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

Anonymous ACL submission

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., 2018) 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, used in the American Bar Association's 2021 Public Target Deal Points Study ("ABA Study").

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 find and categorize deal points in public merger agreements. The ABA Study is an important resource for lawyers to identify merger agreement trends which evolve with regulatory and economic conditions, but the study was suspended between 2016 and 2020 due to the extensive time required to review the merger agreements by highly trained M&A lawyers. Models trained on MAUD can learn to automate parts of this specialized merger agreement review task.

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. Each annotation was verified by three additional annotators to ensure consistency and accuracy. 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.

Legal Entity Extraction. One area of contract review research focuses on legal entity 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 an 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) intro-

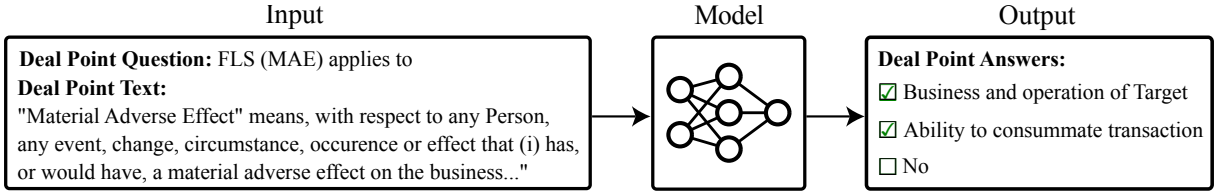


Figure 1: MAUD contains 39,500+ examples for 92 different reading comprehension questions about merger agreements. Given a *deal point question* and *deal point text*, a model learns to predict the correct answer(s) from a list of possible answers standardized by the 2021 ABA Study. The deal point text above is truncated for display.

duce 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. Unlike CUAD, which is a entity extraction task for 16 different types of contracts, MAUD is a multiple-choice reading comprehension task focusing on merger agreements.

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 47,834 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 47,834 annotations based on legal text extracted from 153 English-language public merger agreements. MAUD's merger agreements were sourced from the Electronic Data Gathering, Analysis, and Retrieval system maintained by the U.S. Securities and Exchange Commission.

Terminology. *Deal points* are legal clauses standardized by the ABA 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 text*s. One or more predefined *deal point questions* can be asked about each deal point text. Each deal point question can be answered by one or more predefined *deal point answers*. Deal point questions and texts are grouped into mutually exclusive *deal point categories*.

The 2021 ABA Study includes approximately 130 different deal point questions, 92 of which are represented in MAUD. MAUD contains 8,279 unique deal point text annotations and 39,555 question-answer annotations (i.e. examples), for a total of 47,834 annotations.

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 we take from the ABA Study are in fact multilabel questions, but for uniformity we cast all multilabel questions as binary multiple-choice questions.

3.1 MAUD Datasets and Splits

MAUD contains three datasets (main, additional, and counterfactual) corresponding to three methods of generating examples. See Table 1 for the number of examples contained in each dataset.

Main Dataset. The main dataset contains 20,764 examples with original deal point text extracted from 153 merger agreements by expert annotators.

Additional Dataset. The additional dataset contains 15,111 examples with deal point text extracted from 94 of the 153 merger agreements included in the main dataset. In the additional dataset, deal point texts are abridged to delete portions of legal text in the main dataset that are not pertinent to the deal point question. Because many texts contain answers to multiple questions, we provide the

abridged texts to guide a model to recognize the most pertinent text.

Counterfactual Dataset. The counterfactual 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.10 for example edits.

Train, Dev, and Test Splits. We construct the train-dev-test split as follows. We reserve a random 20% of the combined main and additional datasets as the test split. The remaining main and additional examples are combined with the counterfactual data, and then split 80%-20% to form the train and dev splits.

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

	train	dev	test	overall
main	13,037	3,428	4,299	20,764
add.	10,362	2,104	2,645	15,111
counter.	2,928	752	0	3,680
overall	26,327	6,284	6,944	39,555

Table 1: The number of examples in MAUD, grouped by splits (train, dev, test) and by dataset (main, additional, counterfactual).

4 Experiments

4.1 Setup

Metrics. Because many questions have an imbalanced answer distribution, we use area under the precision-recall curve (AUPR) as our primary metric. 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 respectively. 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 $(Q, A3)$.

Models. We fine-tune pretrained language models on MAUD using the HuggingFace Transformers library (Wolf et al., 2020). For simplicity, we train individual models for each deal point question. We evaluate the performance of fine-tuned BERT-base (110M params), RoBERTa-base (125M params), DeBERTa-v3-base (184M params), and BigBird-base (127M params).

