed in a zero-shot setting with a Chain-of-Thought (CoT) prompt (Wei et al., 2023) we designed for LAR-ECHR (Table 4). The prompt explains the provided input (facts, preceding arguments, continuation options) and the task, asking the LLM to generate an output in three steps. In the first step ('Analysis'), the LLM reflects on the plausibility of each option (candidate next argument). In the second step ('Explanation'), the LLM explains its choice. In the last step ('Answer'), the LLM outputs only the letter (A, B, C, D) of its choice (to facilitate answer collection).

4.3 Experimental results

Table 3 presents the accuracy of the models on the five datasets (LAR-ECHR, LegalBench, MMLU-Law, MMLU-Full, HellaSwag). The large version of `GPT-4o` (L) achieves the best performance on all datasets, with accuracy 75.8% on LAR-ECHR, showing that there is room for model improvement on our dataset. The rankings (in square brackets) of all models on LAR-ECHR are identical to those of LegalBench, even though the two datasets are based on different legal systems (US, EU). Within each family of models, the largest model is the best on all datasets, as expected.

`Mistral` (L) is the second best model on LAR-ECHR and LegalBench, with a larger performance gap from the best model (6 pp), compared to the corresponding gap on LegalBench (2.5 pp), showing that LAR-ECHR distinguishes better than LegalBench the legal reasoning abilities of the top two models. When comparing models of the same family, LAR-ECHR also distinguishes better between `GPT-4o` (L) and `GPT-4o` (S), as shown in Fig. 1, and the same applies between `Mistral` (L) and `Mistral` (M). By contrast, LegalBench distinguishes substantially better between `Mistral` (L) and `Mistral` (S) (Fig. 1), as well as between `Llama`

Prompt:
You will be provided with the introductory Facts in a European Court of Human Rights (ECHR) case, an excerpt of arguments from that case and several possible continuations of these arguments. Your task is to determine which continuation accurately extends the original argument.
Output format:
- Analysis: Provide an in-depth evaluation of how each option relates to the initial argument.
- Explanation: Articulate why your chosen option is the correct continuation.
- Answer: State the letter of your selected answer.
Facts: . . .
Preceding arguments: . . .
Continuation options: . . .
LLM response:

```
Analysis:
(A): Reinforces B's mental health issues
and need for treatment.
(B): Discusses mental hospital confinement,
unrelated to specific case facts.
(C): Introduces irrelevant information
about police actions and examinations.
(D): Diverges to complaints about forced
evaluations, not B's violence.
Explanation:
Option (A) criticizes Croatian authorities
for not acting on B's known psychiatric
issues and violent tendencies, aligning
with the Court's view of ongoing abuse.
Answer: (A)
```

Table 4: Our CoT prompt for LAR-ECHR, a sample input (facts, preceding arguments, continuation options), and the response (analysis, explanation, answer) from `GPT-4o`. The '...' are as in Table 1. The full version of this table can be found in Appendix A.

(L) and `Llama` (S); this is due to the much lower scores the smaller `Mistral` (S) and `Llama` (S) obtain on LegalBench compared to LAR-ECHR.

Table 5 presents the performance of `GPT` and `Llama` models on LAR-ECHR when provided with the original, complete facts of the proceedings, for models with large enough context length. As expected, all models exhibit greater performance when using the complete facts. Notably, `Llama` models benefit more than `GPT` models. Differences between models within the same family are relatively small. These findings suggest that while summaries offer an effective workaround for models with limited context lengths, they can introduce bias, potentially favoring certain models.

5 Related work

The LAR task was inspired by the continuation task introduced by SWAG (Zellers et al., 2018) and later improved by HellaSwag (Zellers et al., 2019). It is a multiple-choice task where the model has to select the most likely continuation of an event description, such as "A woman sits at a piano" is followed by "She sets her fingers on the keys". The corpus is collected from various online sources such as wikiHow[7]. Similarly to LAR, HellaSwag is constructed automatically, via a technique called Adversarial Filtering (AF) which selects the most persuasive LLM-generated continuations as incorrect options. It is shown empirically that accurately predicting the correct continuation of an event in HellaSwag requires skills that are closely related to commonsense reasoning. The primary differences with our work, aside from our focus on the legal domain, are: (a) we employ official content from court proceedings instead of events from online articles of varying credibility, (b) we use (based on the respective annotations) the most appropriate chain of arguments, and (c) we utilize human-generated challenging distractors.

