多言語での判例事実概要からの法的関係性のグラフ可視化

大南英理 [1]　宮西大樹 [2]　前田航希 [3,5]　栗田修平 [4,5]
[1]NAIST　[2] 東京大学　[3] 東京科学大学　[4] 国立情報学研究所　[5]NII LLMC
onami.eri.ob6@is.naist.jp　taiki.miyanishi@weblab.t.u-tokyo.ac.jp
koki.maeda@nlp.c.titech.ac.jp　skurita@nii.jp

概要

裁判官による判決作成では，争点の事実に関連する法律を解釈し事実を当てはめて結論を導く．他方モデルによる判例予測では，法律領域の精度向上とAI の信頼性観点での推論過程の明確化が課題である．本研究では，どのように法律解釈を行ったかを可視化し，モデルが法律解釈を学習することでドメイン知識を向上させるため，EU 法判例を用いて判例内の事実の法的関連性のグラフ生成タスクである LegalViz を提案する．更に，グラフ構造と法律内容の双方を考慮したグラフ可視化評価を行うため評価手法を提案する．実験では，学習したモデルが Few-shot 設定の GPT モデルの性能を上回ることから提案データセットの有効性を示した．

1　はじめに

大規模言語モデル（LLM）やマルチモーダルモデルの急速な発展により [1, 2]，法律や医療，数学等専門領域への自然言語処理の応用が注目される [3, 4, 5]．特に法律分野では従来専門家に担われていた業務の補助・効率化に LLM を活用できる可能性があり [6, 7]，判例予測 [8, 9, 10, 11] やデータセット作成 [12, 13, 14, 15] が研究されてきた．しかし，法律に関する文章は法的三段論法等の特殊な推論方法を用い，文章中に明示的に記載されない他の判例・法律を考慮するため，一般的な文章理解タスクと比べ活用が十分に進んでいない．中でも裁判所の判決は，裁判で争われる事実に関係する法律を解釈し法規範に事実を当てはめて結論を導くタスクを人間の裁判官が行っており，このタスクの LLM による効率化が期待される．この判例予測タスクでは，(1) 問題となっている事実に関する法律を解釈・適用し，事実を当てはめ結論を導くという推論方法が LLM によって十分に理解されておらず，(2) 信頼される AI の観点から，推論過程でなぜその結論が出たかの理由付けがなくブラックボックス化してしまう課題があった．他方 LLM の性能向上に伴い，テキストからコード生成するタスク [16, 17] の研究が行われ，GPT-4 のようなモデルは一定の図表描画も可能とされる [18]．本研究では，従来裁判官により行われた判決の事実を元に関連する法律を解釈し事実を当てはめる作業を，LLM によるグラフ生成タスクとして定式化し，事実関係と解釈結果を可視化することにより，(1)LLM が法律の解釈・適用の推論精度を向上させ，同時に (2) 判決の判断過程を明確化することを目的とし，判例の事実概要を入力として法律上の関係性をグラフ可視化する 7,010 件のデータセット **LegalViz** を提案する．LegalViz は EUR-LEX で公開される 23 言語の EU 法判例を対象とし，判例の事実概要を入力として，(a)法主体，(b) 契約等の法律上の関係性，(c) 判例で解釈対象となった法規範，(d) 判例の法律解釈に関連する事実・主張の要約の **4 つの観点**から，DOT 言語である Graphviz コードを出力と判例の法律上の関係性をグラフ可視化する．更に，グラフ生成と法律内容理解の 2 つの側面から生成したグラフの評価を行うため，グラフ構造と法律内容の 2 つの評価手法を提案する．

2　LegalViz データセット

2.1　タスク

EU 法判例の事実概要を入力として，法律上の関係性をグラフ可視化した Graphviz コードのようなグラフ記述言語 (DOT 言語) を出力するタスクを導入する．図 1 に判例の事実関係箇所を入力として，モデルがグラフ可視化コードを出力する例を示す．

