Can Abstract Meaning Representation Facilitate Fair Legal Judgement Predictions?

Supriti Vijay[*,1] and **Daniel Hershcovich**[2],
[1]Adobe, India
[2]Department of Computer Science, University of Copenhagen
supriti.vijay@gmail.com, dh@di.ku.dk

Abstract

Legal judgment prediction encompasses the automated prediction of case outcomes by leveraging historical facts and opinions. While this approach holds the potential to enhance the efficiency of the legal system, it also raises critical concerns regarding the perpetuation of biases. Abstract Meaning Representation has shown promise as an intermediate text representation in various downstream NLP tasks due to its ability to capture semantically meaningful information in a graph-like structure. In this paper, we employ this ability of AMR in the legal judgement prediction task and assess to what extent it encodes biases, or conversely, abstracts away from them. Our study reveals that while AMR-based models exhibit worse overall performance than transformer-based models, they are less biased for attributes like age and defendant state compared to gender. By shedding light on these findings, this paper contributes to a more nuanced understanding of AMR's potential benefits and limitations in legal NLP.

1 Introduction

Transformer-based language models such as BERT, T5, and GPT have ushered in a new era in NLP. These language models have demonstrated exceptional proficiency in comprehending text with their non-trivial degree of knowledge in every field, propelling them to the forefront of various language-related domains (Chalkidis, 2023). However, despite their impressive performance, language models still face challenges in dealing with context-dependent language, biases in data, and a lack of interpretability (Thakkar and Jagdishbhai, 2023). Such limitations make them unsuitable for domains like legal NLP, which have an abundance of complicated, lengthy, and contextual legal documents. Thus, a system that can capture the intricate semantics of these documents is needed. Semantic representation frameworks have proven to be a promising solution, as they allow for a more nuanced understanding of language and can capture the complex relationships between legal concepts (Abend and Rappoport, 2017; Žabokrtský et al., 2020). Abstract Meaning Representation (Banarescu et al., 2013), one such framework, represents sentence-level meaning in a directed graph-based structure, with nodes representing concepts and edges representing relationships between them. This allows for a more accurate and comprehensive analysis of legal language, which is crucial in fields such as criminal and contract law, where the slightest of ambiguities can have significant consequences. However, limited knowledge exists about how useful these representations are in legal judgement prediction and whether they capture cultural and societal biases along with significant information.

This paper conducts a theoretical analysis of Abstract Meaning Representation, scrutinizing its potential in the realm of law. More concretely, it investigates the critical question of whether AMR can help produce fair legal decisions and reports

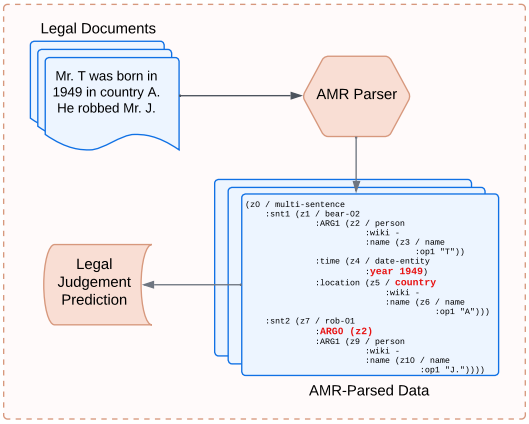

Figure 1: Abstract Meaning Representation in legal judgement prediction (LJP). Here, we demonstrate how AMR parses sensitive attributes like *age, gender identity and defendant state* as well as its ability to resolve co-referential pronouns like *he*, abstracting away gender.

[*]Work done while at Manipal Institute of Technology

Proceedings of the Fifth Workshop on Insights from Negative Results in NLP, pages 101–109

potential biases that may arise from its use. This evaluation helps us determine if AMR is a suitable intermediate representation for legal judgement prediction. We conduct our experiments on the ECtHR Dataset (see Section 4.1), a benchmark for legal judgement prediction which has been annotated with demographic and diversity labels. It proves to be a primary choice due to its inclusion of these labels for fairness evaluation. We utilize the macro F1 score as an evaluation metric for our experiments.

