# MAUD: An Expert-Annotated Legal NLP Dataset for Merger Agreement Understanding

**Steven H. Wang**[1*] **Antoine Scardigli**[1] **Leonard Tang**[2] **Wei Chen**[3] **Dimitry Levkin**[3]
**Anya Chen**[4] **Spencer Ball**[5] **Thomas Woodside**[6] **Oliver Zhang**[7] **Dan Hendrycks**[8]
[1]ETH Zürich [2]Harvard University [3]The Atticus Project [4]The Nueva School
[5]University of Wisconsin, Madison [6]Yale University [7]Stanford University [8]UC Berkeley

## Abstract

Reading comprehension of legal text can be a particularly challenging task due to the length and complexity of legal clauses and a shortage of expert-annotated datasets. To address this challenge, we introduce the Merger Agreement Understanding Dataset (MAUD), an expert-annotated reading comprehension dataset based on the American Bar Association's 2021 Public Target Deal Points Study, with over 39,000 examples and over 47,000 total annotations. Our fine-tuned Transformer baselines show promising results, with models performing well above random on most questions. However, on a large subset of questions, there is still room for significant improvement. As the only expert-annotated merger agreement dataset, MAUD is valuable as a benchmark for both the legal profession and the NLP community.

## 1 Introduction

While pretrained Transformers (Devlin et al., 2019; Brown et al., 2020) have surpassed humans on reading comprehension tasks such as SQuAD 2.0 (Rajpurkar et al., 2018a) and SuperGLUE (Wang et al., 2019), their accuracy in understanding real-world specialized legal texts remains underexplored.

Reading comprehension of legal text can be a particularly challenging natural language processing (NLP) task due to the length and complexity of legal clauses and the difficulty of collecting expert-annotated datasets. To help address this challenge, we introduce the Merger Agreement Understanding Dataset (MAUD), a legal reading comprehension dataset curated under the supervision of highly specialized mergers-and-acquisitions (M&A) lawyers and used in the American Bar Association's 2021 Public Target Deal Points Study ("ABA Study"). The dataset and code for MAUD can be found at github.com/TheAtticusProject/maud.

*Correspondence to stewang@student.ethz.edu

Public target company acquisitions are the most prominent business transactions, valued at hundreds of billions of dollars each year. Merger agreements are the legal documents that enable these acquisitions, and key clauses in these merger agreements are called "deal points."

Lawyers working on the ABA Study perform contract review on merger agreements. In general, contract review is a two-step process. First, lawyers extract key legal clauses from the contract (a span extraction task). Second, they interpret the meaning of these legal clauses (a reading comprehension task). In the ABA Study, the lawyers extract deal points from merger agreements, and for each deal point they answer a set of standardized multiple-choice questions.

Models trained on MAUD's expert-annotated data can learn to answer 92 reading comprehension questions from the 2021 ABA Study, given extracted deal point text from merger agreements. By answering these questions, models interpret the meaning of specialized legal language and categorize the different agreements being made by companies in the contract.

Span extraction and reading comprehension are both important and challenging tasks in legal contract review. A large-scale expert-annotated span extraction benchmark for contract review is already available in Hendrycks et al. (2021b). However, to the best of our knowledge, there is no large-scale expert-annotated reading comprehension dataset for contract review or any other legal task in the English language. Therefore in this short paper, we focus on the legal reading comprehension task. (Appendix A.14 presents a preliminary benchmark for the extraction task for interested researchers.)

Annotating MAUD was a collective effort of over 10,000 hours by law students and experienced lawyers. Prior to labeling, each law student attended 70-100 hours of training, including lectures and workshops from experienced M&A lawyers.



*Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16369–16382
December 6-10, 2023 

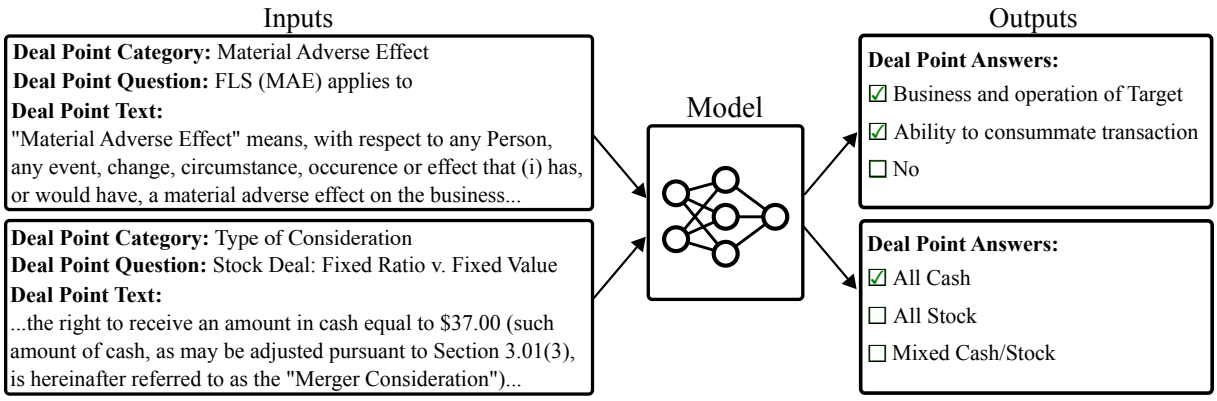


Figure 1: MAUD contains 39,000+ examples for 92 different reading comprehension questions about merger agreements. Given a *deal point question* and *deal point text*, a model learns to predict the correct answer(s) from a list of possible answers standardized by the 2021 ABA Study. The deal point texts above are truncated for display.