2.2　データセット作成

テキスト収集　EUR-LEX は EU 法の判例・命令・裁判官個別意見等を EU が公式に公開するウェブサ

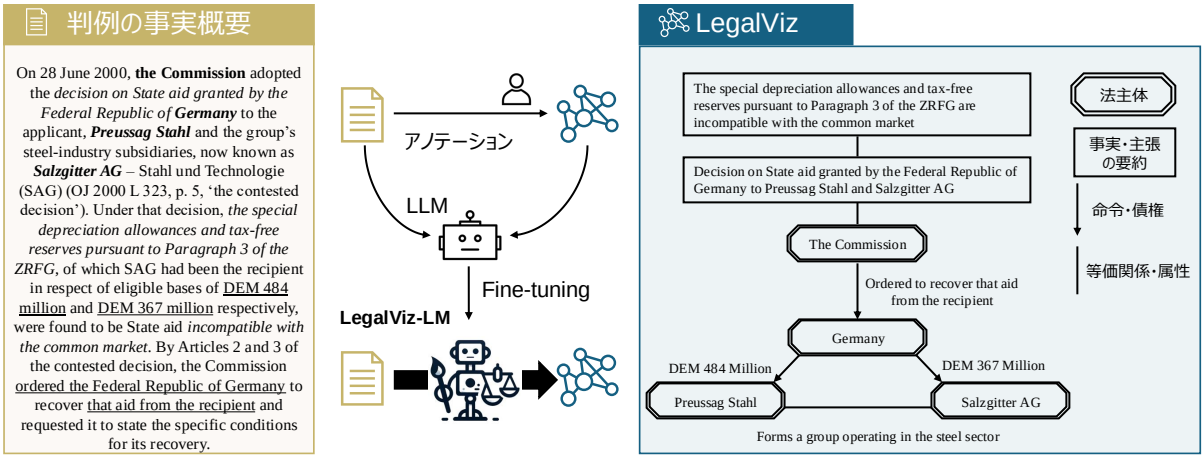

図 1: LegalViz データセットの入力テキスト (左) と出力グラフ (右) 例

イトであり，加盟国全ての言語で翻訳を提供する．ここでは主に過去 20 年を対象とし，EUR-LEX に掲載される全ての言語で判例を取得した．

グラフ可視化コード作成　(1) まず，収集した判例の事実概要部分を対象として，法律関係の図示が可能な単位で短いパラグラフに区切り，各判例の全ての翻訳版で事実概要を作成した．(2) 次に，各判例の英語版を対象に，作成した事実概要から下記のアノテーション規約に従って，DOT 言語でグラフ可視化した．各ノードの形はアノテーション規則で詳細に定義し，DOT 言語コードで形を指定した．(3) 最後に，23 言語でデータセットを作成するため DOT 言語コードを英語から他の 22 言語へ翻訳した．

アノテーション規約　DOT 言語による法律関係の可視化アノテーションでは，視覚的に法律上の意味を把握することを可能にし，モデルが法律上どの意味でノードとエッジを認識しているか確認するため，ノードとエッジの形状を法律上の意味毎に定義する．主なアノテーションの種類は，法主体，判例で解釈対象となった法規範，判例の法律解釈に関連する事実・主張の要約をグラフのノードとし，契約・権利義務等の法律上の関係性の 4 点である．法律関係は，法律上の権利を行使できる法主体同士を契約や金銭債権請求等の法律上の関係性で結び図示でき，更に判例で法規範をどのように解釈し事実を当てはめたかを法規範と事実・主張の要約で記載するためこの 4 点でグラフを構成した．各構成要素は法主体を八角形，法規範を台形，事実・主張の要約を四角形のノードで表記した．更に法律上の関係性ではより詳細に，契約関係を一重実線，家族関係を二重実線，相続を一重点線，一方の法主体から他方への債権義務を実線矢印で表した．