BERT (Devlin et al., 2019) is a bidirectional Transformer that established state-of-the-art performance on many NLP tasks. 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. For these models, we truncate deal point texts to 512 tokens.

27.5% of the unique deal point texts in MAUD and 49.9% of texts across all examples are longer than 512 tokens,[1] so we also evaluate the performance of BigBird-base on our model. 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.

Training. We fine-tune individual models for every question and every architecture using the AdamW optimizer (Loshchilov and Hutter, 2018) with weight decay 0.01. During training, we oversample to give every answer equal proportion.

We chose the learning rate and number of updates by grid search, using the development split for validation. We chose hyperparameters corresponding to the highest mean validation AUPR score over three runs. We trained our final models on the combined training and development splits, averaging AUPR scores on the test split over three runs. See Appendix A.2 for more training details.

4.2 Results

While our fine-tuned models were able to achieve high AUPR scores in the Remedies, General Information, and Operating & Efforts Covenant categories, they achieved lower AUPR scores on other categories, particularly Deal Protections & Related

[1] Number of tokens calculated using `roberta-base` tokenizer.

Deal Point Category	Random	BERT	RoBERTa	DeBERTa	BigBird
Conditions to Closing	16.7%	47.3%	47.2%	46.2%	**54.2%**
Deal Protections and Related Provisions	16.8%	58.2%	61.8%	62.1%	**62.6%**
General Information	20.3%	**97.4%**	97.0%	94.2%	85.0%
Knowledge	17.1%	86.6%	84.0%	85.8%	**87.1%**
Material Adverse Effect	13.9%	43.5%	48.8%	**50.1%**	49.5%
Operating and Efforts Covenant	22.2%	86.9%	89.3%	**93.2%**	91.1%
Remedies	4.8%	79.8%	**100%**	**100%**	96.0%
Overall	15.9%	55.0%	58.8%	59.8%	**60.1%**

Table 2: AUPR scores for each deal point category and fine-tuned model. 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 the appendix for category descriptions.

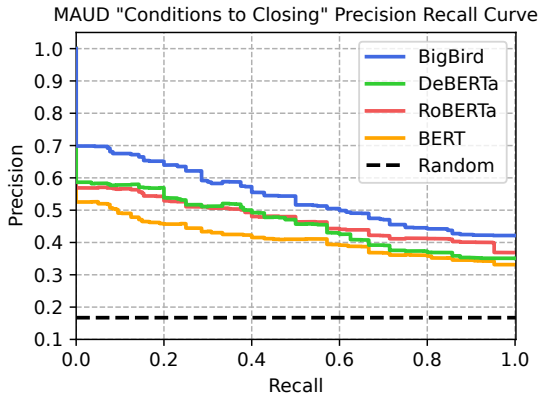


Figure 2: Precision-recall curves for our fine-tuned models on questions from the Conditions to Closing category.

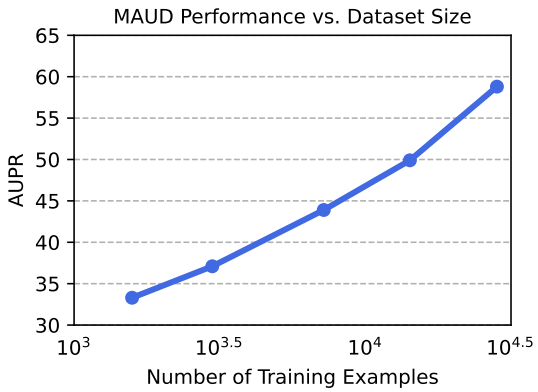


Figure 3: RoBERTa-base AUPR as a function of the number of training examples, highlighting the value of our dataset's size. AUPR is averaged over three runs.

Provisions (best AUPR 62.6%), Conditions to Closing (54.2%), and Material Adverse Effect (50.1%). Our results indicate that there is substantial room for improvement on these three hardest categories, which have the longest text lengths (see Table 5) and which attorneys also find to be the most difficult to review. See Table 2 for full results.

Overall, newer models had higher mean performance on our task, with DeBERTa achieving an overall score of 59.8% AUPR, compared with 58.8% for RoBERTa and 55.0% for BERT. BigBird (60.1% AUPR) achieved only a slightly higher overall score than DeBERTa, but significantly outperformed all other models on the Conditions to Closing category (see Figure 2).