Contributions. We compare AMR's performance parity across different attributes of the ECtHR dataset, including age, gender identity, and defendant state. Our findings reflect that AMR is unable to produce fair outcomes and acts as a random baseline here. While it does report less group disparity for demographic attributes like age and state, it exhibits a low overall performance and a lower worst-case performance. We also release AMR-based models (LegalBERT and DistilRoBERTa) to enable further exploration of AMR in the legal domain.[1]

2 Related Work

While previous research has predominantly focused on AMR parsing of legal documents (Trong and Le, 2018; Vu et al., 2022; Dias et al., 2022), limited attention has been given to assessing AMR's performance in legal tasks. A study by Schrack et al. (2022) explores AMR's ability to identify logical relationships in legal MCQA tasks, revealing challenges posed by AMR parsing. In contrast, our work is the first to investigate whether AMR representations capture social biases alongside linguistic information, emphasizing the need to scrutinize AMR input representations for potential biases in legal judgement prediction tasks.

Research on fairness in machine learning models within the legal domain has also been limited. Previous studies (Angwin et al., 2016; Rice et al., 2019; Wang et al., 2021; Baker Gillis, 2021; Gumusel et al., 2022; Matthews et al., 2022; Wu et al., 2020) have highlighted racial and gender biases in the legal system and language models. More recently, Chalkidis et al. (2022) introduced the FairLex benchmark to assess the fairness of language models. In our study, we leverage one of these datasets to examine whether AMR-based models can effectively mitigate bias, addressing the critical issue of bias reduction in legal language processing.

[1] https://github.com/SupritiVijay/AMR-for-Legal-AI.

3 Abstract Meaning Representation

Abstract Meaning Representation is a structured framework that utilizes graph-like structures to represent sentence meaning, ensuring interpretability for machines and humans. These graphs, conforming to rooted, directed, and acyclic properties, are independent of semantics, grounded in syntax, and annotated using PENMAN notation for textual representation.

For example, the sentence *"Mr. T was born in 1949 in country A. He robbed Mr. J."*, as shown in Figure 1. Here, the sentence can be seen divided into two sub-sentences *(snt1 and snt2)*. In snt1, the event of "being born" *(bear-02)* is associated with Mr. T along with the : *time* and : *location* of birth. While in snt2, the event of "robbing" *(rob-01)* is described. Here, AMR can be seen establishing relationships between entities, connecting Mr. T to both the birth and the act of robbery.

4 Experimental Setup

4.1 Dataset and Metrics

The European Court of Human Rights (ECtHR) dataset (Chalkidis et al., 2021) is a text classification dataset annotated with multiple labels, which map human rights articles potentially violated in each case. It contains 11k legal cases and judgements, which are split into training (9k, 2001–16), development (1k, 2016–17), and test (1k, 2017–19) sets. Additionally, it includes distinct group tags like age, gender and defendant state for each case (See distribution in Appendix A.1). Due to its large sample size, diverse legal texts, and broad attribute coverage, ECtHR is ideal for assessing bias in AMR-based legal judgment prediction.

For a fair comparison with prior work, we adopt the same metrics used by Chalkidis et al. (2022). These include the average macro-F1 score ($mF1$), the group disparity (GD), and the worst-group performance ($mF1_W$). The $mF1$ represents the average macro-F1 score across different groups, providing a comprehensive measure of algorithm performance. The GD is calculated as the group-wise standard deviation, indicating the extent of disparity among the groups. Additionally, the worst-group performance ($mF1_W$), represents the lowest macro-F1 score among the individual groups. This allows us to gauge how poorly the most biased groups may perform.

4.2 AMR Parsing

AMR parsing has been considered a significant bottleneck (Schrack et al., 2022), especially concerning the loss of information in long and multisentence paragraphs. Hence, to overcome this challenge, we utilize the following two pre-processing techniques for our experiments.