3　評価手法

アノテーションされた可視化グラフとモデルにより生成された可視化グラフを，グラフ構造や内部のテキストを考慮し比較する評価手法を導入する．具体的には，(1) 生成された DOT 言語コードに文法エラーがなくグラフ描画できることを確認し，(2) 生成グラフと参照グラフのノード中のテキスト類似度から二部グラフマッチングにより対応するノードを決定し，(3) 対応するノード・エッジのテキスト類似度からグラフの類似度を決定する，という順序を取る．この際に，グラフ構造の正しさのみに着目して，対応するノード間のエッジの一致のみを評価する（`Graph`），加えて対応するノードのテキスト類似度を考慮する（`Graph&Node`），さらにエッジのテキスト類似度をも考慮する（`Graph&Node&Edge`）の 3 種類の評価手法を導入する．

3.1　2 つのグラフの類似度比較

$\mathcal{G}_r$ を参照グラフ，$\mathcal{G}_h$ をモデルによる生成グラフとする．グラフは，エッジ集合 E とノード集合 V から構成され，あるエッジ $e \in E$ が開始ノード v_s と終点ノード v_e をつなぐとき，タプル $e = [v_s, v_e, l]$ と表される．ここで，v_s，v_e はこの有効エッジの始点ノードと終点ノードを表す．ここで l はエッジに付与された説明テキストを示す．なお，以降，参照グラフと出力グラフの各要素には右下添字にそれぞれ r，h を付与する．例えば，参照グラフのエッジの始点ノードと終点ノードは $v_{s,r}$ と $v_{e,r}$ とおく．

DOT 言語としての正しさの判定　まず，生成されたコードが DOT 言語として有効に解釈されるか

を pydot ライブラリ[1]により評価した．

二部グラフマッチングからのノード対応関係導出　次に，出力ノード $\{v_h\}$ と参照ノード $\{v_r\}$ をそれぞれグラフ $\mathcal{G}_h$，$\mathcal{G}_r$ から抽出し，ノード内のテキスト類似度 $s(v_r, v_h)$ をテキスト類似度関数 $s(v_r, v_h) = \mathrm{sim}(v_r, v_h)$ を基いて計算し二部マッチングを行った．ここでテキスト類似度関数には頑健で人手評価と相関する BERTScore [19] を用いた．二部グラフマッチングは NetworkX[2]を用いて行い，ノード間対応関係を取得した．

グラフ構造やラベル類似度による類似度評価　(1)`Graph` 評価では，マッチングしたエッジ同士の F1 スコア比較を行う．グラフ $\mathcal{G}_h$，$\mathcal{G}_r$ のエッジ集合をそれぞれ E_r，E_h，ノード集合を V_r，V_h とおくと，生成グラフから参照グラフへのノードの対応付け関数 $a(\cdot) : V_h \to \{V_r, \phi\}$ およびノードの一致を表すクロネッカーのデルタ記号 $\delta_{\mu\nu} = 1$ iif $\mu = \nu$ otherwise $\delta_{\mu\nu} = 0$ を用いて，参照グラフと生成グラフのエッジ e_h および e_r の一致を

$$f_{\mathrm{Graph}}(e_h, e_r) = \delta_{a(v_{s,h})v_{s,r}} \delta_{a(v_{e,h})v_{e,r}}$$

で計算する．なお，ノードの対応付け関数 $a(\cdot)$ には対応付けが存在しない（$v_h \xrightarrow{a(\cdot)} \phi$）ケースが有り得るが，任意のノード ν に対し $\delta_{\phi\nu} = 0$ とする．このバイナリ関数 $f_{\mathrm{Graph}}(e_h, e_r)$ により

$$\mathrm{TP} = \sum_{e_h \in E_h, e_r \in E_r} f_{\mathrm{Graph}}(e_h, e_r)$$

$$\mathrm{FP} = |E_h| - \mathrm{TP}, \qquad \mathrm{FN} = |E_r| - \mathrm{TP}$$

から F1 値を計算する．この評価によりグラフ構造の類似度を評価することができるが，ノード間対応関係以外に，ノードやエッジ間のテキスト類似度はここでは考慮していない．[3]