Our dataset builds on top of two previous works: LAM:ECHR (Habernal et al., 2024) and ECtHR B (Chalkidis et al., 2021). LAM:ECHR annotated, with the help of legal experts, the arguments of 373 ECHR decisions with *actor* and *argument type* labels, and trained and evaluated their RoBERTa-based models on both tasks. In ECtHR B the goal is to predict the articles of ECHR that were allegedly violated, given the facts of the case. To create LAR-ECHR we aligned the common instances of

[7]https://www.wikihow.com/

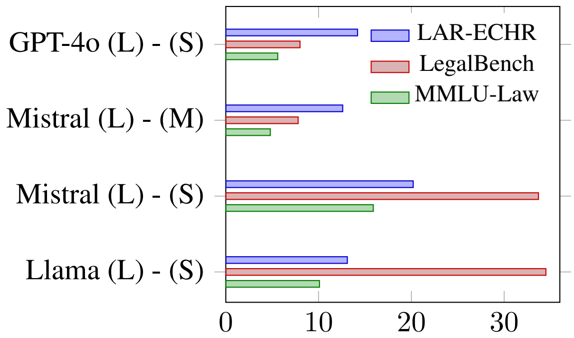

Figure 1: Performance gap within the same LLM family.

Models	LAR-ECHR (complete facts)
GPT-4o	**77.9** (+2.1)
GPT-4o-mini	64.3 (+2.7)
Llama-70B	73.3 (+6.1)
Llama-8B	58.0 (+3.9)

Table 5: Results on LAR-ECHR with complete facts. The difference in performance from the summarized version is shown in parentheses.

these datasets to combine the annotated arguments of LAM:ECHR with the facts of the cases from ECtHR B.

One of the most widely known benchmarks in legal NLP is LexGLUE (Chalkidis et al., 2022). It was one of the first large-scale collection of datasets dedicated to the legal domain. Its creation was inspired by the success of GLUE, a multi-task benchmark dataset (Wang et al., 2018), and the subsequent and more challening SuperGLUE (Wang et al., 2019). LexGLUE includes a variety of (English-only) classification tasks from both US and EU legal systems, however it does not contain any reasoning-specific tasks. LEXTREME (Niklaus et al., 2023) followed with a collection of 11 datasets, featuring tasks similar to those in LexGLUE, to establish a multilingual legal NLP benchmark. LegalLens (Bernsohn et al., 2024) introduced two classification tasks: detecting legal violations and identifying potentially affected individuals. The tasks were created using LLMs and then validated by human experts. The aforementioned benchmarks focus on specific classification tasks. They do not directly measure in-context learning capabilities or the understanding of legal reasoning explanations.

In the broader NLP landscape, several datasets have recently emerged for evaluating the few-shot learning capabilities and advanced reasoning skills of LLMs, replacing GLUE and SuperGLUE as the most widely used benchmarks. These new benchmarks are more aligned with the skills required by chatbot assistants designed to solve a wide range of tasks by following instructions, primarily through generating text rather than predefined labels. The most prominent of these is the Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2021), which is preferred for evaluating the knowledge and general capabilities of LLMs (Liang et al., 2023). It is a multiple-choice dataset that covers 57 tasks across diverse academic subjects, three of them being law-specific. ARC (Clark et al., 2018) is another multiple-choice question-answering dataset that includes science questions requiring various types of reasoning. Big Bench (Srivastava et al., 2023) is a challenging dataset of 204 tasks that focuses on various topics among them arithmetic, logical, common-sense and algorithmic reasoning as well as language understanding and world knowledge.