Dataset Size Ablation. We experimented with fine-tuning RoBERTa models on a random subset of MAUD training data to evaluate the effect of dataset size on performance. We trained individual models for each question on random subsets of 50%, 25%, 10%, and 5% of the training data. We again averaged AUPR scores over three runs and selected hyperparameters for each model by grid search. We found that RoBERTa models trained using all training examples had a overall AUPR score 9.0% higher than those trained on a 50% subset of the training dataset and 25.5% higher than models trained on only 5% of the dataset. These results emphasize the importance of MAUD's size.

5 Conclusion

We introduce MAUD, a large-scale expert-annotated dataset which aims to facilitate 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 the ABA Study, while also serving as a benchmark for assessing NLP models in legal text understanding. We fine-tuned pretrained Transformer models on MAUD and find that while our models exhibit strong performance on some deal point categories, 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. More information on the annotation process can be found in Section A.10.

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 another key component of the ABA Study, which is the extraction of deal point texts from merger agreements. For an expert-annotated span extraction task for legal contracts (but not including merger agreements), we refer the reader to Hendrycks et al. (2021b).

The 153 merger agreements in MAUD involves 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.

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.

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*.

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*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. 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.

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 of 53,000+ legal holdings. 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 Training details

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

To reduce computational costs while fine-tuning BigBird models, we set the maximum input sequence length to the minimum required to encompass all deal point texts associated with the model's deal point question. In particular, this means that questions whose longest deal point text has fewer than 704 tokens were fine-tuned with full attention rather than sparse attention, because google/bigbird-roberta-base requires a sequence length of 704 or higher for sparse attention.

Grid Search For BERT, RoBERTa, and DeBERTa-v3 experiments[2] we used 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 BigBird experiments we used batch size 8. We grid-searched over learning rates $\{1 \times 10^{-5}, 1 \times 10^{-4}\}$ and number of updates $\{200, 400, 600, 800\}$.

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 model, which varied with the maximimum deal point text length. The experiments with the longest deal point text lengths 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.

[2] including RoBERTa dataset size ablation experiments

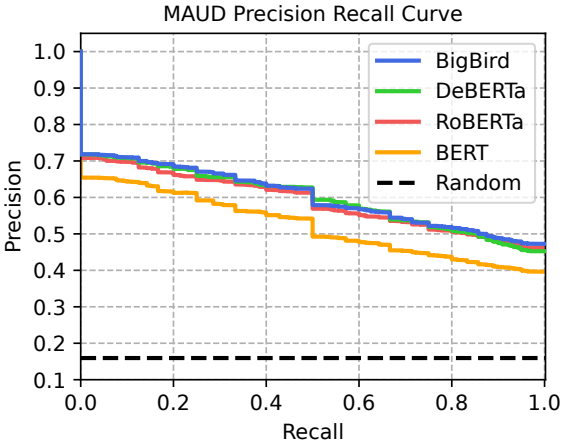


Figure 4: Precision-recall curves for our fine-tuned models on all MAUD questions.

A.3 Best-Performing Hyperparameters

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

A.4 Evaluation Variability

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

A.5 Example annotations in the Datasets

Table 3 shows the dataset structure as well as a few example annotations.

A.6 Number of Examples by Dataset and Category

Table 4 shows the number of examples in each dataset by category. Table 1 shows the number of annotations in each dataset by train-dev-test split.

A.7 Overall Precision Recall Curves

Figure 4 shows overall precision-recall curves for our fine-tuned models (averaged over all MAUD questions).

A.8 Other Dataset Statistics

Table 5 shows the percentage of deal point texts that are longer than 512 tokens and the number of deal point questions in each category.

A.9 Category Descriptions

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

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 □"Could" (reasonably) be expected to □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 □All Stock □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 ☑Authority ☑Approval ⋮ □Other

Table 3: 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.

Category	Main Dataset	Counterfactual Dataset	Additional Dataset	All Datasets
Conditions to Closing	3,436	298	4,102	7,836
Deal Protection and Related Provisions	6,536	2,280	6,016	14,832
General Information	153	17	175	345
Knowledge	391	23	262	676
Material Adverse Effect	8,874	871	3,307	13,052
Operating and Efforts Covenant	1,224	191	1,066	2,481
Remedies	150	0	183	333
All Categories	20,764	3,680	15,111	39,555

Table 4: Number of examples contained in each dataset by category. Each annotation is an answer to a deal point question corresponding to an extracted deal point text.