Each annotation was labelled by three law student annotators, and these labels were verified by an experienced lawyer. See Appendix A.12 for more information on the annotation process. We estimate the pecuniary value of MAUD to be over $5 million using a prevailing rate of $500 per hour in M&A legal fees.

## 2 Related Work

Due to the high costs of contract review and the specialized skills it requires, understanding legal text has proven to be a ripe area for NLP research.

**Information Extraction for Legal NLP.** One area of contract review research focuses on information extraction and document segmentation. Chalkidis et al. (2017) introduce a dataset for extracting basic information from contracts, with follow-up modeling work using RNNs (Chalkidis et al., 2018) and Transformers (Chalkidis et al., 2020). Lippi et al. (2019) introduce a small expert-annotated dataset for identifying "unfair" clauses in 50 online terms of services. Tuggener et al. (2020) introduce a semi-automatically constructed dataset of legal contracts for entity extraction. Leivaditi et al. (2020) introduce an expert-annotated dataset of 2960 annotations for 179 lease agreements. Hendrycks et al. (2021b) introduce CUAD, an expert-annotated contract review dataset containing 13,010 annotations for 150 legal contracts.

**Reading Comprehension for Legal NLP.** Koreeda and Manning (2021) introduce a crowdworker-annotated dataset containing 7191 Natural Language Inference questions about spans of nondisclosure agreements. Hendrycks et al. (2021a) propose a question-answering dataset sourced from freely available online materials, containing questions (including legal exam questions) from dozens of specialized areas. Zheng et al. (2021) present a multiple-choice reading comprehension dataset with 53,317 annotations automatically extracted from US case law citations. Duan et al. (2019) present a Chinese-language legal reading comprehension dataset, with about 50,000 expert-generated annotations of Chinese judicial rulings. In our work we present a legal reading comprehension dataset with 47,457 expert-generated annotations about merger agreements. To the best of our knowledge, MAUD is the only English-language legal reading comprehension dataset that is both large-scale and expert-annotated.

## 3 MAUD: A Legal NLP Dataset for Merger Agreement Understanding

MAUD consists of 47,457 annotations based on legal text extracted from 152 English-language public merger agreements. MAUD's merger agreements were sourced from the EDGAR system maintained by the U.S. Securities and Exchange Commission.

**Terminology.** *Deal points* are legal clauses that define when and how the parties in a merger agreement are obligated to complete an acquisition. We refer to the text of these clauses (extracted by annotators from merger agreements) as *deal point texts*. Multiple *deal point questions* can be asked about each deal point text; the subset of applicable questions is determined by the text's *type*. Each deal point question can be answered by one or more

| Deal Point Category | Main Dataset | Rare Answers Dataset | Abridged Dataset | All Datasets |
|---|---|---|---|---|
| Conditions to Closing | 3,411 | 298 | 4,052 | 7,761 |
| Deal Protection and Related Provisions | 6,491 | 2,280 | 5,937 | 14,708 |
| General Information | 152 | 17 | 173 | 342 |
| Knowledge | 388 | 23 | 258 | 669 |
| Material Adverse Effect | 8,816 | 871 | 3,273 | 12,960 |
| Operating and Efforts Covenant | 1,216 | 191 | 1,054 | 2,461 |
| Remedies | 149 | 0 | 181 | 330 |
| All Categories | 20,623 | 3,680 | 14,928 | 39,231 |

Table 1: Number of MAUD examples contained in each dataset by category. Each example is a question-answer pair corresponding to an extracted deal point text.

predefined *deal point answers*.

**Deal Points in MAUD.** The deal points in MAUD are standardized by the 2021 ABA Study. For the 2021 ABA Study, the American Bar Association appointed an M&A attorney to design 130 deal point questions reflecting recent legal developments and deal trends of interest. Of the 130 different deal point questions in the 2021 ABA Study, 92 are represented in MAUD.

MAUD contains 8,226 unique deal point text annotations and 39,231 question-answer annotations (i.e. examples), for a total of 47,457 annotations. Each text belongs to one of 22 deal point types (see Table 12), and the deal point type determines the subset of questions that pertain to each text. The deal point types are further grouped into seven categories (see Appendix A.11) which we use for scoring purposes.

**Task.** MAUD is a multiple-choice reading comprehension task. The model predicts the correct deal point answer from a predefined list of possible answers associated with each question. (See Figure 1 for an example). Several deal point questions in the ABA Study are in fact multilabel questions, but for uniformity we cast all multilabel questions as binary multiple-choice questions. This increases the effective number of questions from 92 to 144.

### 3.1 MAUD Datasets and Splits

MAUD contains three datasets (main, abridged, and rare answers) corresponding to three methods of generating examples. See Table 2 for the number of examples contained in each dataset.

**Main Dataset.** The main dataset contains 20,623 examples with original deal point text extracted from 152 merger agreements by expert annotators.

**Abridged Dataset.** The abridged dataset contains 14,928 examples with deal point text extracted from 94 of the 152 merger agreements included in the main dataset. Texts in the abridged dataset are joined substrings of the main texts (with the delimiter "<omitted>" indicating skipped text). Because many texts contain answers to multiple questions, we provide the abridged data to guide a model to recognize the most pertinent text.

**Rare Answers Dataset.** The rare answers dataset contains 3,680 examples that have rare answers to a question. Legal experts made small edits to texts in the main dataset to create deal points with rare answers. See Appendix A.12 for an example edit. We introduced the rare answers dataset to ameliorate imbalanced answer distributions in the main dataset. In particular, some answers in the main dataset appear in fewer than 3 contracts, making a train-dev-test split impossible.

| | train | dev | test | overall |
|---|---|---|---|---|
| **main** | 13,256 | 3,471 | 3,896 | 20,623 |
| **abridged** | 9,647 | 2,526 | 2,755 | 14,928 |
| **rare** | 2,924 | 756 | 0 | 3,680 |
| **overall** | 25,827 | 6,753 | 6,651 | 39,231 |

Table 2: The number of examples in MAUD, grouped by splits (train, dev, test) and by dataset (main, abridged, rare answers).

