Contemporary LLMs struggle with extracting formal legal arguments

Lena Held[1,2] **and Ivan Habernal**[2]
Trustworthy Human Language Technologies
[1] Department of Computer Science, Technical University of Darmstadt
[2] Research Center Trustworthy Data Science and Security of the University Alliance Ruhr,
Faculty of Computer Science, Ruhr University Bochum
lena.held@ruhr-uni-bochum.de
www.trusthlt.org

Abstract

Legal Argument Mining (LAM) is a complex challenge for humans and language models alike. This paper explores the application of Large Language Models (LLMs) in LAM, focusing on the identification of fine-grained argument types within judgment texts. We compare the performance of Flan-T5 and Llama 3 models against a baseline RoBERTa model to study if the advantages of magnitude-bigger LLMs can be leveraged for this task. Our study investigates the effectiveness of fine-tuning and prompting strategies in enhancing the models' ability to discern nuanced argument types. Although the tested models succeed at implementing the task in a generative fashion, our findings indicate that neither fine-tuning nor prompting could surpass the performance of a domain-pretrained encoder-only model. This highlights the challenges and limitations in adapting LLMs to the specialized domain of legal argumentation. The insights gained from this research contribute to the ongoing discourse on optimizing NLP models for complex, domain-specific tasks. Our code and data for reproducibility are available at https://github.com/trusthlt/legal-argument-spans.

1 Introduction

Mining legal arguments is the task of identifying, classifying and structuring argumentative units in a legal document. Early works on argument mining in the legal domain considered generic types of arguments, such as claim and premise (Mochales and Moens, 2011). Recent works have shifted towards a legal theory inspired typology of argumentation (Habernal et al., 2024; Lüders and Stohlmann, 2024). Finegrained argument schemes can help legal scholars in structuring and analyzing legal documents, allowing insights into form and strategy of the presented argumentative content. However, the larger inventory of categories, as well as the granularity and complexity make the task of identifying specific argument types more challenging. A layperson would find the categories difficult to comprehend and even domain experts typically require instructions, additional context and training to identify the arguments in text. Despite this, the majority of existing works classify argument spans with a limited context in the given legal document (e.g., only a single paragraph). This design choice is mostly due to the limited input size of the used models. However, the quantitative legal analysis of Habernal et al. (2024, p. 583) clearly shows that the lack of contextual information inevitably leads to a drop in prediction performance.

Contemporary decoder-only LLMs offer two main advantages over the early encoder-based models, namely the capability to ingest very large input documents and the in-context learning (ICL) abilities without the need of expensive fine-tuning. We hypothesize that these features would help us overcome the difficulties in mining legal arguments. However, the research community on legal argument mining lacks the empirical evidence whether LLMs can be used with more complex argumentation schemes. We aim to address this gap in this work.

First, we look into how encoder-decoder and decoder-only models can be used to mine fine-grained legal arguments in long documents. Second, we investigate how additional information and context in the input affects the performance. We hypothesize that the additional context which LLMs are able to ingest will further boost the performance of the argument extraction and classification. We explore the potential of in-context learning and fine-tuning.

We use the LAM:ECHR dataset (Habernal et al., 2024), which contains an expert-crafted legal argument scheme. We reformat the dataset and test it with the contemporary LLMs Flan-T5 and Llama 3, attempting to improve upon the state-of-the-art performance on the dataset. We also explore the in-

Proceedings of the Natural Legal Language Processing Workshop 2025, pages 292–303
November 8, 2025

clusion of additional context to the input paragraph, as well as supplying the models with annotation guidelines to further boost the understanding of the label categories.

2 Related Work

Argument Mining (AM) describes the extraction and analysis of natural language into argumentative components to enable their study in a variety of applications and domains (Palau and Moens, 2009; Cabrio and Villata, 2018). Traditionally, the process of mining arguments is often split into multiple subtasks considering argument components, roles and relationships (Stab and Gurevych, 2017). Although these tasks are usually formulated as a classification problem, the emerge of LLMs has enabled the possibility of using new techniques, such as prompt-based extraction or generative approaches. One such approach is successfully implemented by Kawarada et al. (2024), who extend the idea of generating a structured label format from input which was first introduced by Paolini et al. (2021) to the task of argument mining. By fine-tuning Flan-T5 (Longpre et al., 2023) for the different subtasks span identification, component classification and relation classification, they show the potential of this technique on the established argument mining datasets containing persuasive essays (AAEC) (Stab and Gurevych, 2017), medical abstracts (AbstRCT) (Dernoncourt and Lee, 2017) and comments on Consumer Debt Collection Practices (CDCP) (Park and Cardie, 2018). The key point of this technique is the application of the identification and classification jointly into the output text generated by the decoder. This way the decoder will generate a copy of the input sequence with in-text annotations and labels. It remains to be seen if this approach is transferable to a richer argument typology and longer contextualized input sequences, since the tested datasets contain short documents and broader argument schemes.

A different approach is pursued by Cabessa et al. (2025), in which a range of contemporary LLMs are fine-tuned for individual argument mining subtasks across the same datasets. Contrastive to the previous approach, the desired LLM output is in a structured format. The best performance is achieved by fine-tuning Llama 3.1 8B (Grattafiori et al., 2024) which reaches state-of-the-art performance on all datasets and tasks. Cabessa et al. (2025) also investigate the influence of context, comparing a paragraph-level and an essay-level classification and find that in some cases an extended context can boost the performance.

