Mining Legal Arguments in U.S. Corporate Case Law

Anonymous ACL submission

Abstract

Argument mining in legal texts supports tasks such as passage classification, retrieval, and argument completion. This work introduces a dataset comprising 42 expert-annotated U.S. corporate reorganization cases under I.R.C. §368, together with a syllogism-centered annotation framework designed to represent chained legal reasoning. The framework classifies text spans according to legal function: Rule, Analysis, Conclusion, Background Facts, and Procedural History. It further connects argumentative spans into support trees modeled on IRAC-style reasoning. In contrast to flatter claim and premise schemes, this approach aims to capture intermediate reasoning steps that connect doctrinal rules to case-specific analysis. The corpus is released in span-based, sentence-based, flat, and tree-structured formats, along with annotation guidelines and agreement analysis. Experimental results indicate that the functional labels are learnable in case-disjoint classification, and that structured queries are most helpful for within-case argument completion, whereas broader cross-case retrieval remains challenging. This dataset provides a novel resource for studying legal argument structure within a narrowly defined but significant area of U.S. case law.

1 Introduction

Research on legal argument mining frequently focuses on precedent and judicial reasoning within common-law systems (Valvoda et al., 2021). However, the availability of relevant data remains limited. Most existing resources are derived from European case law and civil-law contexts. For instance, Demosthenes is a corpus comprising 40 decisions from the Court of Justice of the European Union on fiscal state aid, annotated for premises, conclusions, and argument schemes (Lewis, 2023; Grundler et al., 2022). In contrast, the United States operates under a common-law system that emphasizes judicial precedent, where only the majority opinion constitutes binding precedent for future cases (Administrative Office of the U.S. Courts, 2022; Merryman and Pérez-Perdomo, 2019; David and Brierley, 1985). Consequently, well-annotated U.S. judicial opinions are particularly valuable for the study of legal reasoning.

To address this gap, we present a new expert-annotated dataset of U.S. federal case law. The dataset focuses on opinions about corporate reorganizations under I.R.C. §368. This is a useful and distinctive domain. Section 368 defines several specific forms of corporate reorganization, including statutory mergers, stock acquisitions, asset acquisitions, recapitalizations, and changes in corporate form (Office of the Law Revision Counsel, U.S. House of Representatives, 2026). The relevant doctrine is governed by precise legal tests. Typically, a qualifying reorganization must satisfy the continuity of business enterprise requirement and, in many cases, the continuity of interest requirement (Electronic Code of Federal Regulations, 2025). Accordingly, these opinions frequently apply structured statutory rules to detailed factual scenarios in a systematic manner. This characteristic makes the domain particularly suitable for expert annotation of legal reasoning.

We release annotations for 42 decisions, including a double-annotated subset to facilitate agreement analysis. Additionally, we supply a comprehensive description of the annotation process, a description of the annotation guidelines, dataset statistics, and both span-based and sentence-based versions of the corpus. Furthermore, we release both tree-structured and flat annotation formats to support downstream tasks, such as information retrieval and structured reasoning experiments.

We accompany the dataset with a straightforward annotation framework designed to capture chained legal reasoning. Although legal opinions are presented in plain text, their reasoning typically unfolds through a sequence of local infer-