**Train, Dev, and Test Splits.** We construct the train-dev-test split as follows. We reserve a random 20% of the combined main and abridged datasets as the test split. The remaining main and abridged examples are combined with the rare answers data, and then split 80%-20% to form the train and dev splits. All splits are stratified by deal point question-answer pairs. Since each question belongs to one category, it follows that category proportions are

| Deal Point Category | Random | BERT | RoBERTa | LegalBERT | DeBERTa | BigBird |
|---|---|---|---|---|---|---|
| Conditions to Closing | 20.4% | 41.7% | 41.6% | 32.0% | **48.2%** | 46.6% |
| Deal Protections | 17.2% | 53.8% | 57.1% | **58.6%** | 57.9% | 58.0% |
| General Information | 23.4% | 85.7% | 81.7% | 82.0% | **87.2%** | 81.2% |
| Knowledge | 18.8% | 75.6% | **81.4%** | 71.6% | 80.9% | 81.0% |
| Material Adverse Effect | 14.5% | 44.0% | 47.7% | 49.8% | 48.8% | **50.9%** |
| Operating and Efforts Cov. | 22.0% | 84.8% | 85.7% | **89.0%** | 86.9% | 86.6% |
| Remedies | 10.9% | 88.2% | 94.3% | **100%** | 96.6% | 95.0% |
| Overall | 16.8% | 52.6% | 55.5% | 55.9% | 57.1% | **57.8%** |

Table 3: Single-task AUPR scores for each deal point category and fine-tuned model. Scores are calculated over the test split, which includes main and abridged examples but not rare answer examples. Each category score is calculated as the mean minority-class AUPR over all questions in the category and over three runs. The overall score is the mean AUPR score over all questions (not the mean over categories). See Appendix A.11 for category descriptions.

| Deal Point Category | Random | RoBERTa | LegalBERT | DeBERTa |
|---|---|---|---|---|
| Conditions to Closing | 20.4% | 40.3% | **46.2%** | **46.2%** |
| Deal Protections | 17.2% | 48.6% | **53.6%** | 53.0% |
| General Information | 23.4% | **80.2%** | 74.8% | 67.7% |
| Knowledge | 18.8% | 68.3% | **73.0%** | 71.8% |
| Material Adverse Effect | 14.5% | 48.3% | **50.7%** | 47.8% |
| Operating and Efforts Cov. | 22.0% | 80.3% | **87.3%** | 74.2% |
| Remedies | 10.9% | 51.0% | **83.6%** | 77.9% |
| Overall | 16.8% | 51.4% | **55.8%** | 53.0% |

Table 4: Multi-task AUPR scores for each deal point category and fine-tuned model. We omit BERT because the architecturally similar RoBERTa and LegalBERT outperform BERT in the single-task setting, and omit BigBird due to compute limitations.

also balanced across splits.

To avoid data leakage due to main dataset and abridged dataset examples having overlapping text and the same answer, we always split the main examples first and then place abridged examples from the same contract in the same split.

## 4 Experiments

### 4.1 Setup

**Metrics.** Many MAUD questions have an imbalanced answer distribution, so we use area under the precision-recall curve (AUPR) as our primary metric. See Appendix A.5 for details on how we average AUPR across different questions and answers.

**Models.** We fine-tune both single-task and multi-task pretrained language models on MAUD using the Transformers library (Wolf et al., 2020). In the single-task setting, we fine-tune individual pretrained LLMs for every deal point question. We evaluate the single-task performance of BERT-base (110M params), RoBERTa-base (125M params), LegalBERT-base (110M params), DeBERTa-v3-base (184M params), and BigBird-base (127M params).

In the multi-task setting, we fine-tune LLMs with one classification head for every deal point question. We evaluate the multi-task performance of RoBERTa-base, LegalBERT-base, and DeBERTa-v3-base. See Appendix A.4 for full training details.

BERT (Devlin et al., 2019) is a bidirectional Transformer that established state-of-the-art performance on many NLP tasks. LegalBERT (Chalkidis et al., 2020) pretrains BERT on a legal corpus. RoBERTa (Liu et al., 2019) improves on BERT, using the same architecture, but pretraining on an order of magnitude more data. DeBERTa (He et al., 2020) improves upon RoBERTa by using a disentangled attention mechanism and more parameters.

27.6% of the unique deal point texts in MAUD and 50.0% of texts across all examples are longer than 512 RoBERTa-base tokens, motivating our evaluation of BigBird-base. BigBird (Zaheer et al., 2020) is initialized with RoBERTa and trained on longer input sequences up to 4,096 tokens using a sparse attention pattern that scales linearly with the number of input tokens. No deal point texts in MAUD have more than 4,096 tokens. For all

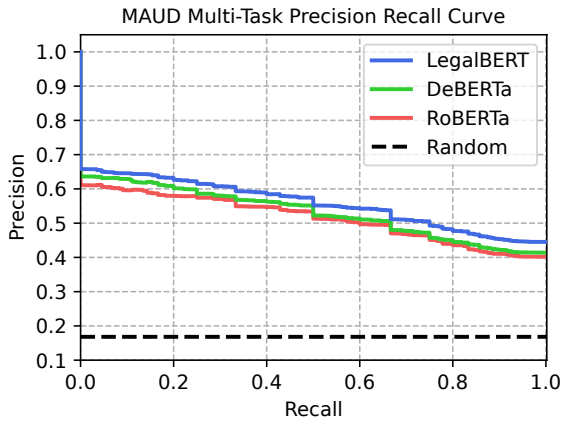


Figure 2: Precision-recall curves for multi-task models averaged over all MAUD questions.