2.1 Legal Argument Mining

Legal Argument Mining as a domain-specific form of AM focuses on extracting arguments from legal texts based on argumentation schemes stemming from the legal domain. These schemes can range from a form similar to AM with premise, conclusion, clause and relation (Poudyal et al., 2020) to domain-specific forms, such as issue, reasons, conclusion (Elaraby and Litman, 2022; Xu et al., 2020). Prompting LLMs like GPT-4 has also been proven to work for LAM using the labels premise and conclusion on the ECHR-AM corpus (Poudyal et al., 2020) as shown by Al Zubaer et al. (2023).

A substantially more granular distinction of legal argument types was proposed by Habernal et al. (2024), who introduced the LAM:ECHR dataset consisting of 373 ECHR judgment documents, annotated by legal experts. The annotations are made up of 15 formal legal argument types and the task is interpreted as a span prediction task using a token-wise BIO-tagging scheme. The current state-of-the-art performance on this dataset is achieved with pretrained RoBERTa (Liu et al., 2019) and Legal-BERT (Chalkidis et al., 2020) models trained for token-wise classification. Highly represented classes such as "O" (i.e., no argument), "Application to the concrete case" and "Precedents ECHR" perform well with over 80% F1-score. However, some labels, most notably the ones with a low representation in the dataset only achieve very low scores or even zero, which leads to 43.13% macro F1-score for the best model, a RoBERTa model which received legal pretraining. To the best of our knowledge, there are no works exploring the LAM:ECHR dataset further in the scope of argument mining.

3 Methodology

Our overall methodology is as follows. First, we cast the BIO-tagged LAM:ECHR dataset into a format which is more suitable to be passed to the LLMs of our choice. Because of this, we also need to adjust the evaluation metric to suit the expected predictions better; see in detail in the next subsection. We then fine-tune Flan-T5-XXL, trying to replicate the good performance the model demonstrated in the related in-text annotation ap-

Case ID 001-61184

Paragraph ID 23

Input 79. There has accordingly been a violation of that provision.

Gold Label 79. [begin_of_annotation] There has accordingly been a violation of that provision. [end_of_annotation] **Decision ECHR** [end_of_label]

Context 77. It should not be forgotten that the interests of the child are paramount in such a case, which is why the Portuguese authorities may be right in considering that parental responsibility must now be granted to the mother. [...] 78. Having regard to the foregoing, and notwithstanding the respondent State's margin of appreciation in the matter, the Court concludes that [...]

Figure 1: An example datapoint from the reformatted LAM:ECHR dataset.

proach (Kawarada et al., 2024). To leverage larger context windows and for the general advantages of contemporary LLMs, we also experiment with two models from the Llama family, Llama 3.1 8B and Llama 3.3 70B, based on the success of those models for similar tasks in other domains (Cabessa et al., 2025). We then explore strategies to enhance and improve our approach by extending the context for each input and including explanations for each label by adding the original annotation guidelines of the dataset.

3.1 Dataset

The original token-based dataset is not suited for prompting or fine-tuning an LLM. Without violating the split intended in the original dataset, we reformat the token-based annotated data into a paragraph-level dataset with annotated spans. Due to the reinterpretation, the distribution of labels is different to the original distribution; the numbers are shown in Table 1. We focus only on the formal legal argument types and ignore the annotated roles, as these were rather well identifiable in the original work already.

Span annotation formalization. The task can now be interpreted as: For a given paragraph of a legal judgment document, identify the argument spans and classify the formal legal argument types. The expected outcome as seen in Figure 1 is the original input text, along with tags which denote the begin and end of an argument as well as a label and a tag for the end of the label.

Follwing Kawarada et al. (2024), we chose the descriptive tags `'[begin_of_annotation]'`, `'[end_of_annotation]'` and `'[end_of_label]'` as delimiters.

Various sizes of context. The annotation guidelines which were used to create LAM:ECHR suggest that in order to correctly label a paragraph, the annotator has to be aware of previous paragraphs and each paragraph "must always be read in context". We want to test whether contextual information which a human expert needs to identify arguments is also helpful in model training. Based on the reformatted dataset, we create four variations which include different amounts of context. The original version only contains the paragraph to be annotated with no additional context, while the variants include a context window of n (unannotated) previous paragraphs before the target paragraph (if applicable). We create the variants with a context window of 2 and 4 paragraphs. The last version consists of the entire text in each judgment, reducing the dataset to 356 annotated full judgment documents in the training set and 37 in the test set. The prompt format alone increases the maximum input size of the dataset to over 2,600 tokens. Adding context further increases the maximum input to more than 3,500 tokens and processing the entire document at once requires up to 260,000 tokens. The annotation guidelines which provide explanations on the label classes need an additional 1,700 tokens. Such an input size is something that only contemporary LLMs with large input windows can handle.

3.2 Evaluation

A fair comparison to the best established baseline on the dataset requires a re-evaluation of the original predictions which are encoded as token-level BIO. While the token-based evaluation has the advantage of accounting for partially correct spans, it also values longer argument spans more than short argument spans.