1. Splitting before parsing (SbP): This approach involves splitting each case in the dataset into sentences before parsing, resulting in single-sentence graphs, as shown in Figure 3. These graphs are then combined to form a multi-sentence graph for a case. While this approach offers advantages in scalability, it may have limitations in terms of maintaining coherence across paragraphs.

2. Splitting after parsing (SaP): In contrast, this alternative approach focuses on creating multi-sentence graphs first, which are then linearized and split into 512 token segments to be sent to the encoder. These graphs capture interdependencies and connections between sentences, enhancing their richness compared to pooling single-sentence graphs. However, it may require more computational resources and time, as illustrated in Table 3.

We utilise the `SpringAMR` parser (Blloshmi et al., 2021) for parsing documents due to its strong and robust parsing quality. It employs a simple Seq2Seq architecture employing a pretrained BART model, trained on the Text-to-AMR task. We further explore the above techniques quantitatively and qualitatively in Appendix C.1 & C.2.

4.3 Baselines

To classify AMR-parsed graphs, we adopt a hierarchical BERT-based architecture similar to Chalkidis et al. (2022), which has been established as the benchmark model for fairness evaluation in legal datasets. This architecture effectively captures the contextual dependencies in legal documents by giving utmost attention to both paragraph and document-level representations. A detailed explanation of fine-tuning the models can be found in Appendix B. Further, we also reproduce the results of the hierarchical architecture with text-only input to evaluate the performance of AMR-based techniques in the subsequent experiments.

4.4 AMR-based models

We utilize `legalbert-base-uncased` and `distilroberta-base`, classifiers trained on textual data, as the primary models in the hierarchical architecture. We also execute continued pre-training on AMR graphs to enhance the performance of transformer models, specifically LegalBERT. We name this model as Dataset-specific LegalBERT$_{SMALL}$. Through this, we examine whether pre-training on AMR graphs captures intricate structural and semantic intricacies inherent to legal language and performs better than other classifiers. We utilize the LegalBERT model as the backbone for pretraining. This model is pre-trained using the ECtHR training subset, employing a sequence length of 128 sub-words for 10 epochs. The AdamW optimizer is used with a maximum learning rate of $1e-4$ and a 10% warm-up ratio.

5 Result Analysis

5.1 Dataset-specific vs Basic Models

In this subsection, we compare the performance of dataset-specific LegalBERT and basic LegalBERT within AMR SaP. The mF1-scores in Table 1 show a significant performance decline with pre-training, attributed to introduced noise and biases inherent in the dataset. In contrast, the basic LegalBERT model, which is trained directly on the specific legal classification task without the additional step of pre-training, can solely focus on learning from the task-specific data. Additionally, we observe that a generalized adaptation to legal knowledge may be more effective than attuning a pre-trained model on the experimental dataset. The vast overview of legal knowledge assists the basic model in acquiring a strong foundation in legal language understanding, allowing it to outperform the dataset-specific model.

5.2 Fairness Analysis

Analysing the results presented in Table 1, it becomes evident that the benchmark DistilRoBERTa$_{FairLex}$ model displays notable group disparities, particularly for Defendant State and Applicant Age. In contrast, most AMR-based models exhibit reduced group disparities in these attributes. However, when it comes to Applicant Gender, AMR-based models consistently demonstrate higher group disparities, with LegalBERT$_{SMALL}$ (AMR SbP) recording the highest GD for it. This phenomenon may be attributed to the parsing of individual sentences, assigning equal weight to all words, including gendered ones, potentially perpetuating implicit biases within the model. In the broader context, we identify a recurring trend where AMR-based