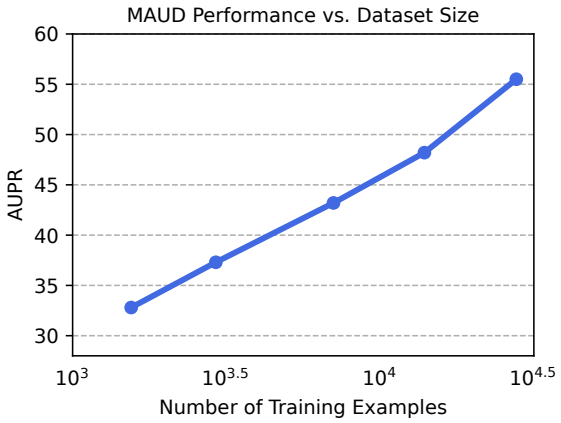


Figure 3: RoBERTa-base AUPR as a function of the number of training examples, highlighting the value of our dataset's size. AUPR is averaged over three runs.

models except BigBird we truncate texts to the first 512 tokens.

### 4.2 Results

Our baseline models achieved high AUPR scores in the Remedies, General Information, and Operating & Efforts Covenant categories, but scored lower on other categories, particularly Deal Protections & Related Provisions (best single-task AUPR 58.6%), Conditions to Closing (48.2%), and Material Adverse Effect (50.9%). Our results indicate that there is substantial room for improvement in these three hardest categories, which have the longest text lengths (see Table 9) and which attorneys also find to be the most difficult to review.

Generally, larger and newer models had higher mean performance on MAUD. In the single-task setting, DeBERTa achieved an overall score of 57.1% AUPR, compared with 55.5% for RoBERTa and 52.6% for BERT. BigBird achieved the highest score of 57.8% AUPR, slightly outperforming DeBERTa. See Tables 3 and 4 for full results on single-task and multi-task models.

**Single-Task versus Multi-Task Performance.** RoBERTa and DeBERTa single-task models outperformed their multi-task counterparts by about 4 pp AUPR. However, for LegalBERT these models had approximately the same performance.

**Effect of Pretraining on Legal Corpus.** In the single-task setting, LegalBERT outperforms BERT and slightly outperforms RoBERTa. In the multi-task setting, LegalBERT also outperforms DeBERTa. The strong performance of LegalBERT suggests that that pretraining on legal data is helpful for MAUD. See Figure 2 for a precision-recall curve comparing multi-task LegalBERT to other models.

**F1 Scores** Tables 7 and 8 present micro- and macro-averaged F1 scores for our multi-task models. LegalBERT has the highest micro F1 score in all categories excluding Knowledge, and the highest macro F1 score in all but three categories.

**Dataset Size Ablation.** We trained single-task RoBERTa models on random subsets of MAUD training data to evaluate the effect of dataset size on performance (see Figure 3). We found that RoBERTa models trained on all training examples had an overall AUPR score 7.3 percentage points higher than those trained on a 50% subset of the dataset and 23.7 percentage points higher than models trained on only a 5% subset.

**Main vs. Abridged Datasets** Table 6 reports multi-task AUPR scores over main and abridged examples separately. As expected, the abridged subscores are higher than the main subscores.

## 5 Conclusion

MAUD is a large-scale expert-annotated dataset which facilitates NLP research on a specialized merger agreement review task, based on the American Bar Association's Public Target Deal Point Study. MAUD can accelerate research towards specialized legal tasks like merger agreement review, while also serving as a benchmark for assessing NLP models in legal text understanding. Fine-tuned LLM baselines exhibit strong performance on some deal point categories, but there is significant room for improvement on the three hardest categories.

## 6 Ethics Statement

### 6.1 Data Collection

Our data was created by volunteer annotators from a non-profit legal organization, who joined the organization in order to create this dataset. None of our annotators were compensated monetarily for their time. Among our 36 annotators, 20 were male and 16 were female. 33 annotators are based in the United States and 3 annotators are based in Europe.

### 6.2 Societal Impact

Advances in ML contract review, including merger agreement review, can reduce the costs of and increase the availability of legal services to businesses and individuals. In coming years, M&A attorneys would likely benefit from having auxiliary analysis provided by ML models.

### 6.3 Limitations

MAUD enables research on models that can automate a specialized labelling task in the ABA Study, but does not target the other task performed in the ABA Study, which is the extraction of deal point texts from merger agreements.

We reserve this task for future work. For researchers interested in the deal point extraction task, we also release the 152 original contract texts and span annotations. Details on the span annotations and a preliminary baseline can be found in Appendix A.14.

The 152 merger agreements in MAUD involve the acquisitions of most but not all of the U.S. public target companies exceeding $200 million in value that were closed in 2021. Merger agreements for private companies or public companies that do not exceed $200 million in value are not included, and consequently models trained on MAUD may be less performant for deal point texts extracted for these merger agreements.

The deal point questions and the list of predefined deal point answers to each question were created by experienced M&A attorneys and standardized by the ABA, but they do not represent all of the deal points that are important in a merger agreement. MAUD should not be used as the sole source for developing AI tools for merger agreement review and drafting.

Many deal point texts exceed the maximum sequence lengths of our baseline models, and therefore we truncated texts to 512 tokens in all models except BigBird.

BigBird has a large GPU memory footprint (see Appendix A.4). Furthermore, prior work has shown that Longformer models (similar sparse attention mechanism) does not significantly outperform a 1024-token BART model on long texts, even though the BART model must truncate the long text (Shaham et al., 2022). Future research could apply other methods for enabling longer sequence lengths like ALiBi (Press et al., 2022) and FlashAttention (Dao et al., 2022) to improve performance on long deal point texts.

