Unlocking Practical Applications in Legal Domain: Evaluation of GPT for Zero-Shot Semantic Annotation of Legal Texts

Jaromir Savelka
Carnegie Mellon University
Pittsburgh, PA, USA
jsavelka@cs.cmu.edu

ABSTRACT

We evaluated the capability of a state-of-the-art generative pre-trained transformer (GPT) model to perform semantic annotation of short text snippets (one to few sentences) coming from legal documents of various types. Discussions of potential uses (e.g., document drafting, summarization) of this emerging technology in legal domain have intensified, but to date there has not been a rigorous analysis of these large language models' (LLM) capacity in sentence-level semantic annotation of legal texts in zero-shot learning settings. Yet, this particular type of use could unlock many practical applications (e.g., in contract review) and research opportunities (e.g., in empirical legal studies). We fill the gap with this study. We examined if and how successfully the model can semantically annotate small batches of short text snippets (10–50) based exclusively on concise definitions of the semantic types. We found that the GPT model performs surprisingly well in zero-shot settings on diverse types of documents (F_1 = .73 on a task involving court opinions, .86 for contracts, and .54 for statutes and regulations). These findings can be leveraged by legal scholars and practicing lawyers alike to guide their decisions in integrating LLMs in wide range of workflows involving semantic annotation of legal texts.

CCS CONCEPTS

• **Applied computing → Law**; **Annotation**.

KEYWORDS

Semantic legal annotation, generative pre-trained transformers, GPT, transfer learning, zero-shot, adjudicatory decisions, contracts, statutory and regulatory provisions

1 INTRODUCTION

We evaluate the effectiveness of GPT-3.5 (`text-davinci-003`) in tasks focused on (i) contract review, (ii) statutory/regulatory provisions investigation and (iii) case-law analysis. We benchmark the performance of the general (not fine-tuned) GPT-3.5 model in performing annotation of small batches of short text snippets coming from the aforementioned types of legal documents against the performance of a traditional statistical machine learning (ML) model (random forest) and fine-tuned BERT model (RoBERTa). The GPT moodel's annotations are based on compact one sentence long semantic type definitions provided to the model as a prompt. Specifically, we analyzed the following research question in the context of the three legal annotation tasks: Given brief type definitions from a single non-hierarchical type system describing short snippets of text, how successfully can a general GPT-3.5 model automatically classify such text spans in terms of the type system's categories?

2 RELATED WORK

Zero-shot GPT in AI & Law. Yu et al. applied GPT to the COLIEE entailment task, improving on the then existing state-of-the-art [45]. Bommarito and Katz successfully applied GPT-3.5 and GPT-4 to the Bar Examination [4, 18]. Other use cases include assessment of trademark distinctiveness [13], legal reasoning [3, 22], and U.S. Supreme court judgment modeling [14].

Rhetorical/Functional Segments in Adjudicatory Decisions. The task involves labeling of smaller textual snippets such as sentences [31] in terms of, e.g., rhetorical roles, functional or argument units. Examples include court [27] or administrative decisions from the U.S. [36], multi-domain court decisions from India [1] or Canada [43, 44], international court [25] or arbitration decisions [6], or even multi-{domain,country} adjudicatory decisions [32]. Identifying a section that states an outcome of the case has also received considerable attention separately [24, 42]. The task sometimes takes a form of identifying a small number of contiguous parts typically comprising multiple paragraphs. Different variations of this task were applied to several legal domains from countries such as Canada [11], the Czech Republic [15], France [5], the U.S. [28], or even in multi-jurisdictional settings [33].