ECtHR (ECHR Violation Prediction)										
Language Models	Average mF1	*Defendent State*			*Applicant Gender*			*Applicant Age*		
		mF1 ↑	GD ↓	$mF1_w$ ↑	mF1 ↑	GD ↓	$mF1_w$ ↑	mF1 ↑	GD ↓	$mF1_w$ ↑
Text Based Models										
DistilRoBERTa	62.9	63.3	2.1	61.2	59.0	2.0	56.3	61.3	2.5	58.5
DistilRoBERTa$_{FairLex}$	NA	53.2	8.3	44.9	57.5	3.1	54.4	54.1	5.9	46.2
AMR Split before Parsing										
LegalBERT$_{SMALL}$	54.8	50.5	1.2	49.3	47.1	5.4	40.4	52.4	4.8	47.2
AMR Split after Parsing										
LegalBERT$_{SMALL}$	57.3	**59.2**	**0.3**	**58.8**	**56.0**	3.5	**52.3**	**56.5**	3.7	**50.1**
(Dataset-specific LegalBERT$_{SMALL}$)	44.2	40.4	5.3	35.0	32.1	**2.5**	28.9	33.3	**0.8**	31.9
DistilRoBERTa	37.6	36.5	0.7	35.7	31.6	4.4	28.3	36.2	5.4	27.6

Table 1: Test results for different baselines and models per ECtHR attribute. We report the average performance across groups (mF1), the group disparity (GD), and the worst-group performance ($mF1_w$). ↑ denotes that higher scores are better and ↓ vice versa. We report results by Chalkidis et al. (2022) as DistilRoBERTa$_{FairLex}$.

models exhibit higher fairness levels compared to text-based models. However, this advantage is offset by lower $mF1$ scores and overall performance metrics. Notably, a subset of AMR-based models, primarily LegalBERT$_{SMALL}$ (AMR SbP), approaches the performance of text-based models but lacks consistency in addressing group disparities across all attributes.

Digging deeper into worst-case performance, we notice that while AMR models inherently prioritize fairness, their lower worst-case performance scores render them impractical for real-world applications. This raises a crucial question: *does a model with greater fairness, at the cost of overall performance, hold value?*. In essence, a model with zero performance yields zero group disparity. This brings to light a paradox: the fairness demonstrated by AMR models, despite having low group disparity, takes on the semblance of a random baseline due to its lack of substantial performance metrics. Consequently, we assert that AMR may not be the optimal choice for ensuring fairness in practice.

5.2.1 Potential Biases

As illustrated in Table 1, we observe that AMR-based models demonstrate lower group disparity than the benchmark DistilRoBERTa$_{FairLex}$ model for defendant state and applicant age and higher group disparity for Applicant Gender. This could be attributed to the fact that other group identifiers, such as defendant state and age, may not be directly linked to the individual during AMR parsing.

For example, the sentence "Mr. T was born in 1949 in country A. He robbed Mr. J." as represented in Figure 1. Here, the accurate recording of the applicant's country (location-$z5$) and year (time-$z4$) establishes a direct link with *:ARG1-*$z1$, while coreference in $z7$ is directly associated with *:ARG0-*$z2$. This distinction implies that while coreferences consistently refer to the individual, contextual details such as time and location are connected to the event itself. Consequently, the presence of pronouns in the case establishes a direct relationship between the gender and personal information of the individual. This dissociation between these contextual elements and the individual prevents the subsequent classification model from making inferences based on these attributes. As a result, age and defendant state exhibit lower group disparity, while gender disparity remains consistent throughout the analysis.

6 Conclusion

In this paper, we explore the application of Abstract Meaning Representation (AMR) in predicting legal judgments. Our analysis has revealed both the benefits and challenges associated with using AMR in this context. While AMRs offer the capability to capture the semantics of legal texts and enable automated analysis and decision-making, providing a promising avenue for fair judgement still remains ambiguous in domains like Applicant Gender. Even so, it clearly demonstrates its efficacy in other group disparities like Age and Defendant State. However, due to their poor performance and low mF1 scores, we conclude that while AMR-based models are fairer by design, they are unsuitable for ensuring fairness in the real world.