(2) `Graph&Node` 評価は，`Graph` 評価を拡張し，エッジの始終点ノードについて，参照グラフと出力グラフでのテキスト類似度を比較し，類似度が高い場合に高いスコアを与える．具体的には，TP を

$$\mathrm{TP} = \sum_{e_h \in E_h, e_r \in E_r} f_{\mathrm{Graph}}(e_h, e_r) \cdot \mathrm{sim}(v_{s,r}, v_{s,h}) \mathrm{sim}(v_{e,r}, v_{e,h})$$

と修正し，F1 値を計算する．

1）https://github.com/pydot/pydot
2）https://networkx.org/
3）なお，この `Graph` 評価は，ノードの対応付け $a(\cdot)$ が正しい場合，つまり V_r と V_h に正確な 1 対 1 対応があるとみなせる場合には，グラフ補完タスクなどで用いられる，グラフのエッジ検出の F1 評価尺度と同一のものと考えられる．

(3)`Graph&Node&Edge` 評価では加えてエッジテキストでも BERTScore による参照グラフと出力グラフのテキスト類似度を比較する．すなわち，TP を

$$\mathrm{TP} = \sum_{e_h \in E_h, e_r \in E_r} f_{\mathrm{Graph}}(e_h, e_r) \cdot \mathrm{sim}(v_{s,r}, v_{s,h}) \mathrm{sim}(v_{e,r}, v_{e,h}) \mathrm{sim}(l_r, l_h)$$

と修正し，F1 値を計算する．

3.2　法律観点からのグラフノード評価

モデルにより出力されたグラフは法律上の意味毎に異なる図形でグラフノードを出力しているため，最初に参照ノードと出力ノードでそれぞれ法主体(八角形)，法規範(台形)，事実・主張の要約(四角形) のノードを抽出し，ノードの形ごとにアラインメントを取得する．次に，法律上の関係性を表すエッジも契約関係・相続関係等で形が異なるため，エッジの形毎にアラインメントを取得する．最後にグラフ構造評価と同様にして，アラインメントをとった法主体，法律上の関係性，法規範，事実・主張の要約の 4 つの観点毎に，テキスト類似度を BERTScore で比較し，F1 スコアを計算した．

4　実験

複数のオープンソース LLM を用いて判例事実概要から法律的関係性を可視化するグラフの生成能力を測定を行い，LegalViz を用いて Fine-tuning を行ったモデルは Few-shot の GPT モデル性能を上回った．

4.1　実験設定

法律関係の可視化タスクを DOT 言語のコード生成タスク形式で行い，Few-shot 形式と Fine-tuning (instruction tuning) 形式の 2 つの形式で実験した．モデルは，公開モデルと OpenAI GPT APIs を使用した．公開モデルは，CodeLlama [20] と Llama 3.1 & 3.2 model [21] および Gemma 2-9B [22] を使用した．

4.2　結果

グラフ構造評価　表 1 の `Graph`，`Graph&Node`，`Graph&Node&Edge` の 3 つの評価は，グラフ・ノード・エッジラベルによるグラフ構造評価である．中でも，ノードとエッジのテキスト類似度とグラフそのもののテキスト類似度が全て評価される `Graph&Node&Edge` が最も難しい評価観点となる．表 1 より，LegalViz により Fine-tuning したモデルは，GPT モデルの方がモデルサイズが大きいと

表 1: Few-shot 及び Fine-tuning での各モデルのスコアを示す．**G**，**G-N**，及び **G-N-E** は，それぞれ `Graph`，`Graph&Node`，及び `Graph&Node&Edge` を意味する．**Top1** と **Top10** は，生成結果の上位 1 件と上位 10 件での有効なグラフ生成確率である．