Our multi-task model adds a new classification head for every question. Using the extended deal point question descriptions (see Appendix A.8), future work may be able to train multi-task models that encode the question description as an input and scale to an arbitrary number of questions, or query answers from generative models using the extended descriptions for prompting.

## References


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*, 33:1877–1901.

Ilias Chalkidis, Ion Androutsopoulos, and A. Michos. 2017. Extracting contract elements. *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2018. Obligation and prohibition extraction using hierarchical RNNs. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 254–259, Melbourne, Australia. Association for Computational Linguistics.

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

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *NAACL-HLT*.

X. Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, D. Wu, S. Wang, T. Liu, Tianxiang Huo, Z. Hu, Heng Wang, and Z. Liu. 2019. CJRC: A reliable human-annotated benchmark dataset for Chinese judicial reading comprehension. *ArXiv*, abs/1912.09156.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. *ArXiv*, abs/2006.03654.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, D. Song, and J. Steinhardt. 2021a. Measuring massive multitask language understanding. In *ICLR*.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021b. CUAD: An expert-annotated NLP dataset for legal contract review. In *NeurIPS*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *ICLR*.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Spyretta Leivaditi, J. Rossi, and E. Kanoulas. 2020. A benchmark for lease contract review. *ArXiv*, abs/2010.10386.

Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018a. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018b. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. Scrolls: Standardized comparison over long language sequences.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1235–1241, Marseille, France. European Language Resources Association.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big Bird: Transformers for Longer Sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 159–168.

# A Appendix

## A.1 Licensing

MAUD is licensed under CC-BY 4.0.

## A.2 Original Merger Agreement Texts

The 152 original merger agreement texts are available as text files in the supplementary data.

## A.3 Output Constraints

There are two conditions by which certain questions or answers in MAUD can be invalidated.

**None-of-the-above Answer Constraint.** For any multi-label questions, if the "No" is label applied, then all other labels must be false. The five questions affected by this constraint are:

- W/N/A/F subject to "disproportionate impact"-Answer
- W/N/A/F applies to-Answer
- A/P/C application to-Answer
- FLS (MAE) Standard-Answer
- FLS (MAE) applies to

**Conditional Questions.** Some questions are only applicable if another question had a particular answer. These three conditional questions follow:

- Stock Deal: Fixed Ratio v. Fixed Value-Answer is valid only if the answer to Type of Consideration-Answer is "All Stock" or "Mixed Cash/Stock".
- Constructive Knowledge-Answer which is valid only if Knowledge Definition-Answer is "Constructive Knowledge".
- COR standard (board determination only)-answer is valid only if COR permitted with board fiduciary determination is "Yes".

If a conditional question is invalidated, then it does not appear in the dataset for the respective contract.

## A.4 Training details

**Training.** We fine-tune both single-task and multi-task models using the AdamW optimizer (Loshchilov and Hutter, 2018) with weight decay 0.01. We oversample to give every answer equal

| contract_name | category | text | question | answer |
|---|---|---|---|---|
| contract_93 | Material Adverse Effect | "Company Material Adverse Effect" shall mean any state of facts, change, condition, occurrence, effect, event, ... | FLS (MAE) Standard-Answer | ☑"Would" (reasonably) be expected to<br>☐"Could" (reasonably) be expected to<br>☐Other forward-looking standard |
| contract_102 | General Information | (i) each share of Company Common Stock (including each share of Company Common Stock described ... | Type of Consideration | ☑All Cash<br>☐All Stock<br>☐Mixed Cash/Stock |
| contract_77 | Conditions to Closing | Section 3.1 Organization, Standing and Power. <omitted>Section 3.2 Capital Stock. <omitted>(b) All outstanding shares of capital stock and other voting securities or ... | Accuracy of Fundamental Target R&Ws-Types of R&Ws | ☑Capitalization-Other<br>☑Authority<br>☑Approval<br>⋮<br>☐Other |

Table 5: MAUD contains three CSV files corresponding to the train, dev, and test splits of the dataset. We illustrate some example rows in the table above, using a subset of the CSV columns. For full details on the dataset's format, we refer the reader to the MAUD Data Sheet or the dataset README.

| Data Subset | Random | RoBERTa | LegalBERT | DeBERTa |
|---|---|---|---|---|
| Abridged | 22.1% | 67.5% | **73.3%** | 64.3% |
| Main | 23.1% | 65.1% | **70.0%** | 63.5% |

Table 6: Mean AUPR scores for multi-task models, computed separately for main test examples and for abridged test examples. For a fair comparison, we compute these scores only over questions that have any abridged examples. As expected, abridged subscores are higher than main subscores.

| Deal Point Category | RoBERTa | LegalBERT | DeBERTa |
|---|---|---|---|
| Conditions to Closing | 66.0% | **66.5%** | 64.0% |
| Deal Protections | 65.5% | **67.1%** | 65.3% |
| General Information | 85.5% | **86.0%** | 82.8% |
| Knowledge | **88.7%** | 87.9% | 87.9% |
| Material Adverse Effect | 79.7% | **81.0%** | 76.6% |
| Operating and Efforts Cov. | 87.5% | **89.9%** | 83.0% |
| Remedies | 90.6% | **97.4%** | 94.3% |
| Overall | 74.8% | **76.1%** | 72.8% |

Table 7: Micro-averaged F1 scores for each deal point category and multi-task model. LegalBERT, the most performant multi-task model by AUPR score (see Table 4), has the highest micro-averaged F1 scores in every category except Knowledge.