In our reformulation of the dataset, we consider a span as classified correctly if it is identified at the correct position in the text and labeled with the correct class. To also consider partially correct spans, in which just a few tokens are outside or additionally inside the argument unit, we introduce a relaxation in the evaluation of a correct span position in the text. We argue that the exact token

Argument type	F1	Freq.
Application to the concrete case	0.80	851
Precedents of the ECHR	0.80	214
Test of the principle of proportionality - Proportionality	0.48	178
Decision ECHR	0.72	130
Test of the principle of proportionality - Legal basis	0.50	71
Non contestation by the parties	0.77	28
Test of the principle of proportionality - Legitimate purpose	0.75	18
Distinguishing	0.43	16
Margin of appreciation	0.74	12
Teleological interpretation	0.14	12
Comparative law	0.50	2
Overruling	0.00	1
Test of the principle of proportionality - Suitability	0.00	1
Textual interpretation	0.00	1
Systematic interpretation	0.00	1
Macro avg	0.41	1536

Table 1: Label frequencies in the test set of our modified version of LAM:ECHR along with the RoBERTa baseline scores predicted by the best performing model which we replicated following Habernal et al. (2024). We evaluated the outputs by the metrics described in section 3.2 using a threshold of 10%.

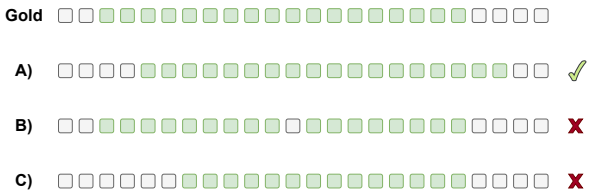

Figure 2: Each box represents a token. An argument span is accepted if the start and end of the span are within a certain threshold % of the original length of the argument. A) shows an accepted partially correct span, B) is separated into two spans and neither span is within the threshold of the ground truth, C) is not within the acceptance threshold.

at which an argument begins or ends is not important, as long as the core meaning of the argument is captured in the predicted span. We therefore allow some variation in the exact delimiters of the span, by accepting a span position as correct if the start and end tokens are within a certain threshold (0, 10, and 20%) of the original length of the argument. Especially for the usually numbered paragraphs in the dataset, this relaxed metric for example allows the argument to begin with the paragraph number or without it. Figure 2 shows an example of this relaxed evaluation. For the evaluation in this work, a threshold of 10% is used and we compare the changes in performance introduced by the relaxation.

Our approach jointly identifies and classifies an argument. The performance is measured based on the final F1-score of the argument component classification (ACC) which by design also includes the prior detection of the argument span. Due to the high imbalance of classes in the dataset, we report the macro-average F1-score alongside the weighted-average F1-score. We can derive the model's ability for argument component identification (ACI) by replacing the exact label with "argument", thus simplifying the classification into a binary problem of argumentative and non-argumentative tokens in the text. This serves as an auxiliary metric to see if the model is able to extract argumentative content correctly at all. The reported score is the weighted-average F1-score. This also helps us to estimate the performance of the classification of 'no argument', which is not considered in our dataset format, contrastive to the original BIO token-level format, which includes the 'O' tag.

Additionally, we need a metric to observe how well the model learns to stick to the required output format. For this, we report the percentage of paragraphs that received a correctly formatted output. If a single token differs after masking the annotation tags and labels of the output, the entire paragraph is considered invalid. The score is reported as 'Output Format Validity'.

3.3 Fine-tuning Flan-T5

We take the most successful model used by Kawarada et al. (2024), Flan-T5-XXL, as a starting point and fine-tune it for the task of identifying and classifying the arguments in the modified LAM:ECHR dataset. We train a LoRa (Hu et al., 2022) adapter for 2 epochs on the dataset. Because

Flan-T5 is not trained on long documents, we only fine-tune the model with paragraph-level input, as defined in the initial version of our modified dataset. This experiment serves as a starting point to see if in-text annotations can be used in conjunction with legal argument types.

3.4 Fine-tuning Llama 3

As Cabessa et al. (2025) have proven the Llama model family to be a viable contender for argument mining, we also fine-tune Llama 3.1 8B Instruct and Llama 3.3 70B Instruct. This gives us the opportunity to test a contemporary decoder-only model and observe differences in performance depending on the model size. Due to the computational costs, we limit all fine-tuning to (Q)LoRa (Dettmers et al., 2023) adapters and 4bit quantization for the 70B model. We train all models for 1 epoch and use a prompt format including instructions, context and input as shown in Figure 3.

Experiments with ICL To test the general capabilities of the models, our first experiment variant makes use of ICL samples to make the model adhere to the output format and test the out-of-the-box performance on identifying complex legal argument types. We hand-pick 4 samples from the training data with different argument type labels. The major difficulty of this task is steering the model towards the desired output format and having it assign a valid label.

Experiments with Fine-tuning In our second round of experiments, we fine-tune the models without additional context, giving them just a single paragraph. This approach is mirroring the method used for fine-tuning Flan-T5.

Experiments with more context In the next round, we include additional context in the prompt with the context windows adding two and four previous paragraphs, respectively. The additional context should help the model better understand the current paragraph. To make use of the large context window of the Llama models, we also train with the full documents for maximum context. This way, the model has all relevant information about the case available. For these experiments we increase the number of epochs to 5 to make up for the decreased amount of training samples. Due to computational constraints, we limit the training data to samples with less than 30,000 tokens.