Inspired by the success of these benchmarks, several benchmarks for the legal domain, with the same orientation, were also made available. For example, the largest legal reasoning benchmark (in terms of number of tasks) is LegalBench (Guha et al., 2023), comprising 162 tasks that cover six different types of legal reasoning and focus on the US legal system. Bongard et al. (2022) created a challenging legal reasoning dataset by adapting questions from a textbook on US civil procedure which however is cast as a binary classification task and does not focus on continuations such as LAR-ECHR. A few datasets that are focused on other legal systems than the US were also made available. For example, LawBench (Fei et al., 2023) consists of 20 tasks on Chinese law that evaluate legal knowledge understanding of LLMs. IL-TUR (Joshi et al., 2024) covers a wide range of multilingual legal text understanding and reasoning tasks for English and 9 Indian languages. Our dataset, LAR-ECHR, differs from the datasets in these benchmarks in that (a) it uses the legal reasoning chain of the arguments of the judges, (b) it refers to EU law and (c) instances are collected semi-automatically from court proceedings using annotations, not handcrafted by legal experts.

6 Conclusion

In this study, we introduced LAR, a legal reasoning NLP task that requires selecting the correct next argument made by judges in a case. We constructed a dataset for this task, called LAR-ECHR, using cases from ECHR. We evaluated seven general-purpose LLMs from three families on this dataset. The best model obtained 75.8% accuracy, indicating significant potential for further model improvement. Model rankings were identical with those of LegalBench, even though the datasets are based on different legal systems. Despite that weak models obtained a substantially lower score in LegalBench, LAR-ECHR distinguished the top models more clearly. The process followed to construct LAR-ECHR can be replicated with cases from any court proceedings, even from different legal systems.

7 Future work

The semi-automatic creation of a LAR dataset requires a few design decisions, two of which we believe are most worth investigating further: (a) the impact of not selecting only the first arguments of the judges (per issue) as target arguments (which in our experience are the most difficult to predict) and (b) the impact of the similarity threshold τ in selecting candidate distractors.

Additionally, we plan to extend the dataset in various directions: (a) collect and align the missing ECHR cases that are annotated from LAM:ECHR, but they do not exist in ECtHR B, (b) include the rest of the articles of ECHR, apart from articles 7 and 8, to cover other domains of legal expertise, (c) annotate more cases to increase the dataset size. These extensions could lead to the inclusion of a training set for fine-tuning LLMs. These LLMs would be either open-source LLMs or smaller BERT-based models that have shown promise in legal reasoning tasks, such as (Chalkidis et al., 2020). Even though these legal-specific models do not exhibit few-shot learning capabilities, they would be ideal baseline models.

Independently of this extension, we plan to evaluate more general-purpose, but also legal-specific LLMs, and update the leaderboard of the dataset. It would be insightful to measure the impact of pre-training on the same or other legal systems. To our knowledge, there is currently only one publicly released family of legal LLMs that can follow instructions, `Saul-7B` (Colombo et al., 2024b), `Saul-54B` and `Saul-141B` (Colombo et al., 2024a).

Finally, the process followed to construct LAR-ECHR could be replicated with cases from other court proceedings to create new LAR datasets that are focused on other legal systems and/or languages.

Limitations

One limitation of our work has to do with the process followed to create the dataset. While the data were originally created by humans, the next statement prediction task is artificial. We employed semi-automatic techniques, based on legal expert annotations and embedding similarity of the arguments, to compile a challenging dataset. We also summarized the facts to fit in the context length of all the models. This process might have introduced biases and/or mistakes, as we have already discussed for the summaries of the facts (§ 4.3). The impact of these biases could only be measured by careful examination from legal experts and extensive comparisons with different variations (e.g. summaries from other models).

Furthermore, it should be noted that, as in many other legal NLP datasets, we are using the 'facts' of ECHR court decisions as if they are the factual information available prior to the final decision. However, due to the details of the legal process and the way that court proceedings are written, this is unrealistic (Medvedeva and Mcbride, 2023). The judges actually publish only the information that is supporting their final decision as the 'facts' of the case; not the original record that they had to consider in that process. To make the task realistic for a real-world application we should include the actual information that the parties had access to before the final judgement took place, but access to this information is very hard to get for most cases.