| Deal Point Category | RoBERTa | LegalBERT | DeBERTa |
|---|---|---|---|
| Conditions to Closing | 41.9% | 41.9% | **44.0%** |
| Deal Protections | 50.9% | **52.8%** | 50.7% |
| General Information | **76.3%** | 68.4% | 61.1% |
| Knowledge | **76.5%** | 75.4% | 75.3% |
| Material Adverse Effect | 61.6% | **62.3%** | 60.1% |
| Operating and Efforts Cov. | 79.1% | **82.6%** | 73.2% |
| Remedies | 68.6% | **91.7%** | 84.7% |
| Overall | 53.3% | **59.7%** | 57.3% |

Table 8: Macro-averaged F1 scores for each deal point category and multi-task model. LegalBERT has the highest F1 score in all but three categories.

| Deal Point Category | Deal Point Questions | Percent Long Texts |
|---|---|---|
| Conditions to Closing | 9 | 43.9% |
| Deal Protection and Related Provisions | 31 | 21.7% |
| General Information | 1 | 5.6% |
| Knowledge | 3 | 16.7% |
| Material Adverse Effect | 39 | 99.0% |
| Operating and Efforts Covenant | 8 | 2.1% |
| Remedies | 1 | 0.0% |
| All Categories | 92 | 50.0% |

Table 9: Number of deal point questions and long text proportions by category. "Percent Long Texts" refers to the proportion of annotations with deal point texts longer than 512 tokens when using a roberta-base tokenizer. Conditions to Closing, Deal Protection and Related Provisions, and Material Adverse Effect have the largest proportion of long texts.

proportion. We trained our final models on the combined training and development splits, averaging test AUPR scores over three runs. For all models except BigBird we truncate deal point texts to 512 tokens.

**Models.** The BERT, RoBERTA, LegalBERT, DeBERTa-v3, and BigBird pretrained language models that we use in our experiments are available on HuggingFace Hub as bert-base-cased, roberta-base, nlpaueb/legal-bert-base-uncased, microsoft/deberta-v3-base, and google/bigbird-roberta-base.

**Multi-Task Models.** For multitask experiments, we attach 144 classification heads to each model, one for each question (including each multilabel binary question) in the dataset.

For every question we maintain a different shuffled queue of training examples. Each training batch fed to the classifier consists of 16 training examples drawn in round-robin order from the question queues.

**Hyperparameter Selection.** For single-task BERT, RoBERTa, LegalBERT, and DeBERTa-v3 experiments, including the RoBERTa dataset size ablation experiment, we use batch size 16. We grid-searched over learning rates $\{1 \times 10^{-5},\ 3 \times 10^{-5},\ 1 \times 10^{-4}\}$ and number of updates $\{100, 200, 300, 400\}$.

For single-task BigBird experiments we use batch size 8. We grid-searched over learning rates $\{1 \times 10^{-5}, 1 \times 10^{-4}\}$ and number of updates $\{200, 400, 600, 800\}$.

For all multi-task models models, we used batch size 16, and grid-searched over learning rates $\{1 \times 10^{-5},\ 3 \times 10^{-5},\ 1 \times 10^{-4}\}$ and number of epochs $\{1, 2, 3, 4, 5, 6\}$. Validation AUPR scores were averaged over 3 runs.

**Infrastructure and Computational Costs.** We trained BERT and RoBERTa experiments in parallel on A5000 GPUs, using about 12GB of GPU memory. Three runs of fine-tuning models for every question with 400 updates took about one GPU-day per learning rate setting.

We trained DeBERTa-v3 experiments in parallel on A4000 GPUs, using about 20GB of GPU memory. Three runs of fine-tuning models for every question with 400 updates took about two GPU-days per learning rate setting.

We trained BigBird experiments in parallel on A4000, A5000, and A100 GPUs, choosing the minimum GPU size required to accomodate the GPU usage of the single-task model, which varied with each model's maximimum deal point text length. The single-task models corresponding to questions with the longest deal point texts required about 75 GB of GPU memory. Three runs of fine-tuning models for every question with 800 updates took 4 to 5 GPU-days per learning rate setting.

Multi-task models were trained on an A100 GPU and used more memory than their single-task equivalents. Three runs of 6 epochs of training took about 6 GPU-hours per learning rate setting and model.

### A.5 Details on MAUD AUPR Score

For every question, we calculate the minority-class AUPR score for each answer and then average to get a mean AUPR score for the question. Then we average over all question scores to get an overall AUPR score for a model.

For example, consider a deal point question $Q$, with three possible answers: $A1$, $A2$, and $A3$, which have 50, 10, and 10 test examples re-

| Deal Point Question | Description |
|---|---|
| Accuracy of Target "General" R&W: Bringdown Timing Answer | When does the general representations and warranties need to be accurate? Does the general representations and warranties need to be accurate at closing or both at signing and at closing? |
| Type of Consideration-Answer | Is the type of consideration all cash, all stock, mixed cash/stock or mixed cash/stock: election? |
| Stock Deal: Fixed Ratio v. Fixed Value-Answer | Is the consideration based on fixed ratio or fixed value? |
| MAE definition includes reference to Target "prospects" (Y/N) | Does the MAE definition includes reference to the word "prospect" or "prospects"? |
| "Ability to consummate" concept is subject to MAE carveouts | Is the impact on the ability to consummate subject to MAE carveouts? |

Table 10: Example Deal Point Question descriptions. We are in the process of writing prompt-friendly descriptions for every deal point question. These descriptions will be released with the MAUD dataset.

spectively. For the unique question-answer pair $(Q, A1)$, we first binarize all answers as $A1$ or $\neg A1$. The minority binarized answer is $\neg A1$, with 20 examples, and so the AUPR score for $(Q, A1)$ is calculated using positive class $\neg A1$. To get the AUPR score for question $Q$, we average the AUPR scores for $(Q, A1)$, $(Q, A2)$, and $(Q, A3)$.