Classification of Legal Norms. Researchers used traditional statistical supervised ML models to classify portions of Italian statutory texts with types, such as definition, prohibition, or obligation [2, 12]. Other groups classified sentences from Dutch statutory texts in terms of categories such as definition, publication provision, or scope of change [10]. Some work focuses on fine-grained semantic analysis of statutory texts in terms of obligations, permissions, subject agents or themes [26, 30, 41], concepts or definitions [40].
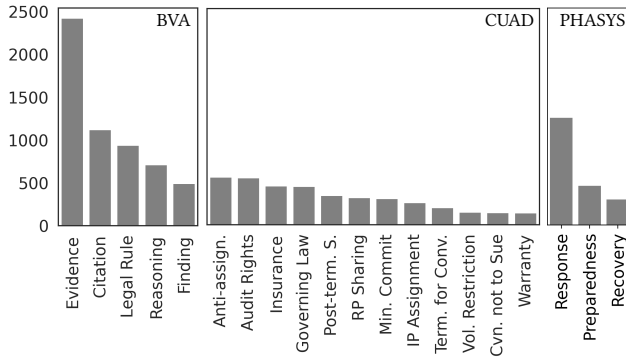
Figure 1: Semantic Types Distribution. The figure shows distribution of the semantic types across the three data sets.

Classification of contractual clauses.. Chalkidis et al. analyzed contractual clauses in terms of types such as termination clause, governing law or jurisdiction [8, 9]. Leivaditi et al. released a benchmark data set of 179 lease agreement documents focusing on recognition of entities and red flags [20]. In this work we focus on twelve selected semantic types from the Contract Understanding Atticus Dataset (CUAD) [16]. Wang et al. assembled and released the Merger Agreement Understanding Dataset (MAUD) [37].

3 DATA

We use three existing manually annotated data sets. Each supports various tasks involving different types of legal documents. All of them are equipped with expert annotations attached to (usually) short pieces of text. We further filtered and processed the data sets to make them suitable for this work's experiments.

The U.S. Board of Veterans' Appeals[1] (BVA) is an administrative body within the U.S. Department of Veterans Affairs (VA) responsible for hearing appeals from veterans who are dissatisfied with decisions made by VA regional offices. The BVA reviews a wide range of issues, including claims for disability compensation, survivor benefits, and other compensation and pension claims. Walker et al. [36] analyzed 50 BVA decisions issued between 2013 and 2017. The decisions were all arbitrarily selected cases dealing with claims by veterans for service-related post-traumatic stress disorder (PTSD). For each decision, the researchers manually extracted sentences addressing the factual issues. The sentences were then manually annotated with rhetorical roles they play in the respective decisions [35]. Figure 1 (left) shows the distribution of the labels.

Contract Understanding Atticus Dataset (CUAD) is a corpus of 510 commercial legal contracts that have been manually labeled under the supervision of professional lawyers. This effort resulted in more than 13,000 annotations.[2] The data set was released by Hendrycks et al. [16] and it identifies 41 types of legal clauses that are typically considered important in contract review in connection with corporate transactions. In this study, we decided to work with
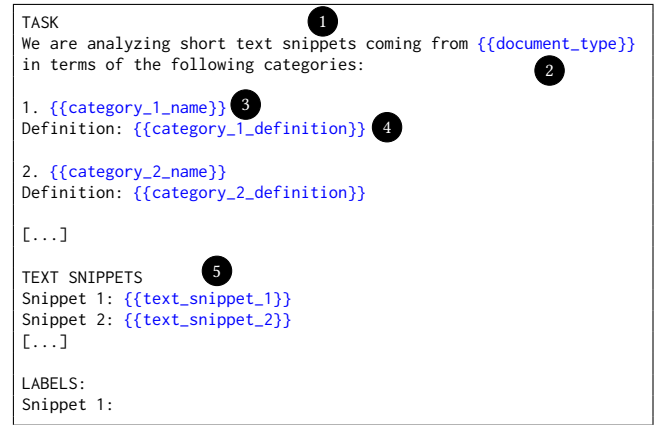

Figure 2: Batch Prediction Prompt Template. The preamble (1) primes the model to generate semantic type predictions. The tokens surrounded by curly braces are replaced with the document type (2) according to the data set, the names of the semantic types (3), the corresponding definitions (4), and the analyzed text snippets (5).

the 12 most common clause-level types present in the corpus the distribution of which is shown in Figure 1 (center).

