**Datasheet for MAUD**

**(Merger Agreement Understanding Dataset) v1**

### **GENERAL**

#### MOTIVATION

##### Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g. company, institution, organization)?

MAUD is created by The Atticus Project, a non-profit organization founded by experienced attorneys whose mission is to harness the power of AI to accelerate accurate and efficient contract review. MAUD is used by the 2021 American Bar Association (ABA) Public Target Deal Points Study, or the ABA Study. An analysis of MAUD can be found in this paper published at https://arxiv.org/abs/2301.00876. Code for replicating the results and the trained model can be found at https://github.com/TheAtticusProject/maud.

##### Did they fund it themselves? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Yes. MAUD was created using unpaid volunteers.

##### For what purpose was the data set created? Was there a specific task in mind? If so, please specify the result type (e.g. unit) to be expected.

MAUD is a collaborative effort of attorneys, law students and AI researchers to contribute to AI advancement in merger agreement review. Specifically, MAUD was created to automate parts of the merger agreement review tasks necessary for the American Bar Association M&A Public Target Deal Points Study, or the ABA Study. 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.

#### COMPOSITION

##### A. What do the instances (of each type, if appropriate) that comprise the data set represent? (e.g. documents, photos, people, countries).

MAUD a reading comprehension dataset of over 39,000 examples and over 47,000 annotations based on legal text extracted from 152 publicly available merger agreements involving publicly traded target companies. See II-E below for more detailed descriptions of the dataset.

The deal point questions and the list of predefined deal point answers to each were created by experienced M&A attorneys and standardized by the ABA. The 2021 ABA Study includes approximately 130 different deal points questions, 92 of which are represented in MAUD.

The underlying merger agreements are sourced from the EDGAR (Electronic Data Gathering, Analysis, and Retrieval) system used by the U.S. Securities and Exchange Commission (SEC). Publicly traded target companies in the United States are required to file these merger agreements under SEC rules. Access to these agreements is available to the public for free at <https://www.sec.gov/edgar>. The 152 merger agreements include all of the 138 acquisitions involving U.S. public target companies not already majority-owned exceeding $200 million in value that were closed in 2021 and 14 acquisitions closed or to be closed in 2022.

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.

##### B. How many instances (of each type, if appropriate) are there in total?

See II-A above.

##### C. Does the dataset contain all possible instances or is it just a sample of a larger set? i.e. Is the dataset different than an original one due to the preprocessing process? In case this dataset is a subset of another one, is the original dataset available?

See II- A above.

##### D. Is there a label or a target associated with each of the instances? If so, please provide a description.

See II-E below.

##### E. What is the format of the data? e.g. .json, .xml, .csv.

We release MAUD as three separate CSV files: MAUD\_train, MAUD\_dev and MAUD\_test, representing the dataset splits used by us to generate the baseline benchmark reflected in this paper https://arxiv.org/abs/2301.00876.

MAUD contains 8226 unique text annotations and 39231 question-answer annotations (i.e. examples), for a total of 47457 annotations.

Each row of the CSVs is an input-output example that can be used to train or evaluate ML models.

The columns of the CSVs are: data\_type, contract\_name, text, answer, label, question, subquestion, text\_type, and category.

\* data\_type: The method by which the example was generated, either “main”, “abridged”, or “rare answers”.

\* contract\_name: Either an anonymized string representing the contract from which the deal point text was sourced, or “<RARE\_ANSWER>”, meaning that the deal point text was created by editing an original deal point text.

\* text: The deal point text in this example.

\* answer: The deal point answer in this example.

\* label: For convenience, an integer value corresponding to the answer value.

\* question: The deal point question in this example.

\* subquestion: If the deal point question is multiple choice, then ‘<NONE>’. If the deal point question is multilabel, then this is the value of a multilabel answer.

\* text\_type: A category describing the type of text extraction. The deal point questions associated with each deal point text is determined by the text\_type.

\* category: The deal point category for the question.

Three example rows in our CSV format follow:

| data\_type | contract\_ name | text | answer | label | question | subquestion | text\_type | category |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| main | contract\_1 | “Company Material Adverse Effect” shall mean any state of facts, change, condition, occurrence, effect, event, ... | “Would” (reasonably) be expected to | 0 | FLS (MAE) Standard-Answer | <NONE> | MAE Definition | Material Adverse Effect |
| main | contract\_2 | (i) each share of Company Common Stock (including each share of Company Common Stock described ... | All Cash | 0 | Type of Consideration | <NONE> | Type of Consideration | General Information |
| main | contract\_3 | 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 ... | Authority | 1 | Accuracy of Fundamental Target R&Ws-Types of R&Ws | Authority | Accuracy of Target R&W Closing Condition | Conditions to Closing |

- main dataset: This subset contains 20,623 examples with original deal point text extracted from 152 EDGAR merger agreements by expert annotators. The deal point text below is truncated for display. An annotator’s task is to pick the correct answer(s) using the deal point text.