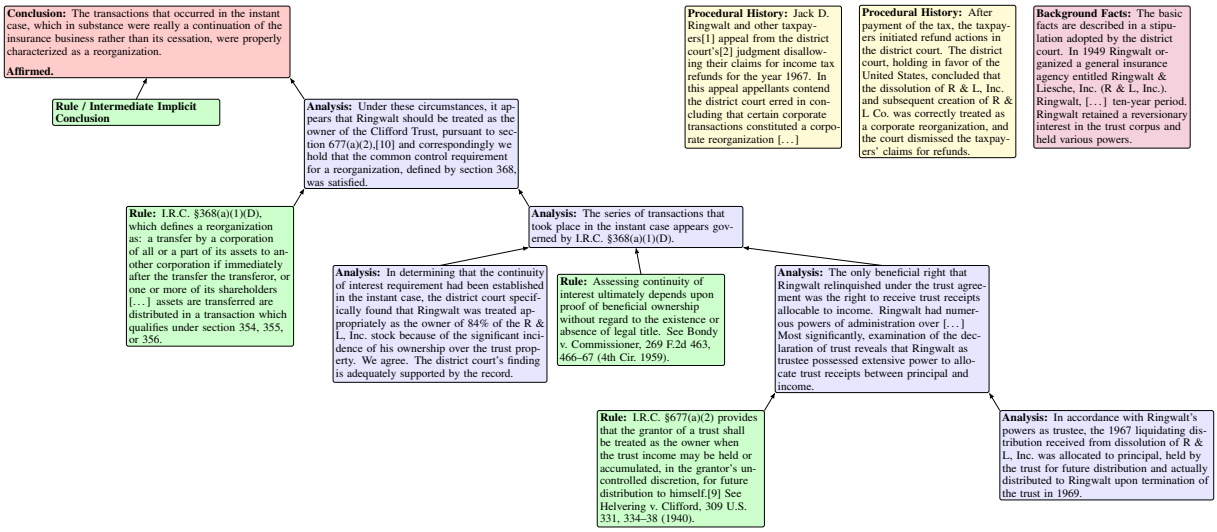

Figure 1: Condensed view of syllogistic argument tree annotation of the case Ringwalt v. U.S., 549 F.2d 89 (8th Cir. 1977). *Background Facts* and *Procedural History* are included in the annotation but are not considered part of the argument structure, as they are defined as contextual spans rather than as support for the argument's claims.

ences, where intermediate conclusions support subsequent steps (Gardner and Bartholomew, 2020). This structure is significant for legal NLP tasks such as identifying passages that justify outcomes, completing missing reasoning steps, and retrieving support for claims (Hou et al., 2025; Šavelka and Ashley, 2016). Our framework labels spans according to their function (*Rule*, *Analysis*, *Conclusion*, *Background Facts*, *Procedural History*) and organizes them into argument trees. The objective is to provide a clear and consistent representation that enhances the dataset's utility for classification, retrieval, and within-case argument completion. A condensed annotation example is presented in Figure 1.

2 Related Work

Argument mining in legal texts is frequently defined as the automatic identification and extraction of inference and reasoning, as expressed in natural language arguments (Lawrence and Reed, 2019). This process involves identifying premises and conclusions and predicting their relationships. The following section outlines methods that closely align with this definition and have been applied to legal arguments similar to those examined in this study.

One line of research focuses on annotating explicit argument structures in court decisions. Yamada et al. (2019) construct a manually annotated corpus of Japanese judgment documents for structure-based summarization. Poudyal et al. (2020) release 42 decisions from the European Court of Human Rights, annotated at the clause level with premise, conclusion, and non-argument labels, as well as links between clauses. Grundler et al. (2022) introduce Demosthenes, a corpus of 40 CJEU fiscal state aid decisions annotated for argumentative elements, their types, and argument schemes. Santin et al. (2023) apply the same corpus for argumentation-structure prediction. Habernal et al. (2024) expand this research with a larger ECHR corpus and a theory-grounded annotation scheme. Collectively, these resources demonstrate that expert annotation of legal reasoning facilitates retrieval, summarization, and structure prediction.

A second line of research focuses on labeling legal texts according to their discourse function. Bhattacharya et al. (2019) annotate Indian Supreme Court judgments with sentence-level rhetorical roles. Malik et al. (2022) expand this approach by introducing expert annotations and 13 fine-grained roles, including facts, arguments, statute, issue, precedent, ruling, and ratio. In the United States, Savelka and Ashley (2018) segment U.S. court opinions into seven functional and issue-specific parts, such as *Background*, *Analysis*, and *Conclusions*. More recently, Csányi et al. (2025) present a human-annotated corpus of sentence-level rhetorical roles for Hungarian judicial decisions. These resources are closely related to the present work because they label the legal function of text and facilitate tasks such as summarization, search, and document structuring. However, most operate at the sentence or segment level and typically do not