At the University of Pittsburgh's Graduate School of Public Health, researchers have manually coded federal, state and local laws and regulations related to emergency preparedness and response of the public health system (PHS). They used the codes to analyze network diagrams representing various functional features of states' regulatory frameworks for public health emergency preparedness. They retrieved candidate sets of relevant statutes and regulations from a full-text legal information service and identified relevant spans of text [34]. They then coded the relevant spans as per instructions in the codebook,[3] representing relevant features of those spans. In this work we focus on the purpose of the legal provision in terms of the three categories the distribution of which is shown in Figure 1 (right). The statutory and regulatory texts were automatically divided into text units which are often non-contiguous spans of text referenceable with citations [29].

4 EXPERIMENTS

We use the Jaccard similarity measure as the baseline (tokens as sets). Each text snippet is compared to the type definitions available in the respective task. It is then assigned the label the definition of which has the highest similarity score with the snippet.

We benchmark the performance of GPT-3.5 [7, 23] against a traditional statistical supervised learning algorithm (random forest [17]) which is evaluated using 10-fold cross-validation at the document level. Within each iteration of the cross-validation, we utilize grid search[4] to select the best set of hyperparameters. The space that is being considered is defined over the type of n-grams to be used, number of estimators and maximum tree depth.

---

[1]U.S. Department of Veterans' Appeals: Board of Veteran's Appeals. Available at: https://www.bva.va.gov/ [Accessed 2023-02-09]
[2]The Atticus Project: Contract Understanding Atticus Dataset (CUAD). Available at: https://www.atticusprojectai.org/cuad [Accessed 2023-02-09]

[3]PHASYS ARM 2 – LEIP Code Book. Available at: https://www.phasys.pitt.edu/pdf/Code_Book_Numerical_Defintions.pdf [Accessed 2023-02-09]
[4]Scikit learn: GridSearchCV. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html [Accessed 2023-02-05]

Table 1: Experimental results in terms of micro-F_1 scores. The Jaccard column shows the performance of the Jaccard similarity baseline. The @N labels denote how many data points were used in the training of the two supervised ML systems (RandF – random forest, BERT – base RoBERTa), where @Max means all the available data points were used. The GPT section reports the performance of the text-davinci-003 model. The blue cells signify the point at which the BERT system matched the performance of either one of the GPT-3.5 models. The green shaded cells do the same for the random forest.