Deal Point Category	Deal Point Questions	Percent Long Texts
Conditions to Closing	9	43.7%
Deal Protection and Related Provisions	31	21.6%
General Information	1	5.5%
Knowledge	3	16.8%
Material Adverse Effect	39	99.0%
Operating and Efforts Covenant	8	2.1%
Remedies	1	0.0%
All Categories	92	49.9%

Table 5: 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.

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 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.10 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, an excerpt of which can be found in Figure 5.

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.

Main and Additional Datasets. To create the main dataset and the additional 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 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. Each annotation was verified by three additional annotators to ensure accuracy.

Counterfactual Dataset. To create the counterfactual 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.

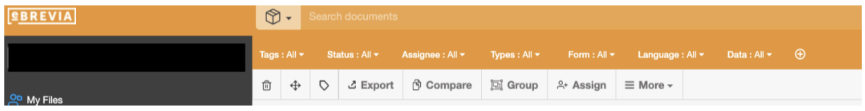
For example, the deal point question "Limitations on Antitrust Efforts" originally had very few examples of "Dollar-based standard" deal point answer. To create examples with this rare answer, the annotators changed phrases in the deal point text similar to "no obligation to divest or take other actions" 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, additional, and counterfactual datasets. For example rows in the dataset, see Table 3.

Chapter 1: Labeling Principles

General

- eBrevia Use
  - Labelling: eBrevia is the online tool that Reviewers will use to label contract clauses.
  - Form: When labelling, use the [redacted] Form (and not the "default" or any other Form).
  - Functions:
    - Only use functions in the orange toolbar below
    - **Do not** use any functions in the white toolbar below
    - **Do not** re-assign the contract to another team member
    - **Do not** change the contract status



  - My File:
    - Go to My Files to find contracts assigned to a Reviewer

Syntax Rules

- ***Note:*** *The Syntax Rules below are general and designed to apply in most cases -- however, all are subject to certain exceptions, and Reviewers should review these Syntax Rules carefully to understand when exceptions may apply and discuss with the Group leads in case there are questions.*
- **Three Types of Labels: Annotation, Answer & Text Input**

Figure 5: An excerpt from the 248-page MAUD annotator instructions handbook. In addition to reading the annotator handbook, volunteer annotators passed quizzes, attended workshops, and watched video lectures from experienced attorneys. We are currently editing the handbook for public release.

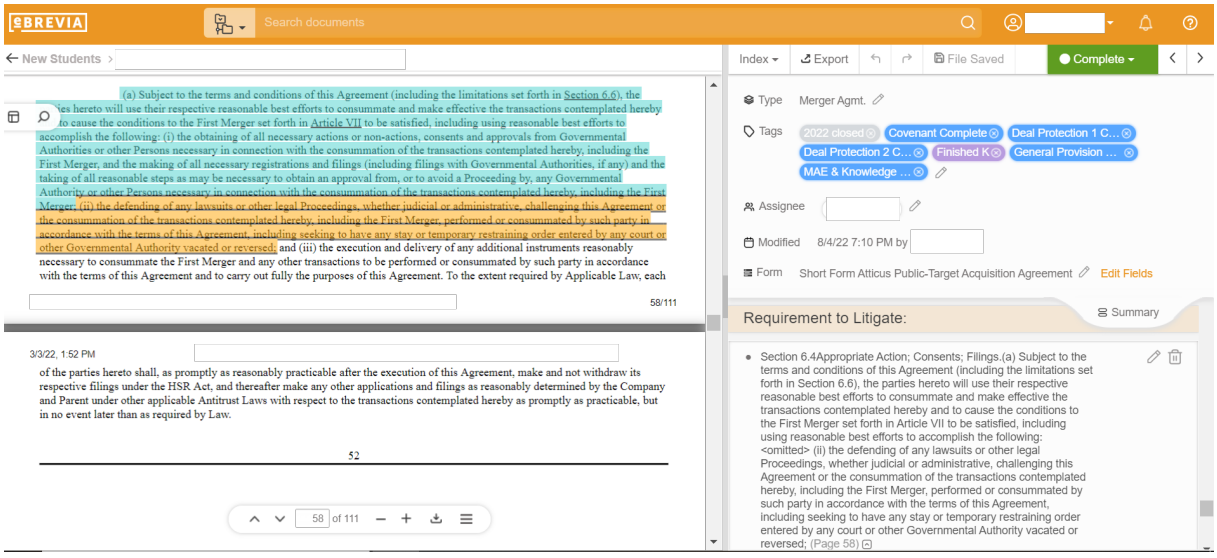


Figure 6: A screenshot of the annotation interface. Our annotators used eBrevia, a proprietary contract review tool.