Experiments with annotation guidelines For the next more advanced experiments, we acquire the annotation guidelines used to create LAM:ECHR. These guidelines contain descriptions of the labels as well as examples. We hypothesize that giving the model a better understanding of the labels should improve the performance. The guidelines can be found in Appendix B. We include the guidelines in our prompt.

3.5 Training details

For reproducibility and full transparency, the code of all experiments is available at `https://github.com/trusthlt/legal-argument-spans`. All the training details, including epochs, parameters, configuration, evaluation, etc. are available in the scripts and `README.md` documents. All experiments were conducted on one NVIDIA A100 80GB GPU.

3.6 Additional experiments with legal LLMs

Apart from the aforementioned models, we also experiment with several other leading legal LLMs, such as Lawma (Dominguez-Olmedo et al., 2024) and SaulLM (Colombo et al., 2024). Unfortunately, we were unable to fine-tune these models for our task, such that we do not include them in the main result section. Nevertheless, we believe that these additional experiments highlight the difficulty of adapting LLMs to a complex task like LAM.

Lawma 8B This legal language model is based on Llama 3 8B and specifically trained for legal text classification tasks, making it a good candidate for our experiments. However, after fine-tuning, the model still defaults to the short classification-style answers which it was originally trained on and ignores the output format that we require entirely.

SaulLM 7B Similar to Lawma, the model is a good candidate because of its specific legal pretraining. After fine-tuning for our task, the model still gives explanatory and "chatty" answers, refusing to adhere to the output format.

4 Results

Table 2 shows the results of our experiments for fine-tuned Flan-T5-XXL, Llama 3.1 8B Instruct and Llama 3.3 70B Instruct compared to the original baseline.

Model	Configuration	ACI	ACC		Output Format Validity
		F1	macro F1	weighted F1	
Leg-RoBERTaL-15k	– context = 0	**0.97**	**0.41**	**0.73**	1.000
Flan-T5-XXL	– context = 0	0.96	0.32	0.69	0.998
Llama 3.1 8B Instruct	– ICL	0.37	0.19	0.29	0.620
	– context = 0	0.95	0.29	0.69	0.992
	– context = 2	0.96	0.27	0.70	0.991
	– context = 4	0.95	0.26	0.70	0.984
	– context = 4 + AG	0.96	0.27	0.69	0.986
	– full-document + AG	0.26	0.09	0.39	0.351
Llama 3.3 70B Instruct (4bit)	– ICL	0.40	0.21	0.37	0.643
	– context = 0	0.95	0.30	0.70	0.992
	– context = 2	**0.97**	0.30	0.72	0.992
	– context = 4	**0.97**	0.30	**0.73**	0.991
	– context = 4 + AG	0.95	0.22	0.64	0.990
	– full-document + AG	0.48	0.20	0.52	0.595

Table 2: F1-scores calculated on the test dataset on Leg-RoBERTaL-15k (Habernal et al., 2024), Flan-T5-XXL, Llama 3.1 8B Instruct and Llama 3.3 70B Instruct (4bit). ACI shows the weighted F1-score of argumentative and non-argumentative components, ACC shows the macro and weighted F1-score for the joint task of identifying and classifying a legal argument. Output format validity shows the percentage of correctly formatted outputs.

Flan-T5 The experiments using Flan-T5 achieve an almost perfect output format validity, suggesting that the encoder-decoder model is able to easily learn how to produce the correct output format. The high score for ACI also suggests that the model can learn how to identify argumentative and non-argumentative components. In terms of classification, the model is able to learn the different argument types decently, but stays below the baseline for both weighted F1-score and macro F1-score. The lower macro F1-score also hints at a better performance for more frequent labels, while infrequent labels are misclassified more often. This is also confirmed when looking at the individual label classification scores in Table 3. Still, Flan-T5 is able to outperform the other models for the labels "Non contestation by the parties", "Decision ECHR" and "Test of the principle of proportionality - Legitimate purpose", but is beaten at all other argument types.

Llama 3.1 8B Instruct The experiments based on Llama 3.1 8B Instruct, although expected to outperform Flan-T5 due to its magnitude larger model size, are just slightly worse in identifying argumentative components as well as sticking to the output format. All experiments with added context are still able to achieve a good output format validity and ACI. The performance of these experiments manages to stay roughly on par with Flan-T5, although there is a slight decrease in macro F1-score for the argument type classification. Added context did not improve nor decrease the performance. Adding annotation guidelines did also not change the outcome for the context experiments.

In-context learning using only the base version of the model without any fine-tuning performs significantly worse compared to the context experiments with only 62% of the outputs even being in the correct format. As a result the performance for argument type classification is also a lot worse compared to the fine-tuned versions.

An especially bad performance can be observed for the configuration using the full judgment document as training data. This experiment only has around 35% correctly formatted outputs and the worst scores in every aspect out of all experiments. Although there was no impact when adding the annotation guidelines to the paragraphed input, it is possible that the input size using the full document alongside the annotation guidelines is simply too large for the model to learn anything meaningful. It is also possible that the large input size increases the difficulty for the model to learn the correct output format.