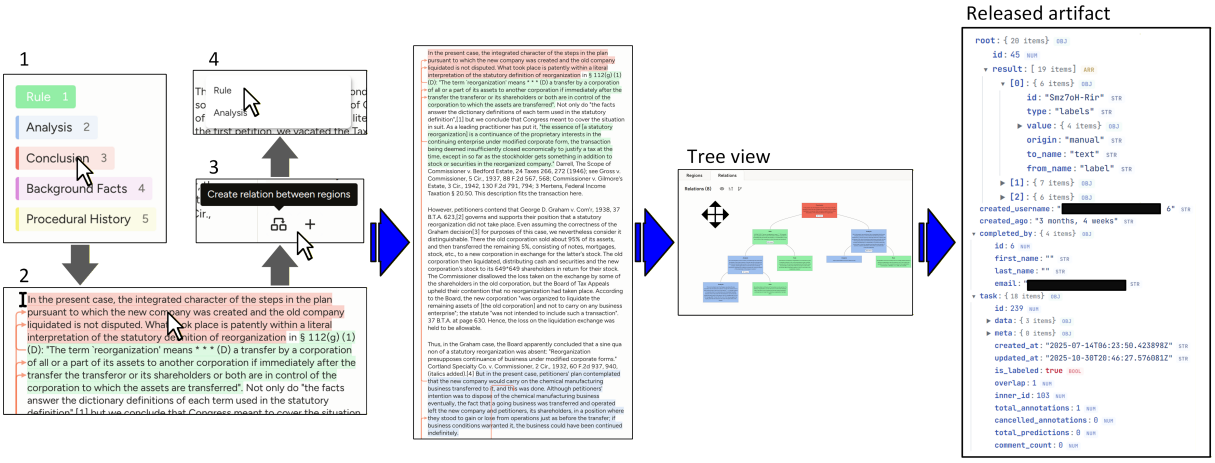

Figure 2: Annotation workflow, visualization of the annotation interface, structure tree visualization, and an example of an annotated case data format. The workflow: (1) selecting the label; (2) highlighting the text, which automatically applies the label's color and saves the annotation; (3) creating edges by pressing the 'create relation' button with one span selected, then selecting another annotation; and (4) for Intermediate Implicit Conclusions, right-clicking displays available labels and inserts a new block. The resulting artifact is a JSON file containing all annotations as positional references, with edges defined by the IDs of the connected annotations.

encode explicit support links between spans.

Prior research in North America has also addressed structured legal reasoning. Walker et al. (2017) introduce the veterans' claims dataset, which annotates U.S. adjudicatory decisions with sentence roles and propositional connective types. Xu et al. (2020) annotate Canadian case summaries and full texts with Issue–Reason–Conclusion triples for legal summarization. While these projects serve as important precursors, they address different materials and research objectives.

This work extends previous efforts by releasing an expert-annotated dataset of U.S. federal tax opinions concerning corporate reorganizations under I.R.C. §368. The annotation framework offers a practical method for representing chained reasoning, and the corpus focuses on a specific but significant doctrinal area. It also provides explicit links between annotated spans, includes a double-annotated subset, and is available in multiple formats to support classification, retrieval, and within-case argument completion.

3 Corpus Creation

We collected 42 U.S. corporate reorganization cases (1k–10k words), focusing on I.R.C. §368(a)(1)(A),(B),(C),(D),(F) and excluding (E),(G) to limit statutory variety. Two law students annotated all 42 documents. Ten documents were double-annotated to estimate inter-annotator agreement, and a professor of law led adjudication. Annotation was conducted in a customized version of the annotation platform Label Studio, that supports span labeling, directed links between spans, and a tree structure visualization of the annotations.

Table 1 summarizes the corpus scale, label distributions, and tree-level structural properties of the released annotations.

3.1 Annotation Procedure

Annotators selected free spans of text expressing atomic units of reasoning and assigned each span a functional label. They then added directed links to connect spans into argument trees that follow a syllogistic pattern (Gardner and Bartholomew, 2020). Each step represents a local inference supporting a downstream claim, enabling the reasoning process to be modeled as a tree. When a premise was implicit, annotators could annotate enthymemes and, when necessary, insert implicit intermediate conclusions as placeholders to maintain structural consistency. Figure 2 provides a graphical representation of the annotation process.

3.2 Argumentation Scheme

Labels. The scheme uses five labels. *Rule* marks generally applicable statements, including legal rules, tests, and other abstract criteria. *Analysis* marks case-specific reasoning that applies rules to the facts and often captures intermediate conclusions by function. *Conclusion* is reserved for the fi-

nal outcome of an argument tree. *Background Facts* and *Procedural History* mark contextual spans that provide narrative support for the opinion but do not participate in the reasoning.