	Jaccard	@20 RandF	BERT	@50 RandF	BERT	@100 RandF	BERT	@250 RandF	BERT	@500 RandF	BERT	@1000 RandF	BERT	@Max RandF	BERT	GPT
BVA	**.35**	**.30**	**.27**	**.31**	**.26**	**.51**	**.52**	**.65**	**.82**	**.65**	**.85**	**.73**	**.89**	**.83**	**.92**	**.73**
Citation	.71	.11	.00	.12	.00	.77	.49	.94	.97	.97	.99	.97	1.0	.99	1.0	.96
Evidence	.28	.58	.59	.61	.60	.69	.66	.77	.91	.76	.92	.81	.93	.87	.94	.77
Finding	.08	.22	.10	.14	.00	.35	.21	.50	.55	.48	.65	.63	.77	.74	.86	.53
Legal Rule	.15	.00	.00	.06	.00	.15	.15	.38	.77	.41	.84	.68	.91	.89	.95	.72
Reasoning	.24	.11	.10	.04	.00	.07	.00	.23	.52	.15	.56	.21	.64	.40	.71	.43
CUAD	**.38**	**.25**	**.12**	**.31**	**.14**	**.52**	**.43**	**.73**	**.68**	**.80**	**.87**	**.86**	**.93**	**.89**	**.95**	**.86**
Anti-assignment	.56	.12	.00	.16	.10	.78	.71	.87	.81	.90	.94	.93	.97	.94	.99	.96
Audit Rights	.50	.39	.12	.35	.06	.49	.71	.71	.82	.81	.94	.86	.97	.90	.96	.93
Cvnt. not to Sue	.31	.04	.00	.00	.00	.04	.00	.73	.26	.76	.87	.76	.93	.90	.96	.78
Governing Law	.52	.80	.17	.80	.22	.96	.92	.94	1.0	.98	1.0	.99	1.0	1.0	1.0	.99
IP Assignment	.05	.40	.00	.51	.00	.53	.08	.73	.13	.76	.84	.83	.88	.87	.94	.84
Insurance	.38	.11	.13	.23	.13	.64	.73	.93	.92	.90	.97	.93	.98	.95	.97	.96
Min. Commitment	.20	.00	.00	.23	.00	.52	.00	.64	.38	.72	.69	.77	.87	.81	.91	.67
Post-term. Services	.43	.21	.15	.29	.10	.44	.29	.60	.57	.72	.77	.77	.84	.78	.85	.69
Profit Sharing	.25	.08	.17	.04	.16	.03	.01	.56	.75	.76	.88	.86	.92	.87	.94	.78
Termination Cnv.	.43	.27	.00	.60	.00	.42	.00	.81	.81	.88	.91	.88	.95	.91	.97	.89
Volume Restriction	.07	.00	.00	.01	.00	.03	.00	.08	.30	.15	.41	.45	.78	.61	.90	.46
Warranty Duration	.21	.02	.00	.12	.00	.15	.00	.48	.21	.56	.76	.79	.91	.84	.93	.81
PHASYS	**.24**	**.52**	**.48**	**.53**	**.48**	**.53**	**.49**	**.56**	**.65**	**.59**	**.68**	**.61**	**.72**	**.63**	**.72**	**.54**
Preparedness	.24	.01	.00	.00	.00	.00	.00	.08	.31	.22	.46	.20	.56	.31	.61	.45
Response	.24	.77	.77	.77	.77	.77	.77	.78	.79	.78	.78	.79	.80	.80	.81	.63
Recovery	.27	.27	.00	.33	.00	.33	.03	.38	.55	.35	.59	.48	.60	.53	.67	.28

To compare the performance of the zero-shot GPT-3.5 model against a fine-tuned LLM, we fine-tune the base RoBERTa model [21] for 10 epochs on the training set within each of the cross-validation folds. The same splits as for evaluating the performance of the random forest are used. We set the batch size to 16 and the length of the sequence to 512. As optimizer we use the Adam algorithm [19] with initial learning rate set to $4e^{-5}$.

To gauge how many labeled documents are needed by the supervised ML system to match and exceed the performance of GPT-3.5, we train the random forest and RoBERTa on training sets of varying sizes. We train the systems on the training sets with 20, 50, 100, 250, 500 and 1,000 data points. These documents are randomly sampled from the training set in each iteration of the cross-validation.

To test the performance of text-davinci-003, we submit a batch of text snippets using the openai Python library[5] which is a wrapper for the OpenAI's REST API. We make the batches as large as possible to achieve maximum cost effectiveness. Their size is determined by the size of the evaluated text snippets that can fit into the prompt (4,097 tokens), leaving enough space for the completion (i.e., the predictions). For the BVA decisions' sentences

[5]GitHub: OpenAI Python Library. Available at: https://github.com/openai/openai-python [Accessed 2023-02-09]

the batch size was set to 50, for the CUAD's contractual clauses to 20, and for PHASYS' statutory and regulatory provisions to 10.

We embed each batch in the prompt template shown in Figure 2. The model returns the list of predicted labels as the prompt completion. The construction of the prompt is focused on maximizing the cost effectiveness of the proposed approach which may somewhat limit the performance of the evaluated GPT-3.5 model.