Llama 3.3 70B Similar to its 8B sibling, the model is not able to adhere to the output format and classify correctly using only ICL, even though the overall classification scores are slightly better than for Llama 3.1 8B in this configuration. Just like the smaller model, the best scores are achieved by finetuning with paragraph-level input. Although the macro F1-score is still lower than Flan-T5, the

weighted F1-score is on par with the baseline.

For the 70B variant, there is also a small increase in performance observable when adding more context with a context window of 4 showing the best scores of all Llama experiments.

Contrary to the experiments on the smaller model, adding annotation guidelines has a more detrimental effect on performance. For the paragraphed configuration, the output format validity and component identification stay intact, which leads us to believe that the annotation guidelines cause the model to label with a more even distribution than the actual training data.

Using the entire document as input makes it difficult for the model to output the correct format and detecting argumentative components also suffers, although the drop is less severe than for Llama 3.1 8B.

A closer look at the individual argument type classifications for the best performing models in Table 3 shows that labels concerning the "Test of the principle of proportionality" appear to be difficult for all models, but both Flan-T5 and Llama 3.3 70B exhibit extreme difficulty with these labels. None of the tested models were able to improve the performance for the underrepresented classes. The best performing Llama model beat the baseline for the three most prevalent labels, but trades off the performance on other labels with a large drop in performance for "Test of the principle of proportionality - Legitimate purpose", "Distinguishing" and "Margin of appreciation".

4.1 Evaluation strategy

To have a more relaxed notion of an identified argument span, we apply the previously introduced relaxed acceptance threshold in our evaluation process. This was originally implemented to ensure that generative models were also credited for partial matches.

Surprisingly, we find that a higher threshold allows for more correct annotation for the baseline model, but the gain in our fine-tuned models is very limited. This shows that our models were able to learn the annotation scheme and adopt rules for identifying span borders from the training data. Table 4 in Appendix C shows the gains for threshold 10% and 20% compared to a strict evaluation with 0% threshold. Nevertheless, a relaxed strategy should be kept in mind for this task format, otherwise "almost correct" matches could be undervalued.

5 Discussion

We can make several observations from our experiments. First of all, fine-tuning LLMs on the task of adding annotations and labels in-text is feasible even with a domain-specific legal dataset. We find that reformatting the task into in-text annotations is a possible avenue for the future of legal argument mining outside of traditional formats like BIO even in complex and difficult annotation schemes. After just 1 epoch of training, the models are able to output the required formats with proper labels. This is an optimistic finding, given that newer and better LLMs are introduced at a high frequency. And even though the overall performance of our best trained model could barely keep up with the RoBERTa baseline, it is possible that scaling to much larger models could outperform the baseline. Using an LLM instead of an encoder model could also open up the possibility to handle more difficult argument annotations (i.e., legal argument relations) through in-text classification.

Secondly, we hypothesized that including additional context or the full document would provide a better understanding of the short paragraph and help with argument classification. From the results, however, we can surmise that the actual effect is minimal. It is conceivable that we did not chose the ideal method to incorporate context and a more sophisticated method than merely adding previous paragraphs is needed to draw the LLM's full potential.

In a similar fashion, annotation guidelines did not boost the performance either, which also leads us to believe that the biggest struggle for LLMs is the interpretation and understanding of the actual labels. This could also imply that providing context and guidelines is not sufficient to understand the argument categories properly.

Despite the advancements of contemporary LLMs in related works, successfully classifying labels like premise and conclusion and even argument relations, the models we tested struggle with handling the nuances of a complex legal argument scheme. We can also hypothesize that due to the nature of the arguments, which are based on the formal meaning of the argument, encoder models might be better suited by design to pick up standardized formulations and keywords which are often used in judgment documents.

Fine-tuning has proven to be the best technique to ensure adherence to the correct output format,

Argument Type	Baseline	Flan-T5	Llama 3.3 70B
Application to the concrete case	0.80	0.77	**0.81**
Precedents of the ECHR	0.80	0.79	**0.83**
Test of the principle of proportionality - Proportionality	0.48	0.38	**0.57**
Decision ECHR	0.72	**0.75**	0.67
Test of the principle of proportionality - Legal basis	**0.50**	0.40	**0.50**
Non contestation by the parties	**0.77**	**0.77**	0.64
Test of the principle of proportionality - Legitimate purpose	0.75	**0.79**	0.29
Distinguishing	**0.43**	0.32	0.12
Margin of appreciation	**0.74**	0.14	0.38
Teleological interpretation	0.00	0.00	0.00
Comparative law	0.50	0.00	0.00
Overruling	0.00	0.00	0.00
Test of the principle of proportionality - Suitability	0.00	0.00	0.00
Textual interpretation	0.00	0.00	0.00
Systematic interpretation	0.00	0.00	0.00
Macro-average	**0.41**	0.32	0.30
Weighted-average	**0.73**	0.69	**0.73**

Table 3: Label-specific F1-scores on the 37 test documents using the best performing configurations for the RoBERTa baseline, Flan-T5-XXL and Llama 3.3 70B.

while ICL is not enough for the complexity of the fine-grained legal argumentation scheme. On the other hand, fine-tuning is also computationally expensive and it is less cost-efficient to train and fine-tune an LLM like Llama 3.3 70B, compared to training a RoBERTa model.