Relations and constraints. Spans can be linked with directed support relations to form trees. *Rule* and *Analysis* spans are treated as part of the argumentative structure and must have a directed path to a terminal *Conclusion*. *Background Facts* and *Procedural History* may be annotated, but remain disconnected from the argument tree. *Conclusion* spans are terminal nodes and cannot support other conclusions.

Intermediate implicit conclusions and enthymemes. Enthymemes are usually defined as arguments that rely on one or more implicit premises (Feng and Hirst, 2011). In our syllogistic approach, an enthymeme is an abbreviated syllogism in which one premise is left unstated and must be inferred from shared background knowledge. More precisely, for annotation purposes, we define it as an atomic syllogism with an unstated *analysis*, or *rule*, as defined in our scheme. In practice, it is often expressed as a causal relation. We instructed annotators to use this label only when strictly necessary and only when the missing component of the argument could not be found explicitly.

3.3 Dataset curation

Adjudication. The adjudicator determines which double-annotated cases are included in the final released dataset. Decisions are made for the entire case rather than for individual components, and the adjudicator does not directly modify annotations once they are finalized. Suggestions and corrections to annotations are provided only when they conflict with the established guidelines.

Span-to-sentence mapping. Span-to-sentence mapping was employed to generate a sentence-level version of the dataset for retrieval experiments and to provide an alternative annotation format. This approach projects span annotations onto sentences using exact character offsets. The process involves verifying that annotated spans align precisely with the source text, splitting the document into sentence units based on offset boundaries, and assigning each sentence to the overlapping span with the greatest offset overlap, according to a predetermined tie-breaking rule. Each sentence inherits the label of its corresponding span, while sentences without a matching span are excluded. This method yields a deterministic sentence-level representation in which a single span may correspond to multiple sentences, but each sentence is assigned no more than one label.

Corpus and structure summary	
Cases	42
Total words	150,040
Average words / case	3,572.38
Median words / case	2,989.50
Total sentences	5,286
Average sentences / case	125.86
Median sentences / case	105.50
Explicit spans	718
Average spans / case	17.10
Median spans / case	14.50
Nodes / edges	800 / 644
Argument trees	43
Implicit insertions	82
Disconnected spans	132
Average depth	2.38
Max depth	10
Average branching	2.18

	Explicit spans ($n = 718$)		**All nodes** ($n = 800$)		**Sentences** ($n = 2715$)	
Label	Count	%	Count	%	Count	%
BF	59	8.22	59	7.38	1,420	52.30
PH	56	7.79	56	7.0	160	5.89
Rule	206	28.69	243	30.38	428	15.76
Analysis	353	49.16	398	49.75	620	22.84
Conclusion	44	6.13	44	5.50	87	3.20

Table 1: Descriptive corpus statistics. Sentence-label percentages are computed over the 2,715 labeled candidate sentences used in sentence-level alignment, not all 5,286 case sentences. BF=*Background Facts*; PH=*Procedural History*.

3.4 Inter-Annotator Agreement

Agreement was evaluated on 10 double-annotated cases. For node labels, Krippendorff's unitized α_u was computed by treating each label as a binary segmentation task over character offsets, utilizing a length-weighted coincidence matrix with background text. Results are reported for both the span-based and sentence-based versions of the dataset. Span-level soft-F1 was also computed using maximum-weight 1-to-1 matching under two pairing strategies: edit-distance and semantic matching. Table 2 presents a summary of the results.

Node-label agreement was high for *Background Facts* (0.879) and *Conclusion* (0.724), and moderate for *Rule* (0.613), *Analysis* (0.567), and *Procedural History* (0.454) in the span view. The sentence view yielded nearly identical α_u values, indicat-

(a) Explicit, text-anchored agreement

Label	α_u^{span}	α_u^{sent}	F1_{edit}	F1_{sem}
Analysis	0.57	0.57	0.46	0.56
Background Facts	0.88	0.88	0.70	0.79
Conclusion	0.72	0.72	0.70	0.85
Procedural History	0.45	0.46	0.47	0.59
Rule	0.61	0.61	0.51	0.64
Macro avg.	0.65	0.65	0.57	0.69