### A.6 Best-Performing Hyperparameters

For brevity we present the over 300 combinations of best hyperparameters as CSV files in the supplementary materials.

### A.7 Evaluation Variability

We find that the average overall AUPR over three runs for our models can vary by 1-2%.

### A.8 Extended Descriptions for Deal Point Questions and Types

The supplementary materials include extended descriptions of each deal point type. These extended descriptions may be useful for prompting and providing additional context to models trained on MAUD. For the Material Adverse Effect category we also include extended descriptions for each deal point question and release these descriptions along with the dataset.

Table 10 and Table 11 present example descriptions of deal point questions and types.

### A.9 Example Annotations in the Datasets

Table 5 shows the dataset structure as well as a few example annotations.

### A.10 Other Dataset Statistics

Table 9 shows the percentage of deal point texts that are longer than 512 tokens and the number of deal point questions in each category.

### A.11 Category Descriptions

We describe the seven categories of deal points found in our dataset.

1. *General Information*. This category includes the type of consideration and the deal structure of an acquisition.

2. *Conditions to Closing*. This category specifies the conditions upon the satisfaction of which a party is obligated to close the acquisition. These conditions include the accuracy of a target company's representations and warranties, compliance with a target company's covenants, absence of certain litigation, absence of exercise of appraisal or dissenters rights, absence of material adverse effect on the target company.

3. *Material Adverse Effect*. This category includes a number of questions based on the Material Adverse Effect definition. Material Adverse Effect defines what types of event constitutes a material adverse effect on the target company that would allow the buyer to, among other things, terminate the agreement.

4. *Knowledge*. This category includes several questions based on the definition of Knowledge. Knowledge defines the standard and scope of knowledge of the individuals making representations on behalf of the target companies.

5. *Deal Protection and Related Provisions*. This category describes the circumstances where a

| Deal Point Type | Description |
|---|---|
| Ordinary course covenant | Whether the acquisition agreement requires the target company to continue its ordinary course of business, how is "ordinary course" defined and the exceptions to this obligation? |
| Negative interim operating covenant | Whether the acquisition agreement prohibits the target company from taking certain actions and the exceptions to this obligation? |
| General Antitrust Efforts Standard | The level of efforts required to obtain anti-trust clearance for the acquisition |
| Limitations on Antitrust Efforts | The limitations to the efforts required to obtain anti-trust clearance for the acquisition |
| Specific Performance | Whether in the event of a breach of the acquisition agreement, the non-breaching party is automatically entitled to specific performance? |

Table 11: Example Deal Point Type descriptions. All 22 Deal Point Type descriptions can be found in the supplementary materials and will be released with the MAUD dataset.

target company's board is permitted to change its recommendation or terminate the merger agreement in order to fulfill its fiduciary obligations.

6. *Operating and Efforts Covenants*. This category includes requirements for a party to take or not to take specified actions between the signing of the merger agreement and closing of the acquisition. The types of covenants include obligation to conduct business in the ordinary course of business and to use reasonable efforts to secure antitrust approval.

7. *Remedies*. This category describes whether a party has the right to specific performance.

### A.12 Labeling Process

MAUD is a collective effort of over 10,000 hours by law students, experienced lawyers, and machine learning researchers. Prior to labeling, each law student attended 70-100 hours of training that included live and recorded lectures by experienced M&A lawyers and passing multiple quizzes. Law students also read an instructions handbook.

Our volunteer annotators are experienced lawyers and law students who are part of a non-profit legal organization. None of the volunteers were compensated monetarily for their time. See Section 6.1 for annotator demographics.

**Data Verification.** The law students who annotated MAUD worked in teams of three. Each annotation in the main and abridged datasets was first annotated collectively, by a consensus established by a law student team. This annotation includes both the text extraction and the deal point answers. When the team of law students could not reach an agreement on an annotation, they escalated to an experienced M&A lawyer. Finally, every law student annotation was reviewed by an experienced M&A lawyer for accuracy.

We unfortunately did not retain the records necessary for us to calculate inter-annotator agreement metrics. However, lawyers reviewing the student annotations report that they agreed about 80% of the time with student annotations.

### A.13 Data Quality

We conducted a post-hoc quality check of deal point answers in one of the most difficult categories, Material Adverse Effect (MAE), with the help of an M&A attorney who was not involved in the original annotation process.

The attorney answered the MAE questions for 10 randomly selected contracts and disagreed with our gold labels 3 times out of a total of 440 labels. After conferring with our annotation team about these disagreements, the attorney decided that all three answers would be better answered by the gold label.

**Main and Abridged Datasets.** To create the main dataset and the abridged dataset, the law students conducted manual review and labeling of the merger agreements uploaded in eBrevia, an electronic contract review tool. On a periodic basis, the law students exported the annotations into reports, and sent them to experienced lawyers for a quality check. The lawyers reviewed the reports or the labeled contracts in eBrevia, provided comments,

and addressed student questions.

Where needed, reviewing lawyers escalated questions to a panel of 3-5 expert lawyers for discussions and reached consensus. Students or the lawyers made changes in eBrevia accordingly.

**Rare Answers Dataset.** To create the rare answers dataset, legal experts copied an example from the main dataset and minimally edited the deal point text to create an example with a rare answer. These edits were then reviewed by an experienced attorney to ensure accuracy.

For example, the deal point question "Limitations on Antitrust Efforts" originally had very few examples of "Dollar-based standard" deal point answer. To create examples with this rare answer, the annotators changed phrases in the deal point text similar to "no obligation to divest or take other actions" with language implying a dollar-based standard, such as "Remedy Action or Remedy Actions with assets which generated in the aggregate an amount of revenues that is in excess of USD 50,000,000."