Future work could focus on better techniques for teaching LLMs the knowledge necessary to understand and apply complex argumentation schemes. The addition of annotation guidelines did not prove to be useful in our case, but it could still be helpful to incorporate them in a different way in future experiments, because the underlying idea of requiring instructions, examples and context to solve the task is still at the core of legal reasoning. It would also be interesting to experiment with an LLM that has received extensive legal pretraining and is still able to be finetuned for specific tasks.

Another difficulty that needs to be overcome is how the training procedure can make up for the massively imbalanced class representation. A training dataset of higher quality and carefully selected, representative examples could be more beneficial than a larger amount of data. We believe that reasoning models or reinforcement learning training methods could also be leveraged to enable this in future work.

6 Conclusion

Our study finds that Flan-T5 and Llama 3 did not outperform the RoBERTa baseline on average in fine-grained legal argument mining, despite using fine-tuning and enriching the prompts with context and annotation guidelines. However, the performance surpasses the baseline for some argument types, showing that instruction-tuned LLMs generally have the potential to learn legal argument classification. Contrary to the findings of related works, there was no strong indication that adding context helps, but passing the entire document actually reduced the performance. We suggest that the exploration of long contexts in the legal domain should be explored further in this regard. Underrepresented and unbalanced labels remain a challenge, highlighting the difficulty of the task. A better integration of the annotation guidelines could be a good future direction to teach LLMs the specific skills to understand and apply such a complex legal argument scheme. The specialized nature of legal argumentation and its connection to legal theory presents unique challenges that current LLMs struggle to understand, emphasizing the need for further research and potential domain-specific adaptations.

Limitations

Although we design the prompts to the best of our knowledge and draw inspiration from the related works, it is possible that a different prompt design could achieve better results. The fine-tuned models are able to perform the task, but generally, the performance then degrades on the tasks they were originally trained for.

Ethics statement

To the best of our knowledge, our work falls under the umbrella of empirical legal studies with the aim

to better understand the nature of argumentation in human rights cases in the EU, and therefore we see no risks in misusing our research. Moreover, all datasets are public and open.

Acknowledgements

This work has been supported by the German Research Foundation as part of the ECALP project (HA 8018/2-1) and by the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within the https://uaruhr.de.

References

Abdullah Al Zubaer, Michael Granitzer, and Jelena Mitrović. 2023. Performance analysis of large language models in the domain of legal argument mining. *Frontiers in Artificial Intelligence*, 6.

Jérémie Cabessa, Hugo Hernault, and Umer Mushtaq. 2025. Argument Mining with Fine-Tuned Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6624–6635, Abu Dhabi, UAE. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2018. Five Years of Argument Mining: a Data-driven Analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5427–5433. International Joint Conferences on Artificial Intelligence Organization.

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.

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. 2024. SaulLM-7B: A pioneering Large Language Model for Law. *Preprint*, arXiv:2403.03883.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Ricardo Dominguez-Olmedo, Vedant Nanda, Rediet Abebe, Stefan Bechtold, Christoph Engel, Jens Frankenreiter, Krishna Gummadi, Moritz Hardt, and Michael Livermore. 2024. Lawma: The Power of Specialization for Legal Tasks. *Preprint*, arXiv:2407.16615.

Mohamed Elaraby and Diane Litman. 2022. ArgLegalSumm: Improving Abstractive Summarization of Legal Documents with Argument Mining. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6187–6194, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Grattafiori et al. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

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*, 32:557–594.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Masayuki Kawarada, Tsutomu Hirao, Wataru Uchida, and Masaaki Nagata. 2024. Argument Mining as a Text-to-Text Generation Task. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2002–2014, St. Julian's, Malta. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

Kilian Lüders and Bent Stohlmann. 2024. Classifying Proportionality - Identification of a Legal Argument. *Artificial Intelligence and Law*.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation Mining. *Artificial Intelligence and Law*, 19(1):1–22.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *International Conference on Artificial Intelligence and Law*.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured Prediction as Translation between Augmented Natural Languages. In *International Conference on Learning Representations*.

Joonsuk Park and Claire Cardie. 2018. A Corpus of eRulemaking User Comments for Measuring Evaluability of Arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. ECHR: Legal Corpus for Argument Mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75, Online. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, 43(3):619–659.

Huihui Xu, Jaromír Šavelka, and Kevin D. Ashley. 2020. Using Argument Mining for Legal Text Summarization. In *Proceedings of the 33rd International Conference on Legal Knowledge and Information Systems (JURIX 2020)*, pages 184–193, Virtual event. IOS Press.

A Input prompt example

Figure 3 shows an example prompt used with a context window of four previous paragraphs.

B Annotation guidelines

The annotation guidelines and instructions in Figure 4 were obtained from Habernal et al. (2024) and used for augmenting the prompts for the Llama models.

C Effect of leniency on span threshold

Table 4 shows the evaluation with different levels of thresholds for accepted span border variations. There is essentially no increase score even when accepting up to 20% offset from the original argument beginning and end for our fine-tuned models. The largest difference is visible for the original baseline, suggesting that the original model was not as good at finding the exact begin and end of an argument as our fine-tuned models.