(b) Implicit IC insertion: contingency counts

Match	Ctx	YY	YN	NY	NN
Edit-distance	62	2	7	15	38
Semantic	64	2	7	16	39

(c) Implicit IC insertion: agreement summary

Match	P_o	P_+	P_-	κ	$\mathbf{R}_{A1/A2}$	$\mathbf{E}_{A1/A2}$
Edit-distance	0.65	0.15	0.78	-0.04	0.15 / 0.27	9 / 24
Semantic	0.64	0.15	0.77	-0.05	0.14 / 0.28	9 / 26

Table 2: Inter-annotator agreement on the 10 double-annotated cases. Panel (a) reports agreement on explicit, text-anchored spans. Sentence-level α_u projects span annotations to sentence units. F1 is computed using either edit-distance-based or semantic span pairing. Panels (b–c) report agreement on whether an implicit intermediate conclusion (IC) is inserted between aligned explicit spans. P_o is observed agreement, and P_+/P_- are positive/negative agreement for insertion vs. non-insertion. $\text{R}_{A1/A2}$ is the per-annotator insertion rate over candidate contexts; $\text{E}_{A1/A2}$ is the number of usable inserted implicit nodes that can be mapped to an evaluable comparison context for agreement scoring.

Pairing	Contexts	Observed	Expected	κ
Edit distance	95	0.074	0.414	-0.581
Semantic	97	0.072	0.418	-0.593

Table 3: Direct-edge agreement on matched explicit-span contexts. Observed agreement remains low and κ is negative under both pair-matching strategies.

ing that converting spans to sentences has minimal impact on unitized agreement. In contrast, overlap-based F1 was notably higher in the sentence view than in the span view for *Analysis*, *Background Facts*, and *Rule*, suggesting that disagreement for these labels is partly attributable to span-boundary variation. Agreement was substantially weaker for implicit intermediate-conclusion insertion ($\kappa \approx -0.05$) and weakest for direct-edge annotation (observed agreement ≈ 0.07, $\kappa \approx -0.59$; Table 3).

4 Experiments and Results

4.1 Classification Experiments

We classify passages into functional roles under two label settings. The first uses the original five labels. The second merges *Conclusion* into *Analysis*, which gives a four-class variant. We exclude implicit intermediate conclusions from all runs. This leaves 719 explicit passages from 42 cases.

We employ five-fold case-disjoint `StratifiedGroupKFold` cross-validation, grouping passages by case ID to ensure that no case appears in both training and test sets. This approach evaluates the models' ability to generalize to previously unseen cases.

We compare TF-IDF and three embedding families: SBERT, LegalBERT, and ModernBERT. We also report GPT-5-mini as a separate comparison row. Its setup differs from the embedding models, so it is not a strict baseline but we consider it as context-rich upper bound.

Table 4 presents Macro-F1 and per-class F1 scores. Merging *Conclusion* with *Analysis* increases Macro-F1 across all models. LegalBERT achieves the highest performance among embedding-based models in both labeling schemes, with Macro-F1 scores of 0.71 for the five-class task and 0.80 for the four-class task. GPT-5-mini attains the highest Macro-F1 in the five-class setting at 0.76, while LegalBERT slightly outperforms it in the four-class setting at 0.80 compared to 0.79. Random and majority baselines remain much lower in both settings.

4.2 Within-case argument completion and split-global retrieval

Task. We cast argument completion as passage retrieval. Given an incomplete argument, the model must retrieve the missing supporting sentence or sentences.

Each query has one missing slot. The processed dataset contains 403 training queries, 28 validation queries, and 40 test queries. Splits are case-disjoint. The test split contains 4 held-out cases. Each test query has 1 to 6 positive sentences.

Queries are built from span nodes and directed support links. Each terminal *Conclusion* node is treated as a motion root, and each query is built from the subtree under that root.