Ethics Statement

The primary objective of this research is to advance legal NLP and more specifically the use of LLMs as tools that assist–without replacing–legal professionals. A diverse set of communities can be benefited from our research: (a) the NLP community can challenge existing and future LLMs on an advanced legal reasoning dataset and even build new datasets for other courts, (b) legal practitioners can improve their understanding of the way these models make decisions and (c) the legal tech community can gain useful insights into LLM capabilities across different courts and legal systems, enabling them to design appropriate use cases and develop more accurate tools.

Most previous work in legal NLP, including both benchmarks and models, (Guha et al., 2023; Niklaus et al., 2023; Chalkidis et al., 2020), advocate that they do not aim to replace judges, but instead to assist them in reaching more informed decisions. However, most of them are trying to predict the outcome of legal decisions, without providing or evaluating legal reasoning explanations. In contrast, our work evaluates the ability of LLMs to identify the correct next statement in a judge's chain of legal arguments, which is closely linked to their capacity to produce valid legal reasoning.

When introducing a new legal NLP task, it is vital to consider the intended use cases for potential models designed for it (Medvedeva and Mcbride, 2023; Tsarapatsanis and Aletras, 2021). In our case, we advocate that such models be used solely as supporting tools to review the reasoning of legal professionals, rather than to produce their own legal reasoning (let alone predict the outcome of a case).

For example, we propose developing a legal verification tool, i.e. a tool that can verify the validity of the legal reasoning of an argument chain. This tool could be used by judges to validate the 'Law' section of the proceedings (after the final decision is taken) before publishing them. If a potential reasoning weakness is located by the model, then it could provide its own CoT explanation to pinpoint the root cause of the problem. The judges would then evaluate if they agree with the model or not, and if their reasoning requires revision. In this example ethical risks are almost completely mitigated, because the decision is already taken. This tool would help the judges prepare the proceedings faster and it could decrease oversights.

We recognize the ethical importance of data privacy and confidentiality. All data is obtained from publicly accessible online sources, without infringing any proprietary rights, and in accordance with the licenses under which they were released. The data from LAM:ECHR were released under the '`Apache 2.0`' license along with the respective software.[8] The data from ECtHR B were released under the 'Creative Commons Attribution-NonCommercial-ShareAlike 4.0' ('`CC BY-NC-SA 4.0`') license.[9] In accordance to '`CC BY-NC-SA 4.0`', we released our dataset under the same license as well.[10]

[8] https://github.com/trusthlt/mining-legal-arguments/blob/main/LICENSE

[9] https://huggingface.co/datasets/AUEB-NLP/ecthr_cases

[10] https://creativecommons.org/licenses/

Acknowledgments

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. In addition, AWS resources were provided by the National Infrastructures for Research and Technology GRNET and funded by the EU Recovery and Resiliency Facility.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: a natural language processing perspective. *PeerJ Comput. Sci.*, 2:e93.

Dor Bernsohn, Gil Semo, Yaron Vazana, Gila Hayat, Ben Hagag, Joel Niklaus, Rohit Saha, and Kyryl Truskovskyi. 2024. LegalLens: Leveraging LLMs for legal violation identification in unstructured text. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2129–2145, St. Julian's, Malta. Association for Computational Linguistics.

Leonard Bongard, Lena Held, and Ivan Habernal. 2022. The legal argument reasoning task in civil procedure. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 194–207, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

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.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Pierre Colombo, Telmo Pires, Malik Boudiaf, Rui Melo, Dominic Culver, Sofia Morgado, Etienne Malaboeuf, Gabriel Hautreux, Johanne Charpentier, and Michael Desa. 2024a. Saullm-54b & saullm-141b: Scaling up domain adaptation for the legal domain. *Preprint*, arXiv:2407.19584.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024b. Saullm-7b: A pioneering large language model for law. *Preprint*, arXiv:2403.03883.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2024. Mining legal arguments in court decisions. *Artificial Intelligence and Law*, 32(3):1–38.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language

by-nc-sa/4.0/

understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

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, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, et al. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Abhinav Joshi, Shounak Paul, Akshat Sharma, Pawan Goyal, Saptarshi Ghosh, and Ashutosh Modi. 2024. IL-TUR: Benchmark for Indian legal text understanding and reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11460–11499, Bangkok, Thailand. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, et al. 2023. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *Preprint*, arXiv:2405.20362.