	グラフ構造			有効なグラフ		法律観点			
モデル	**G**	**G-N**	**G-N-E**	**Top1**	**Top10**	**法主体**	**法律上の関係性**	**法規範**	**事実・主張の要約**
Few-shot 結果									
CodeLlama 7B	12.88	9.10	3.70	16.78	85.22	48.94	5.17	10.11	1.24
CodeLlama 7B it.	15.67	11.78	6.07	37.65	89.39	55.10	8.01	11.00	1.29
CodeLlama 13B	15.33	10.90	5.23	17.30	85.04	51.46	7.34	11.89	2.38
CodeLlama 13B it.	16.37	12.35	6.47	33.39	88.70	55.00	8.54	10.76	2.21
Llama3.1 8B	26.10	20.32	11.18	30.00	83.22	64.06	14.21	16.85	2.84
Llama3.1 8B it.	24.47	17.91	10.32	24.00	84.00	62.96	13.95	16.22	2.21
Llama3.2 3B	22.20	17.06	8.65	27.13	80.52	57.35	11.18	12.24	2.28
Llama3.2 3B it.	25.64	19.80	11.38	56.26	92.09	54.11	14.51	10.93	2.78
Gemma2 9B	15.35	11.28	5.28	35.30	93.30	54.88	7.03	9.18	2.56
Gemma2 9B it.	27.22	22.44	12.64	70.70	94.17	73.27	15.16	17.21	1.82
GPT-3.5-Turbo	26.66	22.28	13.51	94.26	**100.0**	73.02	16.18	13.81	**3.88**
GPT-4	**33.46**	**28.70**	**19.96**	**99.13**	**100.0**	**75.31**	**23.24**	**21.52**	3.30
GPT-4o	23.58	20.10	13.42	95.22	**100.0**	75.15	15.82	19.93	2.97
Fine-tuning 結果									
CodeLlama 7B	30.56	23.04	16.34	94.43	99.57	76.73	21.54	39.81	8.59
CodeLlama 7B it.	33.47	25.85	18.68	96.61	99.65	76.90	24.00	34.61	9.03
CodeLlama 13B	34.44	25.94	17.70	97.13	99.83	76.73	23.23	42.23	7.43
CodeLlama 13B it.	35.61	27.75	19.65	96.17	99.65	77.68	24.87	46.32	9.85
Llama3.1 8B	30.09	19.86	13.25	94.70	**100.0**	68.22	19.75	29.01	9.39
Llama3.1 8B it.	29.59	20.32	13.42	87.91	99.83	70.57	18.98	31.51	9.28
Llama3.2 3B	33.38	24.29	17.56	92.78	99.83	73.29	23.83	47.47	9.89
Llama3.2 3B it.	30.37	21.51	14.70	87.22	99.83	71.93	20.38	43.24	10.08
Gemma2 9B	**43.38**	**36.47**	**27.52**	**98.00**	**100.0**	**81.85**	**32.53**	**50.97**	**12.75**
Gemma2 9B it.	42.30	34.26	25.95	96.17	**100.0**	81.02	31.80	42.05	11.92

考えられているにもかかわらず，Few-shot の GPT モデルの結果を上回ることから，LegalViz データセットの有効性が確認できる．また，Gemma2-9B が Fine-tuning 後の実験で `Graph`，`Graph&Node`，`Graph&Node&Edge` の全ての評価で最も良い性能を示した．Few-shot 実験では Gemma2-9B の性能は Gemma-2-9B-it. より低かったことからも，LegalViz データセットの有効性が確認できる．

有効なグラフ生成率　有効なグラフ生成率の評価では，コード生成を 1 回行った場合と 10 回行った場合で文法エラーなく有効な DOT 言語を生成する確率を評価する．前者では，GPT-4 が最も有効なグラフを生成する確率が高く，次に性能が高いのは Gemma2-9B である．後者では，3 つの Fine-tuning 済みモデル (Llama-3.1-8B，Gemma2-9B，Gemma2-9B-it.) が 100 %の確率で有効なグラフを生成した．GPT 以外のモデルで Few-shot と Fine-tuning の結果を比較すると LegalViz による学習で有効なグラフ生成確率が大幅に改善していると言える．