Limitations

We experimented with one AMR parser (with two sentence-splitting strategies), `SpringAMR`. While this is a widely used and highly accurate AMR parser, other parsers might exhibit different behavior with respect to encoding demographic attributes such as those we investigate here.

Furthermore, while AMR is the most popular meaning representation framework, other meaning representation frameworks may again behave differently. For example, UCCA (Abend and Rappoport, 2013) represents semantic structure without attempting to capture lexical disambiguation at all.

Finally, we only investigated one of the datasets included in FairLex, namely ECtHR, targeting the age, defendant state and gender attributes. Different conclusions may be drawn regarding other datasets, tasks and attributes—for example, the SCOTUS dataset indicates whether the respondent is a person, public entity, organization, facility or other. FSCS contains the language and region of the case. Further investigation is required to better understand and address the limitations of what is represented in the parsed AMRs and what is not to ensure fair and accurate predictions across all demographic groups.

Ethics Statement

Automating legal judgement prediction raises ethical implications and warrants a thorough examination of potential biases. Our AMR-based models have shown promising improvements in group disparity. However, the parsed AMR may nevertheless unintentionally overlook or misrepresent certain group identifiers, leading to biased predictions we are not yet aware of. Furthermore, the remaining performance disparities observed across demographic groups, particularly in Applicant Gender, highlight the need for continuous evaluation, improvement in fairness considerations and stronger guarantees before deploying such models in legal contexts.

The ECtHR dataset is released as part of FairLex under the CC-BY-NC-SA-4.0 license. We only use it for our experiments and do not redistribute it. Furthermore, the original dataset is anonymized, and we do not add any new data—particularly no personal information.

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

Acknowledgements

References

Omri Abend and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.

Omri Abend and Ari Rappoport. 2017. The state of the art in semantic representation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 77–89, Vancouver, Canada. Association for Computational Linguistics.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica*, 23:77–91.

Noa Baker Gillis. 2021. Sexism in the judiciary: The importance of bias definition in NLP and in our courts. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 45–54, Online. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Rexhina Blloshmi, Michele Bevilacqua, Edoardo Fabiano, Valentina Caruso, and Roberto Navigli. 2021. SPRING Goes Online: End-to-End AMR Parsing and Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 134–142, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilias Chalkidis. 2023. ChatGPT may pass the bar exam soon, but has a long way to go for the LexGLUE benchmark.

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.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard.

2022. FairLex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, Dublin, Ireland. Association for Computational Linguistics.

Martha Chamallas. 2019. Feminist legal theory and tort law. In *Research Handbook on Feminist Jurisprudence*, pages 386–405. Edward Elgar Publishing.

João Dias, Pedro A. Santos, Nuno Cordeiro, Ana Antunes, Bruno Martins, Jorge Baptista, and Carlos Gonçalves. 2022. State of the art in artificial intelligence applied to the legal domain.

Ece Gumusel, Vincent Quirante Malic, Devan Ray Donaldson, Kevin Ashley, and Xiaozhong Liu. 2022. An annotation schema for the detection of social bias in legal text corpora. In *Information for a Better World: Shaping the Global Future: 17th International Conference, iConference 2022, Virtual Event, February 28–March 4, 2022, Proceedings, Part I*, pages 185–194. Springer.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR.

Sean Matthews, John Hudzina, and Dawn Sepehr. 2022. Gender and racial stereotype detection in legal opinion word embeddings.

Douglas Rice, Jesse H. Rhodes, and Tatishe Nteta. 2019. Racial bias in legal language. *Research & Politics*, 6(2):2053168019848930.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Jackson Sargent and Melanie Weber. 2021. Identifying biases in legal data: An algorithmic fairness perspective.

Nikolaus Schrack, Ruixiang Cui, Hugo López, and Daniel Hershcovich. 2022. Can AMR assist legal and logical reasoning? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1555–1568, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Krishna Thakkar and Nimit Jagdishbhai. 2023. Exploring the capabilities and limitations of gpt and chat gpt in natural language processing. *Journal of Management Research and Analysis*, 10:18–20.