Masha Medvedeva and Pauline Mcbride. 2023. Legal judgment prediction: If you are going to do it, do it right. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 73–84, Singapore. Association for Computational Linguistics.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, and othes. 2022. Text and code embeddings by contrastive pre-training. *Preprint*, arXiv:2201.10005.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. LEXTREME: A multi-lingual and multi-task benchmark for the legal domain. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3016–3054, Singapore. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, , et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599, Online. Association for Computational Linguistics.

Santosh T.y.s.s., Rashid Haddad, and Matthias Grabmair. 2024. ECtHR-PCR: A dataset for precedent understanding and prior case retrieval in the European court of human rights. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5473–5483, Torino, Italia. ELRA and ICCL.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A Prompt

The following is the complete version of the prompt presented in Table 4, that we designed for LAR-ECHR.

```
You will be provided with the
introductory Facts in a European Court
of Human Rights (ECHR) case, an excerpt
of arguments of that case and several
possible continuations of these
arguments. Your task is to determine
which continuation accurately extends
the original argument from the case. To
complete this task successfully:

1. Thoroughly analyze each provided
option to identify its connection to the
 initial argument presented.
2. Choose the option that not only
maintains the theme and context of the
initial argument but also follows
logically and seamlessly from it.
3. After selecting the most appropriate
continuation, provide a detailed
rationale for your choice.
4. Clearly state your answer by
specifying the letter corresponding to
the correct option.

Output format:
- Analysis: Provide an in-depth
evaluation of how each option relates to
 the initial argument.
- Explanation: Articulate why your
chosen option is the correct
continuation.
- Answer: State the letter of your
selected answer.
```

B Complete Example

The following is the complete version of the instance presented in Table 1. The facts of the case are presented first, then the examples as they were given to the LLMs for evaluation. After the input of the example, the complete CoT response of GPT-4o follows to help the reader understand the reasoning of the LLM. The explanations should not be considered an accurate representation of the inner thinking of the LLM, but still can help us understand its reasoning and can help non-experts understand the legal terms.

Facts:

```
The applicant, born in 1979, married B
in April 2001, giving birth to a
daughter, C, shortly after. B, who
suffered from psychiatric disorders due
to his traumatic experiences during the
Homeland War, subjected the applicant to
 verbal and physical abuse between 2003
and 2005. He was arrested in November
2005 and indicted for violent behavior
but released in December 2005. The
applicant moved to a women's shelter
with C in January 2006 for safety. B
continued his abusive behavior, leading
to further legal proceedings, including
charges of making death threats against
the applicant and a police officer, for
which he was found guilty and sentenced
to imprisonment in 2006. A restraining
order was issued, and subsequent appeals
 by B and the State Attorney's Office
were dismissed in 2007, though the
sentence was not enforced. Multiple
other criminal and minor offenses cases
were initiated against B over the years
for domestic violence and threatening
behavior. Meanwhile, the marriage of the
 applicant and B was dissolved in
November 2006. The applicant faced
difficulties in securing safe
accommodation due to B's continued
harassment, including hiring a private
detective to locate her. The legal
proceedings against B were ongoing, with
 several hearings adjourned due to B's
absence, and no psychiatric treatment
had been ensured despite recommendations
. Overall, the applicant struggled with
legal enforcement and protective
measures against B's continued threat
and harassment, affecting her and her
daughter's safety and stability.
```