We set the temperature to 0.0, which corresponds to no randomness. The higher the temperature the more creative the output but it can also be less factual. We set max_tokens to 500 (a token roughly corresponds to a word). This parameter controls the maximum length of the output. We set top_p to 1, as is recommended when temperature is set to 0.0. This parameter is related to temperature and also influences creativeness of the output. We set frequency_penalty to 0, which allows repetition by ensuring no penalty is applied to repetitions. Finally, we set presence_penalty to 0, ensuring no penalty is applied to tokens appearing multiple times in the output, which is especially important for our use case.

5 RESULTS AND DISCUSSION

Table 1 shows the results of our experiments on applying the text-davinci-003 model to the three tasks involving adjudicatory

Table 2: Confusion Matrices of GPT-3.5 Predictions. The columns show the true labels as assigned by human experts, while the rows report the predictions of the system.

CUAD

	AA	AR	CS	GL	IP	IN	MC	PT	PS	TC	VR	WD
Anti-assignment	320	0	0	0	0	0	0	2	1	0	0	2
Audit Rights	0	489	0	0	0	2	0	18	0	0	0	4
Cvnt. not to Sue	5	6	97	0	3	4	1	13	0	0	1	2
Governing Law	1	0	0	439	0	0	0	1	1	0	1	0
IP Assignment	16	1	22	0	242	0	0	28	11	0	2	0
Insurance	0	5	0	0	0	423	0	10	0	0	0	0
Min. Commitment	0	4	0	0	0	1	200	8	32	1	44	6
Post-term. Services	0	10	0	0	1	1	3	142	1	3	14	3
Profit Sharing	0	8	0	0	1	2	30	28	252	1	9	1
Termination Cnv.	7	0	1	1	0	2	16	8	0	188	0	3
Volume Restriction	5	7	0	0	1	0	41	7	9	0	64	3
Warranty Duration	0	3	0	0	0	6	6	5	1	0	3	109

BVA

	CIT	EVD	FND	LR	RSN
Citation	1050	9	1	10	2
Evidence	7	1582	33	3	84
Finding	0	168	274	18	87
Legal Rule	35	3	28	572	25
Reasoning	20	654	149	332	504

PHASYS

	PRP	RES	REC
Preparedness	308	518	106
Response	139	699	134
Recovery	2	13	50

opinions (BVA), contract clauses (CUAD) and statutory and regulatory provisions (PHASYS). Firstly, the GPT-3.5 model outperforms the baseline based on Jaccard similarity by a large margin on all the three tasks (.35 vs .73 on BVA, .38 vs .86 on CUAD, and .24 vs .54 on PHASYS). While the magnitude of the difference might be somewhat surprising the better performance of the GPT-3.5 model as compared to the baseline is to be expected. The baseline only matches the exact words from the type definitions to the exact words in the evaluated text snippets, whereas the sophisticated GPT-3.5 model has access to their semantics.

When compared to the supervised algorithms trained on the in-domain data, the performance of the GPT-3.5 model is surprisingly high. While the GPT-3.5 model clearly falls short when compared to the supervised models trained on large portions of the available data, it is quite competitive, perhaps beyond what could be reasonably expected, when the size of the training data is limited. Overall, it appears that the RoBERTa model needs at least several hundred training data points to match the performance of the GPT-3.5 model (light-blue cells in Table 1). The random forest requires even more data, often close to a thousand (light-green cells in Table 1).

The three confusion matrices in Table 2 offer detailed view into the performance of the GPT-3.5 model. The performance on CUAD's contractual clauses appears very promising overall. There is a small number of classes that the system struggles to distinguish from each other, such as *Minimum Commitment*, *Profit Sharing*, or *Volume Restrictions*. As for the BVA's adjudicatory documents, the *Reasoning* class appears to be the most problematic. There is especially large number of *Evidence* sentences (654) that have been misclassified as *Reasoning*. Finally, the PHASYS' statutory and regulatory provisions seem to be the most challenging. A large number of *Emergency Response* provisions are labeled as *Emergency Preparedness*.