テキスト内容の法律的観点評価　法律的な観点の評価では，3 節で説明した 4 つの評価観点で比較を行った．法主体はほとんどの場合入力した事実概要から取得できるため，4 つの評価観点のうち最もスコアが高い．事実・主張の要約は判例で考慮された重要な事実や主張の要約のテキスト生成タスクであり，他の観点よりスコアが低い傾向にある．これは，法規範の抽出では「法」「規則」等の単語，法律上の関係性の抽出では「契約」「通知」等の単語を伴う確率が高いが，判例で考慮された重要な事実や主張には決まって伴われる用語がほぼ無いためだと考えられる．Fine-tuning した `Gemma2-9B` モデルが法律の 4 つの観点全てで最も良いスコアを示し，特に事実・主張の要約のスコアは全てのモデルで LegalViz の学習により約 3 倍の上昇が見られた．

5　結論

23 言語の EU 法判例を対象に，判例の事実概要を入力し法律上の関係性をグラフ可視化するタスクである LegalViz を提案し，グラフ構造と法律内容の 2 つの側面から新しい評価手法を提案した．その結果，LegalViz による学習がグラフ構造及び法律内容の両面で精度向上に寄与することが示された．

謝辞

本研究は JST CRONOS JPMJCS24K6 の助成を受けたものです.

参考文献

[1] Tom Brown *et al*. Language models are few-shot learners. In **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[2] OpenAI. GPT-4 technical report. Technical report, 2023.

[3] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In **The 36th Conference on Neural Information Processing Systems (NeurIPS)**, 2022.

[4] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. **PLoS digital health**, Vol. 2, No. 2, p. e0000198, 2023.

[5] 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, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In **Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)**, 2023.

[6] Jonathan H Choi, Kristin E Hickman, Amy B Monahan, and Daniel Schwarcz. Chatgpt goes to law school. **J. Legal Educ.**, Vol. 71, p. 387, 2021.

[7] Jens Frankenreiter and Julian Nyarko. Natural language processing in legal tech. **Legal Tech and the Future of Civil Justice**, 2022.

[8] Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In **Proceedings of the Natural Legal Language Processing Workshop 2021**, pp. 19–35, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[9] Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1854–1864, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[10] Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. Charge-based prison term prediction with deep gating network. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 6362–6367, Hong Kong, China, November 2019. Association for Computational Linguistics.

[11] Chaojun Xiao, Haoxiang Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. Cail2018: A large-scale legal dataset for judgment prediction. In **ArXiv**, Vol. abs/1807.02478, 2018.

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

[13] Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders Søgaard. LeXFiles and LegalLAMA: Facilitating English multinational legal language model development. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15513–15535, Toronto, Canada, July 2023. Association for Computational Linguistics.

[14] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 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)**, pp. 4310–4330, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[15] Hiroaki Yamada, Takenobu Tokunaga, Ryutaro Ohara, Akira Tokutsu, Keisuke Takeshita, and Mihoko Sumida. Japanese tort-case dataset for rationale-supported legal judgment prediction. **ArXiv**, Vol. abs/2312.00480, , 2023.

[16] Freda Shi, Daniel Fried, Marjan Ghazvininejad, Luke Zettlemoyer, and Sida I. Wang. Natural language to code translation with execution. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 3533–3546, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[17] Fenia Christopoulou, Guchun Zhang, and Gerasimos Lampouras. Text-to-code generation with modality-relative pre-training. In **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1194–1208, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[18] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4. **ArXiv**, Vol. abs/2303.12712, , 2023.

[19] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In **8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020**. OpenReview.net, 2020.

[20] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, I. Evtimov, Joanna Bitton, Manish P Bhatt, Cristian Cantón Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre D'efossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code. **ArXiv**, Vol. abs/2308.12950, , 2023.

[21] Abhimanyu Dubey, et al. The llama 3 herd of models, 2024.

[22] Gemma Team. Gemma 2: Improving open language models at a practical size, 2024.