Sinh Vu Trong and Minh Nguyen Le. 2018. An empirical evaluation of AMR parsing for legal documents.

Sinh Trong Vu, Minh Le Nguyen, and Ken Satoh. 2022. Abstract meaning representation for legal documents: an empirical research on a human-annotated dataset. *Artificial Intelligence and Law*, pages 1–23.

Yuzhong Wang, Chaojun Xiao, Shirong Ma, Haoxi Zhong, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2021. Equality before the law: Legal judgment consistency analysis for fairness. *arXiv preprint arXiv:2103.13868*.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-biased court's view generation with causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–780, Online. Association for Computational Linguistics.

Zdeněk Žabokrtský, Daniel Zeman, and Magda Ševčíková. 2020. Sentence meaning representations across languages: What can we learn from existing frameworks? *Computational Linguistics*, 46(3):605–665.

A Fairness in Legal Judgement Prediction

The legal domain represents a complex and multifaceted system shaped by various social, cultural, and historical factors. Portrayed as blind, unbiased, and objective, justice is often plagued by systemic biases ingrained in the language of judicial opinions, case outcomes, and the personal predispositions of its practitioners (Rice et al., 2019). While for NLP in law, these biases manifest in either representational harms where certain social groups are over or underrepresented or sentencing disparities across certain groups (Sargent and Weber, 2021). In our evaluation of fairness, we adopt an *equal risk or equal odds* (Hashimoto et al., 2018) approach where we define bias as the disproportionate performance of a classifier across different groups with similar risk profiles. Such parity conclusively establishes sensitive traits like *age, nationality, and gender* as significant attributes when forming an outcome. Therefore, we embrace this asymmetry in efficacy as a measure of fairness across input representations in the legal judgement prediction domain.

For instance, victims of domestic violence, rape, and sexual assault have little recourse to obtain tort compensation due to the installation of recovery restrictions (Baker Gillis, 2021; Chamallas, 2019). This is merely one situation where failing to provide equal weight to all genders in the law results in severe damage.

A.1 FairLex & the ECtHR dataset

We use prior work conducted under FairLex (Chalkidis et al., 2022) as our baseline for text

Applicant Age				*Applicant Gender*			*Defendant State*	
N/A	>35	<65	>65	N/A	Male	Female	E.C.	West
2,794	839	4,246	1,121	3,306	4,407	1,287	7,224	1,776

Table 2: Group distribution in training set for each attribute of ECtHR dataset. These are the statistics presented in the FairLex paper (Chalkidis et al., 2022).

classification and fairness. The study partitions the ECtHR dataset on the following attributes:

1. **Defendant States:** These comprise European nations accused of breaching the ECHR. Each case's defendant states form a subset of the 47 Council of Europe Member States. To establish statistical significance, the defendant states are categorized into two groups: Central-Eastern European states and all other states, as delineated by the EuroVoc thesaurus.

2. **Applicant's Age:** The applicant's birth year is gleaned from case facts whenever possible, leading to classification within age groups ($\leq$ 35, $\leq$ 64, or older).

3. **Applicant's Gender:** Extracted from case details, gender is categorized as male or female based on pronouns or other gender-specific terminology. We will add these attribute distributions to the dataset description as well.

B Problem Formulation

In this section, we introduce the notations used for the task of predicting legal judgments. Let $(X_i, Y_i)_{i=1}^{N}$ represent a training set comprising N samples. Each sample consists of an input list of facts denoted as $X_i = \{t_1, t_2, ..., t_m\}$, pertaining to a single legal case. To capture the semantic and relational nature of the text, we feed these text paragraphs into an AMR parser, which generates the respective graphs, i.e., each t_j creates its own encoded graph f_j. Therefore, if initially each sample was represented by $X_i = \{t_1, t_2, ..., t_m\}$, where each X_i was an entire legal case and each t_j were its individual facts, after encoding by AMRs, they can be represented as $X_i = \{f_1, f_2, ..., f_m\}$. With this, we have restructured the problem statement as judgement prediction using AMR-graphs. The corresponding labels for the multi-label classification task are represented by $Y_i = \{y_1, y_2, ..., y_{10}\}$. Our objective is to maximize the posterior probability $p(Y|X)$ for each case. However, due to the presence of lengthy textual content within each case and the inherent token limit of transformer-based language models, we adopt a hierarchical approach to address this challenge.