While the results of our experiments are promising, limitations clearly exist. First, the performance of the models is far from perfect and there is a considerable gap between the performance of the zero-shot LLM compared to the performance of the supervised ML systems trained on hundreds or thousands of example data points. Hence, in workflows with low tolerance towards inaccuracies in semantic annotation the zero-shot LLM predictions may need to

be subjected to a human-expert QA. The outcome of such human-computer interaction may be a high-quality data set of the size that enables fine-tuning of a powerful domain-adapted LLM.

There are considerable differences in performance of GPT-3.5 across the three data sets. While the performance on the CUAD data set seems to be very reasonable there are some limitations when it comes to the performance on the BVA data set. The model struggles with the *Reasoning* type in terms of mislabeling many sentences of other types as *Reasoning* as well as not recognizing many *Reasoning* sentences as such (Table 2). This is consistent with the performance of the supervised ML models. While the fine-tuned base RoBERTa is clearly more successful in handling this semantic type compared to the GPT-3.5, it still struggles (F_1 = .71). The random forest model under-performs GPT-3.5. Hence, the correct recognition of this type may require extremely nuanced notions that may be difficult to acquire through a compact one-sentence definition (GPT-3.5) or word occurrence features (random forest). For such situations, the proposed approach might not (yet) be powerful enough and the only viable solution could be fine-tuning an LLM.

The performance of GPT-3.5 on the PHASYS data set is not satisfactory. We identified several challenges this data set poses that makes it difficult even to the supervised ML models (Table 1). The data set is imbalanced with the *Response* type constituting 62.4% of the available data points. Second, the definitions of the semantic types appear to be somewhat less clear and lower quality than for the other two data sets. Hence, we hypothesize that the manual annotation of this data set heavily relied on the informal expertise of the human annotators which was not fully captured in the annotation guidelines. Finally, the fine-grained distinctions between what counts as emergency *Response* as opposed to *Preparedness* may simply be too nuanced to be captured in a compact definition.

6 CONCLUSIONS

We evaluated `text-davinci-003` on three legal annotation tasks, involving adjudicatory opinions, contractual clauses, and statutory and regulatory provisions. The model was provided with a list of compact definitions of the semantic types. The tasks were to assign a batch of short text snippets with the defined categories. The

results of the experiment are very promising, where the model achieved (micro) $F_1 = .73$ for the rhetorical roles of sentences from adjudicatory decisions, .86 for the types of contractual clauses, and .54 for the purpose of public-health system's emergency response and preparedness statutory and regulatory provisions. Our findings are important for legal professionals, educators and scholars who intend to leverage the capabilities of state-of-the-art LLMs to lower the cost of existing high-volume workloads, involving semantic annotation of legal documents, or to unlock novel workflows that would have not been economically feasible to be performed manually or using supervised ML. We also envision that the approach could be successfully combined with high-speed similarity annotation frameworks [38, 39] to enable highly cost efficient annotation in situations where resources are scarce.

REFERENCES

[1] Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2019. Identification of rhetorical roles of sentences in Indian legal judgments. In *JURIX 2019*, Vol. 322. IOS Press, 3.

[2] Carlo Biagioli, Enrico Francesconi, Andrea Passerini, Simonetta Montemagni, and Claudia Soria. 2005. Automatic semantics extraction in law documents. In *10th International Conference on Artificial Intelligence and Law*. 133–140.

[3] Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can GPT-3 perform statutory reasoning? *arXiv preprint arXiv:2302.06100* (2023).

[4] Michael Bommarito and Daniel Martin Katz. 2022. GPT Takes the Bar Exam. *arXiv preprint arXiv:2212.14402* (2022).