Arguments:

```
The Government argued that in Croatia
the protection of victims of domestic
violence was ensured through the
mechanisms of criminal law, and in
particular the Protection against
Domestic Violence Act. In the present
case the relevant authorities had
```

reacted to the incidents of violence against the applicant by B, had instituted several sets of both criminal and minor offences proceedings and had applied such criminal sanctions and protective measures against B as they had considered proper and suitable in the circumstances. The Government submitted that the prison term imposed on B for not paying in full the fine imposed in the decision of the Z. Minor Offences Court of 2 October 2006 had not been enforced because Z. Prison had been full to capacity. Likewise, the measure of compulsory psycho-social treatment imposed on B in the same decision had not been implemented owing to the lack of licensed individuals or agencies able to execute such a protective measure (see paragraphs 31 and 34 above).
In addition, the Government had adopted two national strategies for protection against domestic violence (the first one covering the period between 2005 and 2007 and the second covering the period between 2008 and 2010) which included, inter alia, the education of all those involved in cases of domestic violence and cooperation with the non-governmental organisations working in that field as well as financial and other support for them. Thus, in 2008 only sixteen new shelters with a total of 329 places for the victims of violence had been established, of which six were State-funded.
The Court takes note of B's repeated violent behaviour towards the applicant. The facts in issue concern frequent episodes of violence in the period between November 2003 and June 2006, amounting to some two years and seven months. The violence was both verbal, including serious death threats, and physical, including hitting and kicking the applicant in the head, face and body, causing her injuries. In view of the fact that all the incidents of domestic violence in the present case concerned the same perpetrator and occurred in a continual manner, the Court will examine them as a continuous situation.

Continuation options:

(A) The Court takes further note of the psychiatric reports concerning B which indicated that he suffered from several mental disorders, including a severe form of PTSD, emphasised his tendency towards violence and his reduced ability to control his impulses, and reiterated the recommendation for continuing compulsory psychiatric treatment (see paragraphs 6, 12 and 13 above).

(B) Insofar as the applicant's complaint under Article 5 - 1 of the Convention is concerned, his confinement in the mental hospital can be divided into two periods: between 9 and 20 December 2002, and after 20 December 2002, when the hospital became his guardian.

(C) The Court notes, and this is not in dispute, that in August and October 2009 the police collected information about the applicant's character from his neighbours and relatives, pursuant to the instruction of the President of the Krasnogvardiyskyy Court (see paragraphs 13-15 and 20 above). That information was apparently used by the psychiatric hospital's specialists in the context of the forensic psychiatric examination of the applicant.

(D) The applicant firstly complained that (i) the decisions to order her to undergo a psychiatric examination and to arrest her and (ii) her detention for a total period of eighty-three days were in breach of Article 8 of the Convention . However, the Court considers that these complaints do not raise any separate issue which is not covered by the finding of a violation of Article 5 - 1 of the Convention.

GPT-4o Response:

Analysis:
To find the correct continuation of the initial argument, we need to focus

on both the factual context and legal analysis provided.
- Option (A) aligns closely with the initial arguments by addressing the psychiatric reports about B, reinforcing the situation's consistency with his diagnosed mental disorders. It maintains the theme of continuous violence and the need for psychiatric treatment, linking back to the facts concerning B's mental health.
- Option (B) discusses periods of confinement in a mental hospital, but this appears unrelated to the specific facts of the case involving B's violent behavior toward the applicant and does not further the argument regarding domestic violence or its handling by Croatian authorities.
- Option (C) mentions the police collecting information about the applicant's character and a forensic psychiatric examination, which appears out of context, as there is no mention of such actions linked to either the applicant or B in the provided facts.
- Option (D) pertains to complaints about forced psychiatric evaluations and unrelated detention, diverging from the continuity of violence by B and responses by Croatian authorities, and thus, it does not logically follow the given argument.

Explanation:
Option (A) is the most logical continuation as it directly ties into and extends the argument concerning the failure of the Croatian authorities to implement appropriate protective measures and treatment for B despite his documented psychiatric issues and violent tendencies. This reflects the Court's approach of understanding the situation as a "continuous situation" of abuse, emphasizing the importance of ongoing psychiatric intervention, which was previously recommended and not enacted.

Answer: A ✔