Instruction:
Annotate the given ECtHR judgment with formal argument types. Mark the arguments by inserting the tags [begin_of_annotation] and [end_of_annotation]. After these tags, append the label for the argument type and close the label with [end_of_label]. Do not modify the original text otherwise and do not append additional information or explain anything. Only attach a label from the following list: ['Application to the concrete case', 'Decision ECHR', 'Non contestation by the parties', 'Precedents of the ECHR', 'Test of the principle of proportionality - Legal basis', 'Margin of appreciation', 'Test of the principle of proportionality - Proportionality', 'Test of the principle of proportionality - Legitimate purpose', 'Textual interpretation', 'Distinguishing', 'Teleological interpretation', 'Overruling', 'Systematic interpretation', 'Test of the principle of proportionality - Suitability', 'Comparative law']
Context:
64. The second section, entitled "Transitioning to Interrogation - The Initial Interview", deals with the stage before the application of EITs. It reads: "Interrogators use the Initial Interview to assess the initial resistance posture of the HVD and to determine – in a relatively benign environment – if the HVD intends to willingly participate with CIA interrogators. The standard on participation is set very high during the Initial Interview. The HVD would have to willingly provide information on actionable threats and location information on High-Value Targets at large not lower level information for interrogators to continue with the neutral approach. [REDACTED] to HQS. Once approved, the interrogation process begins provided the required medical and psychological assessments contain no contra indications to interrogation." 65. The third section, "Interrogation", which is largely redacted, describes the standard combined application of interrogation techniques defined as 1) "existing detention conditions", 2) "conditioning techniques", 3) "corrective techniques" and 4) "coercive techniques". 1) The part dealing with the "existing detention conditions" reads:
Input:
"Detention conditions are not interrogation techniques, but they have an impact on the detainee undergoing interrogation. Specifically, the HVD will be exposed to white noise/loud sounds (not to exceed 79 decibels) and constant light during portions of the interrogation process. These conditions provide additional operational security: white noise/loud sounds mask conversations of staff members and deny the HVD any auditory clues about his surroundings and deter and disrupt the HVD's potential efforts to communicate with other detainees. Constant light provides an improved environment for Black Site security, medical, psychological, and interrogator staff to monitor the HVD."

Figure 3: Example of a prompt providing a context window = 4, which adds the 4 paragraphs *before* the target input paragraph.

Model	**Config**	**ACC**					
		macro-avg			weighted-avg		
	Threshold	0%	10%	20%	0%	10%	20%
Leg-RoBERTaL-15k	– context = 0	0.37	0.41	0.42	0.71	0.73	0.74
Flan-T5-XXL	– context = 0	0.32	0.32	0.32	0.69	0.69	0.69
Llama 3.1 8B Instruct	– ICL	0.19	0.19	0.19	0.27	0.29	0.30
	– context = 0	0.29	0.29	0.29	0.68	0.69	0.69
	– context = 2	0.27	0.27	0.27	0.70	0.70	0.70
	– context = 4	0.26	0.26	0.27	0.69	0.70	0.70
	– context = 4 + AG	0.27	0.27	0.27	0.69	0.69	0.69
	– full-document + AG	0.09	0.09	0.10	0.38	0.39	0.40
Llama 3.3 70B Instruct (4bit)	– ICL	0.21	0.21	0.22	0.35	0.37	0.39
	– context = 0	0.30	0.30	0.30	0.70	0.70	0.71
	– context = 2	0.30	0.30	0.31	0.72	0.72	0.73
	– context = 4	0.29	0.30	0.30	0.72	0.73	0.73
	– context = 4 + AG	0.22	0.22	0.22	0.63	0.64	0.64
	– full-document + AG	0.20	0.20	0.20	0.51	0.52	0.52

Table 4: F1-scores calculated on the test dataset on Leg-RoBERTaL-15k (Habernal et al., 2024), Flan-T5-XXL, Llama 3.1 8B Instruct and Llama 3.3 70B Instruct (4bit). The argument component classification is calculated with a span identification threshold of 0% (10% and 20%) for macro F1 and weighted F1 score.