[5] Paul Boniol, George Panagopoulos, Christos Xypolopoulos, Rajaa El Hamdani, David Restrepo Amariles, and Michalis Vazirgiannis. 2020. Performance in the Courtroom: Automated Processing and Visualization of Appeal Court Decisions in France. In *Proceedings of the Natural Legal Language Processing Workshop*.

[6] Karl Branting, Brandy Weiss, Bradford Brown, Craig Pfeifer, A Chakraborty, Lisa Ferro, M Pfaff, and A Yeh. 2019. Semi-supervised methods for explainable legal prediction. In *ICAIL 2019*. 22–31.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[8] Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2017. Extracting contract elements. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*. 19–28.

[9] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2021. Neural contract element extraction revisited: Letters from sesame street. *arXiv preprint arXiv:2101.04355* (2021).

[10] Emile de Maat, Kai Krabben, and Radboud Winkels. 2010. Machine learning versus knowledge based classification of legal texts. In *Legal Knowledge and Information Systems*. IOS Press, 87–96.

[11] Atefeh Farzindar and Guy Lapalme. 2004. LetSum, an Automatic Text Summarization system in Law field. JURIX 2004.

[12] Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia. 2010. *Integrating a bottom–up and top–down methodology for building semantic resources for the multilingual legal domain*. Springer.

[13] John Goodhue and Yolanda Wei. 2023. Classification of Trademark Distinctiveness using OpenAI GPT 3.5 model. *Available at SSRN 4351998* (2023).

[14] Sil Hamilton. 2023. Blind Judgement: Agent-Based Supreme Court Modelling With GPT. *arXiv preprint arXiv:2301.05327* (2023).

[15] Jakub Harašta, Jaromír Šavelka, František Kasl, and Jakub Míšek. 2019. Automatic Segmentation of Czech Court Decisions into Multi-Paragraph Parts. *Jusletter IT* 4, M (2019).

[16] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. In *35th Conference on Neural Information Processing Systems: Track on Datasets and Benchmarks*.

[17] Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1. IEEE, 278–282.

[18] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. Gpt-4 passes the bar exam. *Available at SSRN 4389233* (2023).

[19] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[20] Spyretta Leivaditi, Julien Rossi, and Evangelos Kanoulas. 2020. A benchmark for lease contract review. *arXiv preprint arXiv:2010.10386* (2020).

[21] 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. *arXiv preprint arXiv:1907.11692* (2019).

[22] Ha-Thanh Nguyen, Randy Goebel, Francesca Toni, Kostas Stathis, and Ken Satoh. 2023. How well do SOTA legal reasoning models support abductive reasoning? *arXiv preprint arXiv:2304.06912* (2023).

[23] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

[24] Alina Petrova, John Armour, and Thomas Lukasiewicz. 2020. Extracting Outcomes from Appellate Decisions in US State Courts. In *JURIX 2020*. 133.

[25] Prakash Poudyal, Jaromír Šavelka, 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*. 67–75.

[26] Jaromír Šavelka and Kevin D Ashley. 2015. Transfer of predictive models for classification of statutory texts in multi-jurisdictional settings. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*. 216–220.

[27] Jaromir Savelka and Kevin D Ashley. 2017. Using conditional random fields to detect different functional types of content in decisions of united states courts with example application to sentence boundary detection. In *Workshop on Automated Semantic Analysis of Information in Legal Texts*. 10.

[28] Jaromir Savelka and Kevin D Ashley. 2018. Segmenting US Court Decisions into Functional and Issue Specific Parts.. In *JURIX 2018*. 111–120.

[29] Jaromír Šavelka, Matthias Grabmair, and Kevin D Ashley. 2014. Mining information from statutory texts in multi-jurisdictional settings. In *Legal Knowledge and Information Systems*. IOS Press, 133–142.