**Final Annotation Formatting.** We exported the final annotations as three CSV files corresponding to the main, abridged, and rare answers datasets. For example rows in the dataset, see Table 5.

### A.14 Extraction Dataset and Baseline

We also release a deal point extraction dataset in SQuAD2.0 JSON format (Rajpurkar et al., 2018b), as commonly used for Extractive Question Answering datasets. The extraction dataset and the training and evaluation code for a roberta-base baseline can be found at github.com/TheAtticusProject/maud-extraction.

The annotated contract spans in this extraction dataset correspond to deal point texts in the primary reading comprehension task's main dataset.

The deal point texts extracted by the annotators are often discontiguous, with gaps ranging from a sentence to several pages in length.

**Deal Point Types.** As described in Section 3, each example in MAUD's primary reading comprehension task contains an extracted deal point text, a deal point question that can be asked of this text, and a deal point category.

Another field in each example is the deal point type. There are 22 deal point types, and each deal point type belongs to exactly one deal point category. The same set of questions are asked of all deal points with the same type.

In MAUD's extraction task, the Extractive QA model is asked to identify spans of text inside the full contract text that were annotated by legal experts as belonging to a particular deal point type.

See Table 12 for a list of deal point types and the number of examples of each type.

**Extraction Data Formatting.** For every deal point type and every contract, we format the question as follows: "Highlight the parts of the text (if any) related to "<Deal Point Type>" that should be reviewed by a lawyer."

**Extraction Dataset Splits.** We build train, dev, and test splits by splitting the 152 contracts in a 80-10-10 ratio. The deal point texts as extracted by the annotators are often discontiguous, with gaps ranging from a sentence to several pages in length. Therefore the number of spans of each type can be exceed than the number of questions.

**Metrics.** MAUD's extraction task is very similar to the contract review extraction task from Hendrycks et al. (2021b). In both tasks there is a large imbalance between the number of negative examples and positive examples in each contract. Consequently, we use the area under the precision recall curve (AUPR), averaged over three runs, as our primary metric. Following Hendrycks et al. (2021b), we consider a candidate span from our model to be a match for a lawyer-annotated span if the Jaccard similarity index is at least $50\%$.

**Training Setup.** Since the most spans in the contract text are negative examples, we oversample positive examples to create a balanced training dataset.

We fine-tune a roberta-base model on the combined train and dev datasets using an A100 GPU. We use Adam optimizer (Kingma and Ba, 2014) and batch size $40$. We use validation AUPR to select the best learning rate from $\{1 \times 10^{-5}, 3 \times 10^{-5}, 1 \times 10^{-4}\}$ and the best number of training epochs from $\{4, 6, 8\}$. The best learning rate was $1 \times 10^{-4}$ and the best number of epochs was $4$. We average our test AUPR score over three runs.

**Results.** Our RoBERTa model has an AUPR score of $19.7\%$. This is far lower than the $42.6\%$ baseline AUPR score achieved by RoBERTa in Hendrycks et al. (2021b), suggesting that our contract review extraction task is much more challenging.

| Deal Point Type | Train | Dev | Test | RoBERTa-base (AUPR) |
|---|---|---|---|---|
| Absence of Litigation Closing Condition | 25 | 6 | 2 | 5.8% |
| Accuracy of Target R&W Closing Condition | 1,120 | 153 | 161 | 29.6% |
| Agreement provides for matching rights in connection with COR | 469 | 67 | 53 | 14.6% |
| Agreement provides for matching rights in connection with FTR | 424 | 56 | 46 | 14.8% |
| Breach of Meeting Covenant | 102 | 14 | 6 | 4.3% |
| Breach of No Shop | 269 | 28 | 35 | 18.6% |
| Compliance with Covenant Closing Condition | 231 | 32 | 26 | 86.7% |
| FTR Triggers | 241 | 45 | 27 | 44.0% |
| Fiduciary exception to COR convent | 454 | 58 | 50 | 31.4% |
| Fiduciary exception: Board determination (no-shop) | 276 | 43 | 35 | 47.9% |
| General Antitrust Efforts Standard | 181 | 27 | 22 | 33.6% |
| Intervening Event Definition | 114 | 15 | 12 | 73.9% |
| Knowledge Definition | 121 | 16 | 16 | 100.0% |
| Limitations on FTR Exercise | 385 | 47 | 45 | 58.8% |
| MAE Definition | 367 | 59 | 40 | 3.2% |
| Negative interim operating convenant | 169 | 25 | 22 | 61.2% |
| No-Shop | 302 | 45 | 39 | 8.9% |
| Ordinary course covenant | 167 | 28 | 26 | 63.2% |
| Specific Performance | 134 | 18 | 20 | 81.5% |
| Superior Offer Definition | 183 | 30 | 23 | 61.4% |
| Tail Period & Acquisition Proposal Details | 360 | 47 | 40 | 16.7% |
| Type of Consideration | 217 | 26 | 21 | 69.9% |
| Overall | 6,311 | 885 | 767 | 19.7% |

Table 12: Span counts and RoBERTa-base AUPR for the different splits of the MAUD extraction dataset, grouped by deal point type. The overall deal point scores are calculated using the PR curve over all spans (not average over AUPR scores).

**Limitations.** Our RoBERTa model can only process windows of up to 512 tokens in size. However, for some deal point types, including Material Adverse Effect, most deal point texts are longer than 512 tokens. Our baseline model does not aggregate adjacent span predictions. If a gold-label span is more than twice as long as RoBERTa's context window, then it's impossible to for the model to predict a matching span with at least 50% Jaccard similarity. Future work can explore models which longer context windows or which can aggregate span extraction across multiple context windows to improve performance on these deal point spans.