[23] Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6974–6996, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

付録

A　統計

下表に計 7,010 件の LegalViz データセットを学習，検定，テストセットに分割した際の統計量を示す．

Split	# Instances	# Nodes	# Relations
Train	4,710	12,624	16,367
Validation	1,150	3,404	2,717
Test	1,150	3,128	3,589
Total	7,010	19,156	22,673

また EUR-LEX にて使用される 23 言語別の統計量を表 2 に示す．

B　各言語毎の評価

表 3 は EU 公式 23 言語の言語別に Few-shot と Fine-tuning を比較し，評価手法を用いて最も性能が良かった `Gemma2-9B` モデルのスコアを示す．また，GPT 系モデルを除いた 10 モデルの平均スコアを算出することにより，各モデルの影響を最小化し，LegalViz による学習の前後のスコア差を確認する．言語毎の比較では，マルタ語，ラトビア語，リトアニア語，ハンガリー語等の利用可能な言語資源が比較的少ない言語 [23] でスコアが低い傾向にあり，比較的言語資源が多い英語とフランス語でスコアが高い傾向にある．言語グループの観点では，ウラル語族のハンガリー語とフィンランド語で共にスコアが低い傾向にあり，ロマンス語族のルーマニア語，フランス語，スペイン語，イタリア語，ポルトガル語はウラル語族のグループやその他の英語・フランス語より言語資源が少ない言語よりもスコアが高い傾向にあるため，言語的な差異がスコアに影響している可能性がある．法律の 4 つの評価軸のうち，法規範と事実・主張の要約では，グラフ生成のために入力テキストの要約が必要なので複数の言語で Few-shot のスコアが 0 となっている．Few-shot の結果と比較して，LegalViz による学習により全ての法律の評価軸・全ての言語でスコアが上昇している．

表 2: 言語別の統計量．L_{word} は事実概要の平均単語数，L_{char} は事実概要の平均文字数を表す．L_{code} は Graphviz コードの文字数を表す．

Lang.	ISO	# Ins.	L_{word}	L_{char}	L_{code}
All	-	7,010	109.0	644.2	759.8
Bulgarian	BG	290	113.4	625.5	759.8
Spanish	ES	307	133.7	693.4	708.4
Czech	CS	307	102.8	582.9	832.5
Danish	DA	307	110.9	640.7	766.8
German	DE	312	108.9	683.0	759.1
Estonian	ET	307	83.9	588.8	809.4
Greek	EL	307	121.4	698.6	779.2
English	EN	312	122.6	629.2	623.0
French	FR	312	128.6	674.8	766.9
Croatian	HR	263	103.2	577.7	718.7
Italian	IT	312	123.3	705.1	788.7
Latvian	LV	307	94.4	598.8	725.7
Lithuanian	LT	307	94.6	609.4	749.8
Hungarian	HU	307	97.4	670.2	809.7
Maltese	MT	305	100.4	706.3	777.7
Dutch	NL	312	122.0	687.0	784.7
Polish	PL	307	106.7	655.0	759.2
Portuguese	PT	307	125.2	653.1	778.0
Romanian	RO	290	118.3	672.0	791.8
Slovak	SK	308	101.0	585.7	727.9
Slovenian	SL	308	106.6	580.0	730.5
Finnish	FI	308	78.5	649.6	808.7
Swedish	SV	308	108.9	639.1	748.2

表 3: **法主体**，**法律上の関係性**，**法規範**，**事実・主張の要約**の 4 つの評価軸のスコア．“fs.” は Few-shot，“ft.” は Fine-tuning を意味する． Avg. fs. は GPT 系モデルを除いた Few-shot モデルの平均スコアを表す．