[30] Jaromír Šavelka, Gaurav Trivedi, and Kevin D Ashley. 2015. Applying an interactive machine learning approach to statutory analysis. In *Legal Knowledge and Information Systems*. IOS Press, 101–110.

[31] Jaromir Savelka, Vern R Walker, Matthias Grabmair, and Kevin D Ashley. 2017. Sentence boundary detection in adjudicatory decisions in the united states. *Traitement automatique des langues* 58 (2017), 21.

[32] Jaromír Šavelka, Hannes Westermann, and Karim Benyekhlef. 2020. Cross-Domain Generalization and Knowledge Transfer in Transformers Trained on Legal Data. (2020).

[33] Jaromir Savelka, Hannes Westermann, Karim Benyekhlef, Charlotte S Alexander, Jayla C Grant, David Restrepo Amariles, Rajaa El Hamdani, Sébastien Meeùs, Aurore Troussel, Michał Araszkiewicz, et al. 2021. Lex Rosetta: transfer of predictive models across languages, jurisdictions, and legal domains. In *Proceedings of the eighteenth international conference on artificial intelligence and law*. 129–138.

[34] Patricia M Sweeney, E Bjerke, MA Potter, H Guclu, CR Keane, KD Ashley, M Grabmair, and R Hwa. 2013. Network analysis of manually-encoded state laws and prospects for automation. *Winkels, Lettieri & Faro (eds.) Network Analysis in Law* (2013), 53–77.

[35] Vern R Walker, Ji Hae Han, Xiang Ni, and Kaneyasu Yoseda. 2017. Semantic types for computational legal reasoning: propositional connectives and sentence roles in the veterans' claims dataset. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*. 217–226.

[36] Vern R Walker, Krishnan Pillaipakkamnatt, Alexandra M Davidson, Marysa Linares, and Domenick J Pesce. 2019. Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning. *ASAIL@ ICAIL* 2385 (2019).

[37] Steven H Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dimitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2023. MAUD: An Expert-Annotated Legal NLP Dataset for Merger Agreement Understanding. *arXiv preprint arXiv:2301.00876* (2023).

[38] Hannes Westermann, Jaromír Šavelka, Vern R Walker, Kevin D Ashley, and Karim Benyekhlef. 2019. Computer-Assisted Creation of Boolean Search Rules for Text Classification in the Legal Domain. In *JURIX 2019*, Vol. 322. IOS Press, 123.

[39] Hannes Westermann, Jaromír Šavelka, Vern R Walker, Kevin D Ashley, and Karim Benyekhlef. 2020. Sentence Embeddings and High-Speed Similarity Search for Fast Computer Assisted Annotation of Legal Documents. In *JURIX 2020*, Vol. 334. IOS Press, 164.

[40] Radboud Winkels and Rinke Hoekstra. 2012. Automatic Extraction of Legal Concepts and Definitions. In *Legal Knowledge and Information Systems: JURIX 2012: the Twenty-Fifth Annual Conference*, Vol. 250. IOS Press, 157.

[41] Adam Wyner and Wim Peters. 2011. On rule extraction from regulations. In *Legal knowledge and information systems*. IOS Press, 113–122.

[42] Huihui Xu, Jaromír Šavelka, and Kevin D Ashley. 2020. Using Argument Mining for Legal Text Summarization. In *JURIX 2020*, Vol. 334. IOS Press.

[43] Huihui Xu, Jaromir Savelka, and Kevin D Ashley. 2021. Accounting for sentence position and legal domain sentence embedding in learning to classify case sentences. In *Legal Knowledge and Information Systems*. IOS Press, 33–42.

[44] Huihui Xu, Jaromir Savelka, and Kevin D Ashley. 2021. Toward summarizing case decisions via extracting argument issues, reasons, and conclusions. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*. 250–254.

[45] Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. Legal Prompting: Teaching a Language Model to Think Like a Lawyer. https://doi.org/10.48550/ARXIV.2212.01326