These guideline annotations serve as an instruction manual for the annotation of ECtHR judgments. Judgments of the ECtHR shall be annotated according to this guideline. Thereby, it is in the nature of things that a classification on the basis of the categories provided in the guideline can only be made based on a critical appraisal of the full argumentation of the ECtHR.
It has to be distinguished between the fifteen possible types of legal arguments.
In many cases, the category of a paragraph of the decision does not result exclusively from the text of that paragraph itself, but only in connection with further paragraphs, so that each paragraph must always be read in context, i.e. in relation to what has been addressed in a previous one.
1. "Non contestation by the parties": Procedural arguments are generally marked as such due to their special nature. This means that even in the part of the judgment on the application to the concrete case, the relevant sentences are not marked as such – although they are nevertheless part of this section – but are grouped into the category of procedural arguments (Non contestation by the parties).
2. "Textual interpretation": The wording is the first indication, but not a rigid boundary for the regulatory content of a norm (which can go beyond its wording). The textual interpretation is complemented by other methods of interpretation (see below). It can be referred to the meaning of the norm wording at the time of its origin or its application considering technical or (most subsidiarily) colloquial language. According to the final clause of the ECtHR, only English and French are "authentic" languages, i.e. only these are to be used (other languages only subsidiarily) for the interpretation.
3. "Systematic interpretation": Systematic interpretation is based on the ideal (!) of an in itself consistent legal system. Each legal norm is thus "to be interpreted only from its position and function within the complete legal system". On European level the relevant law/contract itself, the overall legal order or other international treaties as well as distant influences such as a constitution, the Charter of fundamental human rights, etc. can be taken into account.
4. "Teleological interpretation": Moreover, the category "intent and purpose" includes three further subcategories: the teleological interpretation, the efficiency of the protection (Art. 33 para. 4 VCLT) as well as the (judicial) development of the law. It is controversial if the teleological interpretation is a mean of interpreting a norm or the goal of the interpretation itself. It brings up the question which objective (telos) is to be achieved by the legal norm? The decisive factor is not the historical intention of the legislator, but the objective purpose expressed in the norm. The objective of the norm is characterized significantly by the wording, the systematic and the history (means of interpretation). Regarding the ECtHR the teleological interpretation is specified as a „dynamic" or „evolutive" interpretation. It takes into account the specialties of the ECtHR as a "living instrument, which must be interpreted in the light of present day conditions", i.e. gives the judges a bigger margin of appreciation.
5. "Comparative law": Legal situation in the Contracting parties/Legal situation in the EU/Autonomous definitions. References to the case law of other courts belong here as well.
Only the following categories (6.-9.) are to be used when there is a proportionality test. At the end, there is a decision of the ECtHR (cf. under C), which is to be annotated accordingly as "decision of the ECtHR". Occasionally, however, the category "application to the concrete case" may also be used for the legal basis (1.) as well as the legitimate purpose (2.). This depends on the respective individual case and must be assessed critically. For all other points (3.-4.) the categories specified here have to be used while the category "application to the concrete case" is never used.
6. "Test of the principle of proportionality - Legal basis": "In a constitutional democracy, a constitutional right cannot be limited unless such a limitation is authorized by law. This is the principle of legality. From here stems the requirement – which can be found in modern constitutions' limitation clauses, as well as in other international documents – that any limitation on a right be "prescribed by law". At the basis of this requirement stands the principle of the rule of law"
7. "Test of the principle of proportionality - Legitimate purpose": "The proper purpose component examines whether a law (a statute or the common law) that limits a constitutional right is for a purpose that justifies such limitation"
8. "Test of the principle of proportionality - Suitability": "The requirement is that the means used by the limiting law fit (or a rational connected to) the purpose the limiting law was designed to fulfill. The requirement is that the means used by the limiting law can realize or advance the underlying purpose of that law; that the use of such means would rationally lead to the realization of the law's purpose. It is therefore required that the means chosen be pertinent to the realization of the purpose in the sense that the limiting law increases the likelihood of realizing its purpose". The means used must at least further the achievement of the legitimate purpose.
9. "Test of the principle of proportionality - Proportionality": Since the ECtHR – in contrast for instance to the Federal Constitutional Court – does not strictly differentiate between the categories of necessity and proportionality in a strict sense, considerations of necessity – if present – are annotated within this category. "The next component of proportionality is the necessity test. It is also referred to as the requirement of "the less restrictive means". According to this test, the legislator has to choose – of all those means that may advance the purpose of the limiting law – that which would least limit the human right in question". The suitable means must be necessary to achieve the legitimate purpose, that is the least restrictive of all equally effective means available. "According to proportionality stricto sensu, in order to justify a limitation on a constitutional right, a proper relation ("proportional" in the narrow sense of the term) should exist between benefits gained by the public and harm caused to the constitutional right from obtaining that purpose. This test requires a balancing of the benefits gained by the public and the harm caused to the constitutional right through the use of the means selected by law to obtain the proper purpose". In an assessment of the benefits of the measure and the impairment of the affected persons, it must be determined whether the applied measures are appropriate, meaning reasonable for the persons concerned.
10. "Overruling": Overruling is referred to the (re-)adjustment of a precedent on a horizontal level, only under the premise of fundamental deficits of the previous precedent.
11. "Distinguishing": Distinguishing happens if an essential difference of facts is assessed by the judges, which leads to a non-transfer of a precedent to the new case.
12. "Margin of appreciation": The margin of appreciation is a margin of discretion granted by the ECtHR to the judiciary, legislature and executive of the Member States before a violation of the ECtHR is assumed.
13. "Precedents of the ECtHR": Binding effect of the legal content of earlier judgments of the ECtHR for later judgments. Only decisions (of all kind: GC, Chamber, Committee, Commission) of the ECtHR itself belong in this category.
14. "Application to the concrete case": Determination of the relation between the concrete case and the abstract legal norm. Subsumption of the facts of the case under a legal norm, i.e. examination whether the offence is fulfilled and the legal consequence thereby triggered.
15. "Decision ECtHR": The final sentence of the interpretation of a norm as well as the final sentence of the part of the judgment on the application to the concrete case may be concerned. If a section on the application of the Convention to the concrete case presents a pure reproduction of the facts – even though this extends over several paragraph - this is also marked as "application". This category is generally to be understood broadly.

Figure 4: Original LAM annotation guidelines from prompt augmentation in Llama experiments