	Parsing Time (seconds)	*Average No. of Tokens (case)*
Split Before	444960	47387.96
Split After	648000	68439.15

Table 3: Statistics for the two parsing strategies: sentence splitting before/after parsing.

This architecture uses a transformer-based backbone model, such as LegalBERT (Chalkidis et al., 2020) or DistilRoBERTa (Sanh et al., 2020), to generate embeddings for each fact (f_k) in the input. This enables us to obtain contextualized representations for each fact. Instead of using pooling techniques at the word level, we consider the representation of the $[CLS]$ token as the fact embedding (e_k), capturing the global context of the entire fact. Subsequently, a segmentation-encoder layer is employed to process the fact embeddings ($E = \{e_1, ..., e_k, ..., e_m\}$) and capture the longform structure of the legal case. This layer combines the fact embeddings using attention weights, generating a multi-vector representation for each fact in the case ($SE = \{se_1, ..., se_k, ..., se_m\}$). These representations are then pooled and fed into a classification layer to generate the probability (p) of a violation (Y) given the input (X).

C AMR Parsing

C.1 Quantitative Analysis

We compare the length of parsed strings using two AMR parsing techniques, "Splitting before parsing" (X-axis) and "Splitting after parsing" (Y-axis), as shown in Figure 2. The plot illustrates a significant difference, with a distinct upper bound on the Y-axis (1.4M characters) and a lower bound on the X-axis (391k characters), taking into account characters and whitespaces. This trend persists even after removing whitespaces using regex, indicating that "Split After" consistently results in longer strings. Additionally, we compute statistics on depth and the

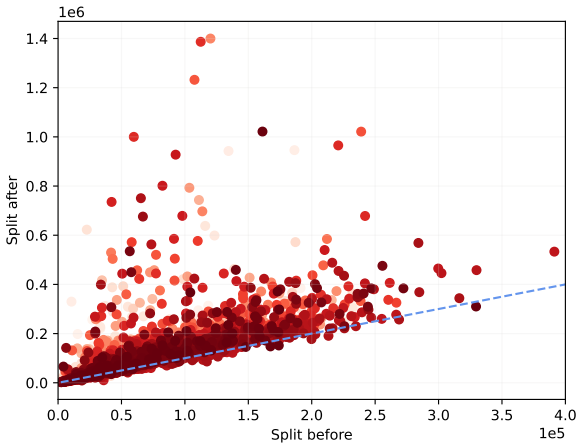

Figure 2: A scatter plot depicting the length of parsed strings using AMR parsing techniques, "Splitting before parsing" (X-axis) and "Splitting after parsing" (Y-axis), reveals a noticeable difference between them. The plot shows a wider dispersion of data points on the Y-axis. Here, length of parsed strings refer to the number of characters in the entire case, i.e., all linearized graphs that are concatenated together.

number of relations. The average depth for "Split Before" is 17.76, while for "Split After," it is 44.81, suggesting higher structural complexity in the latter. Likewise, "Split After" exhibits an average number of relations at 44.90, compared to 17.87 for "Split Before," indicating more interactions in the former.

Additionally, the scatter plot demonstrates that the "Splitting after parsing" technique exhibits a wider dispersion of data points on the Y-axis, indicating its ability to retain a more significant amount of knowledge. These findings highlight the effectiveness of the "Splitting after parsing" technique in capturing more information.