Model	BG	ES	CS	DA	DE	ET	EL	EN	FR	HR	IT	LV	LT	HU	MT	NL	PL	PT	RO	SK	SL	FI	SV
法主体																							
Gemma 2 9B fs.	59.22	59.20	52.53	54.47	56.24	53.51	53.22	59.95	53.77	55.32	55.42	51.33	55.04	42.26	47.55	60.06	51.17	61.68	59.21	52.85	50.72	55.69	57.21
Gemma 2 9B ft.	80.62	84.33	80.06	82.53	83.31	78.57	80.44	86.98	82.90	81.14	80.94	81.11	77.16	81.66	82.59	82.70	82.58	85.64	83.97	81.27	79.16	79.23	83.22
Avg. fs. models	60.78	63.32	57.12	59.67	61.30	57.67	53.59	65.77	60.61	60.48	59.06	56.85	55.53	55.73	56.64	61.37	60.18	61.36	60.89	59.50	57.20	59.68	61.04
Avg. ft. models	73.19	77.05	73.59	74.14	74.28	72.62	72.31	78.78	75.5	74.87	75.82	71.87	72.08	72.16	73.36	75.2	74.25	76.72	76.22	74.08	72.77	72.47	75.17
法律上の関係性																							
Gemma 2 9B fs.	10.64	7.24	8.97	10.44	2.28	7.31	7.31	13.39	5.23	7.11	4.42	3.54	6.79	2.81	3.71	9.27	8.01	11.18	6.90	6.81	5.94	2.40	6.73
Gemma 2 9B ft.	32.00	33.88	31.19	37.13	32.34	26.13	30.42	36.53	23.06	36.43	27.14	30.74	28.06	38.13	31.88	29.68	40.27	34.39	37.17	35.15	29.50	34.40	33.02
Avg. fs. models	11.22	11.80	10.28	13.90	11.26	8.74	12.56	10.76	11.67	10.47	10.12	12.78	12.19	11.24	9.22	13.66	12.18	11.88	10.62	10.42	11.27	10.02	11.68
Avg. ft. models	23.66	23.18	22.96	25.47	22.98	23.0	23.01	27.72	22.09	25.56	20.87	24.4	25.79	21.45	23.09	20.95	24.13	24.08	23.2	23.36	23.3	21.94	24.86
法規範																							
Gemma 2 9B fs.	0.00	0.00	15.15	10.37	13.85	17.95	15.91	17.96	0.00	9.92	0.00	4.97	0.00	6.89	0.00	6.20	19.31	12.47	6.05	14.77	0.00	7.51	23.25
Gemma 2 9B ft.	58.98	49.07	41.06	52.69	56.19	54.21	47.31	61.32	61.87	51.14	49.90	64.56	38.86	42.34	48.10	44.85	53.68	42.08	54.92	48.00	47.03	54.46	54.68
Avg. fs. models	12.19	11.78	14.39	13.89	13.58	12.11	12.38	14.07	15.02	15.15	13.04	14.51	14.84	13.35	11.65	16.77	14.65	13.66	12.89	15.30	9.52	11.08	15.20
Avg. ft. models	36.8	37.3	36.52	37.38	45.55	38.04	33.58	36.04	36.92	40.67	36.5	40.69	45.54	39.53	37.58	35.53	42.92	34.26	35.04	44.14	36.46	34.54	38.53
事実・主張の要約																							
Gemma 2 9B fs.	0.00	3.25	3.27	3.74	4.42	2.30	1.53	1.54	1.59	3.85	0.00	0.00	1.95	2.41	4.40	3.94	3.03	4.86	1.40	1.32	4.43	4.54	1.45
Gemma 2 9B ft.	16.78	17.10	9.17	12.76	6.78	9.86	20.95	5.90	11.09	11.13	16.32	11.21	15.90	15.22	11.10	8.92	12.03	11.90	16.09	17.65	13.17	7.46	14.89
Avg. fs. models	2.10	2.66	1.57	2.55	2.50	1.62	2.15	3.75	2.98	1.99	1.38	2.50	3.04	1.75	1.71	2.31	3.63	2.24	2.49	1.82	2.34	1.25	1.86
Avg. ft. models	10.52	12.05	6.94	9.62	8.64	7.33	13.1	11.23	8.9	9.07	10.92	9.05	10.68	8.47	10.1	8.32	10.77	9.28	8.89	9.22	8.67	6.45	9.04