- abridged dataset: This subset contains 14,928 examples with deal point text extracted from 94 of the 152 merger agreements included in the main dataset. In the abridged 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 deal point texts contain answers to multiple deal point questions, we provide the abridged texts to guide a model to recognize the most pertinent text.

- rare answers dataset: This subset contains 3,680 annotations that have rare answers to a question. Legal experts made small edits to texts in the main dataset to create a deal point with a rare answer. This ameliorates the imbalanced answer distribution problem prevalent in the main dataset.

The number of examples by splits (train, dev, test) and by dataset (main, rare answers, abridged) is listed below:

|  | Train | Dev | Test | Overall |
| --- | --- | --- | --- | --- |
| main | 13,256 | 3,471 | 3,896 | 20,623 |
| abridged | 9,647 | 2,526 | 2,755 | 14,928 |
| rare answers | 2,924 | 756 | 0 | 3,680 |
| overall | 25,827 | 6,753 | 6,651 | 39,231 |

The number of examples in each dataset by category is listed below:

| Category | Main Dataset | Rare Answers Dataset | Abridged Dataset | All Datasets |
| --- | --- | --- | --- | --- |
| Conditions to Closing | 3411 | 298 | 4052 | 7761 |
| Deal Protection and Related Provisions | 6491 | 2280 | 5937 | 14708 |
| General Information | 152 | 17 | 173 | 342 |
| Knowledge | 388 | 23 | 258 | 669 |
| Material Adverse Effect | 8816 | 871 | 3273 | 12960 |
| Operating and Efforts Covenant | 1,216 | 191 | 1054 | 2461 |
| Remedies | 149 | 0 | 181 | 330 |
| All Categories | 20,623 | 3,680 | 14,928 | 39,231 |

##### F. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g. because it was unavailable). This does not include intentionally removed information, but might include, e.g. redacted text.

See II-A above.

In addition, the abridged dataset and the main dataset have overlapping deal point texts and the same answers. To avoid data leakage, we place all abridged and main annotations from the same merger agreements in the same train-dev-test split.

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

The deal point text includes reference to extraneous text such as “<omitted>” and “(Page [number])” that were not part of the underlying merger agreements:

* Some deal point text in the datasets includes multiple legal clauses extracted from different parts of a merger agreement that, only when read together, represent the full text needed to answer a deal point question. In these cases, we insert the symbol “<omitted>” to indicate that the text was not continuous in the merger agreement; and
* Page numbers in the format of “(Page [number])” were added to indicate the page numbers of the merger agreement where the deal point text was extracted from.

##### G. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. Do not include missing information here.

See II-F above.

##### H. Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.

See II-D above.

##### I. Are there recommended data splits, e.g. training, development/validation, testing? If so, please provide a description of these splits explaining the rationale behind them.

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 annotations are combined with the rare answers data, and then split 80%-20% to form the train and dev splits.

To avoid data leakage due to main dataset and abridged dataset examples having overlapping text and the same answer, we place main and abridged examples from the same contract in the same split. All splitting is stratified by question-answer pairs.

##### J. Is the dataset self-contained, or does it link to or otherwise rely on external resources? e.g., websites, tweets, other datasets. If it links to or relies on external resources, a) Are there any guarantees that they will exist, and remain constant over time? b) Are there official archival versions of the complete dataset? i.e. including the external resources as they existed at the time the dataset was created. c) Are there any restrictions (e.g. licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, if appropriate.

MAUD includes deal point text extracted from the underlying 152 merger agreements. These agreements are publicly available for free at <https://www.sec.gov/edgar>. We do not anticipate that the U.S. Securities and Exchange Commission will remove, change or restrict the access to these agreements.

##### K. Does the dataset contain data that might be considered confidential? e.g. data that is protected by legal privilege or by doctor patient confidentiality, data that includes the content of individuals non-public communications. If so, please provide a description.

No.

##### L. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

##### M. Does the dataset relate to people? If so, please specify

No.

##### N. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

##### 

No.

#### COLLECTION PROCESS

##### How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

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 must attend 70-100 hours of training that included live and recorded lectures by experienced M&A lawyers, follow a 250-page annotation guideline, and pass multiple quizzes.

To create the main dataset and the abridged dataset, the law students then 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 abridged annotators to ensure consistency and accuracy. The final annotations were exported into the main dataset and the abridged dataset in CSV files.

To create the rare answers dataset, legal experts copied deal point text from the main dataset and minimally edited it to derive the rare answers. The annotations were then reviewed by an experienced attorney to ensure accuracy.

We combined the main, abridged, and rare answers datasets, and split them into the MAUD\_train.csv, MAUD\_dev.csv, and MAUD\_test.csv files (See II-E above for formatting details).

##### What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

See III-A above.

##### If the dataset is a sample from a larger set, what was the sampling strategy? i.e. deterministic, probabilistic with specific sampling probabilities.

Yes, see II-A above.

##### Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Our annotators are experienced lawyers and law students. All of them are unpaid volunteers in creating MAUD.

##### Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The merger agreements were downloaded from EDGAR between April 2021 and January 2022, and manually labeled between April 2021 and January 2022. The final datasets were created between April 2021 through December 2022.

##### Did the collection process involve the participation of individual people? If so, please report any information available regarding the following questions: Was the data collected from people directly? Did all the involved parties give their explicit consent? Is there any mechanism available to revoke this consent in the future, if desired?

No.

##### Has an analysis of the potential impact of the dataset and its use on data subjects been conducted? i.e. a data protection impact analysis. If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Not applicable. The dataset is based on publicly available merger agreements.

##### Were any ethical review processes conducted?

Not applicable. The dataset is based on publicly available merger agreements.

##### Does the data come from a single source or is it the result of a combination of data coming from different sources? In any case, please provide references.

The contracts are sourced from the EDGAR system maintained by the SEC at <https://www.sec.gov/edgar>.

##### Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?

MAUD enables research on models that can automate the reading comprehension 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 the merger agreements. Accordingly, MAUD achieves an important step toward automating the ABA Study, but still has a long way to go to achieve the desired motivation.

#### IV. PREPROCESSING/CLEANING/LABELLING

##### Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

See II-A and III-A above.

##### Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

Yes, the underlying merger agreements sourced from the EDGAR system are publicly available for free at <https://www.sec.gov/edgar>.

##### Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the software used to label the contracts is provided by eBrevia (https://ebrevia.com).

#### USES

##### Has the dataset been used already? If so, please provide a description.

Yes, MAUD was used by the 2021 American Bar Association Public Target Deal Points Study.

##### Is there a repository that links to any or all papers or systems that use this dataset? If so, please provide a link or any other access point.

Visit The Atticus Project website at [atticusprojectai.org](https://www.atticusprojectai.org/), and The Atticus Project github account at: <https://github.com/TheAtticusProject>

##### What (other) tasks could the dataset be used for? Please include your own intentions, if any.

See I-C above.

##### Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset does not constitute legal advice. The purpose of MAUD is to facilitate development of AI tools to aid legal professionals, but should not be used as a replacement for legal advice. The deal point questions and the list of predefined deal point answers to each 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.

We make no representations or warranties regarding the underlying merger agreements or the deal point text extracted therefrom, which are publicly available and downloadable from EDGAR.

In addition, the use of the dataset is subject to our privacy policy at <https://www.atticusprojectai.org/privacy-policy> and disclaimer at <https://www.atticusprojectai.org/disclaimer>.

##### Are there tasks for which the dataset should not be used? If so, please provide a description.

See V-D above.

##### If others want to extend/augment/build on this dataset, is there a mechanism for them to do so?

MAUD is a highly specialized dataset that requires M&A legal expertise. The Atticus Project has an Open Data Fellow Program. Anyone interested in extending and augmenting MAUD for public use can apply to the Open Data Fellow Program by submitting your application at <https://www.atticusprojectai.org/open-data-fellow-programs>.

#### VI. DISTRIBUTION

##### A. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes. The dataset is open-sourced and publicly available.

##### B. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset can be downloaded from our website at [https://www.atticusprojectai.org/maud](https://www.atticusprojectai.org/).

##### C. When will the dataset be distributed?

MAUD v1 was released on January 2, 2023.

##### D. Is the dataset distributed under a copyright or other intellectual property (IP) license? And/or under applicable terms of use (ToU)? Please cite a verified source.

The dataset is licensed under the [Creative Commons Attribution 4.0 (CC BY 4.0)](https://creativecommons.org/licenses/by/4.0/) license and free to the public for commercial and non-commercial use. https://creativecommons.org/licenses/by/4.0/

Note: we make no representations or warranties regarding the license status of the underlying merger agreements or the deal point text extracted therefrom, which are publicly available and downloadable from EDGAR.

In addition, the use of the dataset is subject to our privacy policy at <https://www.atticusprojectai.org/privacy-policy> and disclaimer at <https://www.atticusprojectai.org/disclaimer>.

##### E. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

Not applicable. Note: we make no representations or warranties regarding the underlying merger agreements or the legal text extracted therefrom, which are publicly available for free and downloadable from EDGAR.

##### F. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Not applicable.

#### VII. MAINTENANCE

##### Who will be supporting/hosting/maintaining the dataset?

The dataset is hosted and maintained by The Atticus Project, a non-profit organization.

##### How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Contact us via email at info@atticusprojectai.org.

##### Is there an erratum? If so, please provide a link or other access point.

##### 

None at this time. All changes to the dataset will be published on The Atticus Project website (https://www.atticusprojectai.org).

##### Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

All changes or updates to the dataset will be published on The Atticus Project website (https://www.atticusprojectai.org).

##### If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Not applicable.

##### Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes. Older versions of the data will continue to be publicly available on the Atticus Project website (<https://www.atticusprojectai.org>) and (<https://github.com/TheAtticusProject>) unless otherwise indicated therein.

##### If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We welcome your contribution to the dataset. Email us at [info@atticusprojectai.org](mailto:info@atticusprojectai.org)