C.2 Qualitative Analysis

In this section, we study different techniques of parsing from the perspective of structure, coreference, and context retention. The first technique, "Splitting before parsing," offers scalability, although it also limits context understanding and coherence across paragraphs. For instance, as shown in Figure 3, individual sentences may not capture the associations between entities, leading to a lack of comprehensive analysis. Furthermore, we observe that while splitting a paragraph into component sentences, certain short phrases enclosed between two periods tend to be skipped. In the example presented in Figure 3, person "T." can be seen eliminated during graph generation. We have validated these errors in "Splitting before parsing" method using a naive approach of '.' detection, as well as, using NLTK's $sentence - splitter$. The issue of co-reference still persists across both splitters as expected.

In contrast, the second technique, "Splitting after parsing," retains entity and event-coreference and maintains a stronger connection to the original context. Here, splitting is based on the limitation provided by the LLM model, since we are using $BERT$, it is the $max - tokens$ which can be fed into that model. This, allows the graphs to strongly associate and encode large amounts of text data, including their co-references irrespective of the sentence structure. Upon feeding it further for classification, since we use a HAN architecture it continues to carry-forward the same co-references in its predictions. Therefore, as demonstrated in Figure 3, the multi-sentence graph represents the same content but with a different organization, capturing diverse associations and temporal relationships. It is able to better capture the interrelation between the individuals involved, the event, and the timing of the event. This technique contributes to more accurate parsing results and a deeper understanding of legal entities and their relations.

While our findings suggest the "Splitting after parsing" method is a more effective parsing strategy for AMR graphs, we still witness occasional oversights by the approach. Such as the graph on the left (*split before*) uses the same variable $z0$ for the person "J.", the action of placing, and the action of visiting. This is incorrect as they are distinct entities or events. The person "T." who visited is not represented in the graph. The graph does not capture that both the placing and the visiting happened on the same day, 23 June 1993. The graph uses ($z1$ / she) to represent "her," but it's not clear that "her" refers to "J". The graph separates the events of placing and visiting into different sub-graphs but does not establish any relationship between them. Also, the date "23 June 1993" is associated only with the person "J." and not with the events of placing and visiting. The graph on the right (*split after*) uses a single variable $z1$ to represent both "J." and "T." under : $name$. This is incorrect as they are distinct entities. While the graph includes the date entity $z6$, it is only linked to the $place - 01$ event. It should also be linked to the $visit - 01$ event to indicate that both events happened on the same day. The graph still does not make it clear that "her" refers to "J." Coreference should be explicitly represented.

31. On 23 June 1993 J. was placed in the family centre. T. visited her the same day.

Split-Before

```
(z0 / person
    :wiki -
    :name (z1 / name
              :op1 "J")
    :time (z2 / date-entity
              :day 23
              :year 1993))
(z0 / place-01
    :ARG2 (z1 / center
              :mod (z2 / family)))
(z0 / visit-01
    :ARG1 (z1 / she)
    :time (z2 / day
              :ARG1-of (z3 / same-01)))
```

Split-After

```
(z0 / visit-01
    :li 31
    :ARG0 (z1 / person
              :wiki -
              :name (z2 / name
                        :op1 "J."
                        :op2 "T."))
    :ARG1 (z3 / place-01
              :ARG1 z1
              :ARG2 (z4 / center
                        :mod (z5 / family))
              :time (z6 / date-entity
                        :year 1993
                        :month 6
                        :day 23
                        :time-of z0)))
```

Figure 3: AMR graphs, in PENMAN format, obtained through sentence splitting before (left) and after parsing (right), showing the differences in graph structure. In the former, sentence splitting errors result in an incorrect AMR. The latter results in an AMR with less severe errors, which also demonstrates cross-sentence co-reference resolution of the time expression. For distinction, we present segments of the image in red, which are clearly contrasted within the “Split-Before” side of the image. We see that "T.", "month 6", and "time-of z0" are better co-referenced and associated by the “Split-After” technique.