Candidate pools. All well-formed sentences in a case are retrievable candidates, not only annotated

Model	5 classes						4 classes				
	Macro Avg	Analysis	BF	Conclusion	PH	Rule	Macro Avg	Analysis	BF	PH	Rule
TF-IDF	0.69	0.75	0.77	0.42	**0.82**	0.69	0.78	0.81	0.80	0.79	0.70
SBERT	0.65	0.74	0.72	0.37	0.70	0.73	0.74	0.81	0.70	0.72	0.74
Legal-BERT	**0.71**	**0.77**	**0.82**	**0.45**	0.79	**0.74**	**0.80**	**0.83**	**0.81**	**0.83**	**0.75**
Modern-BERT	0.65	0.73	0.78	0.39	0.65	0.70	0.71	0.79	0.76	0.60	0.69
GPT-5-mini†	0.76	0.74	0.71	0.69	0.85	0.80	0.78	0.81	0.68	0.84	0.80
Random	0.17	0.29	0.13	0.06	0.12	0.24	0.22	0.38	0.10	0.13	0.26
Majority	0.13	0.66	0.00	0.00	0.00	0.00	0.18	0.71	0.00	0.00	0.00

Table 4: Classification experiments with five and four classes (F1-score). Macro Avg denotes macro-averaged F1. Experiments use case-disjoint `StratifiedGroupKFold` as cross-validation iterator. †GPT-5-mini uses label descriptions and additional case context, so it should be read as a context-rich upper-bound row. BF=*Background Facts*; PH=*Procedural History*.

spans. We evaluate three candidate pools: (i) *same-case*, which ranks against sentences from the same case after removing positives for other queries; (ii) *same-case full*, which ranks against all sentences from the same case; and (iii) *global split*, which ranks against all 581 sentences in the 4 held-out test cases. The average candidate-pool sizes on the test split are 138.8, 157.0, and 581.0, respectively.

Query structure. We evaluate three query designs: *structured*, *flat-masked*, and *flat-plain*. We compare five retrievers: BM25, E5, MB-base, FT-flat, and FT-struct. The two fine-tuned retrievers are dual encoders initialized from ModernBERT-base. They use a multi-positive contrastive objective.

Evaluation. We report hit rate@K, mean per-query recall@K, and exact set match@K for $K \in \{1, 5, 10, 20\}$. Figure 3 plots hit rate@K.

Results. In the *same-case* regime, BM25 and E5 are already strong. Both reach hit rate@20 of 57.5%. FT-struct reaches 55.0%, FT-flat 50.0%, and MB-base 37.5%. Mean recall@20 shows a similar pattern, with E5 and BM25 slightly ahead of FT-struct.

In the *same-case full* regime, FT-struct is the strongest system. It reaches hit rate@20 of 50.0%, mean recall@20 of 31.3%, and exact set match@20 of 20.0%. FT-flat is second among the fine-tuned models. BM25 and E5 remain competitive on hit rate, but are lower on mean recall and exact set match.

In the *global split* regime, BM25 performs best on hit rate@20 at 45.0%, and E5 is second at 37.5%. The fine-tuned retrievers are much weaker. FT-flat reaches 15.0% hit rate@20 and FT-struct 12.5%. MB-base is weaker than both fine-tuned ModernBERT systems in all regimes.

5 Discussion

Agreement patterns and limitations. The agreement results suggest that node annotations demonstrate greater reliability compared to structural annotation. The near-identical α_u values observed in both span and sentence views indicate that annotators generally agree on the discourse role of a text segment once the relevant content is identified.

Semantic pairing consistently yields higher F1 scores compared to edit-distance pairing; however, this result reflects a more permissive matching condition rather than a more robust estimate of annotation reliability. We interpret this as a sensitivity analysis, as we cannot conclude that it provides a more appropriate metric.

The weakest results pertain to implicit insertion and direct-edge annotation. In the case of implicit insertion, agreement is primarily attributable to shared decisions not to insert, while positive insertion decisions exhibit inconsistency among annotators. Direct-edge agreement is lower and remains poor under both pairing strategies, indicating that the issue is not primarily related to lexical matching. Collectively, these findings suggest that edge annotations are comparatively unstable and should be used with greater caution than node labels, potentially motivating tighter guidelines or adjudication in future annotation rounds.

Classification. The primary improvement results from removing the infrequent *Conclusion* label. This label has only 44 passages, and it creates a hard boundary with *Analysis*. The four-class results suggest that this boundary is one of the main sources of error.

The most significant remaining confusion occurs between *Rule* and *Analysis*. GPT-5-mini frequently

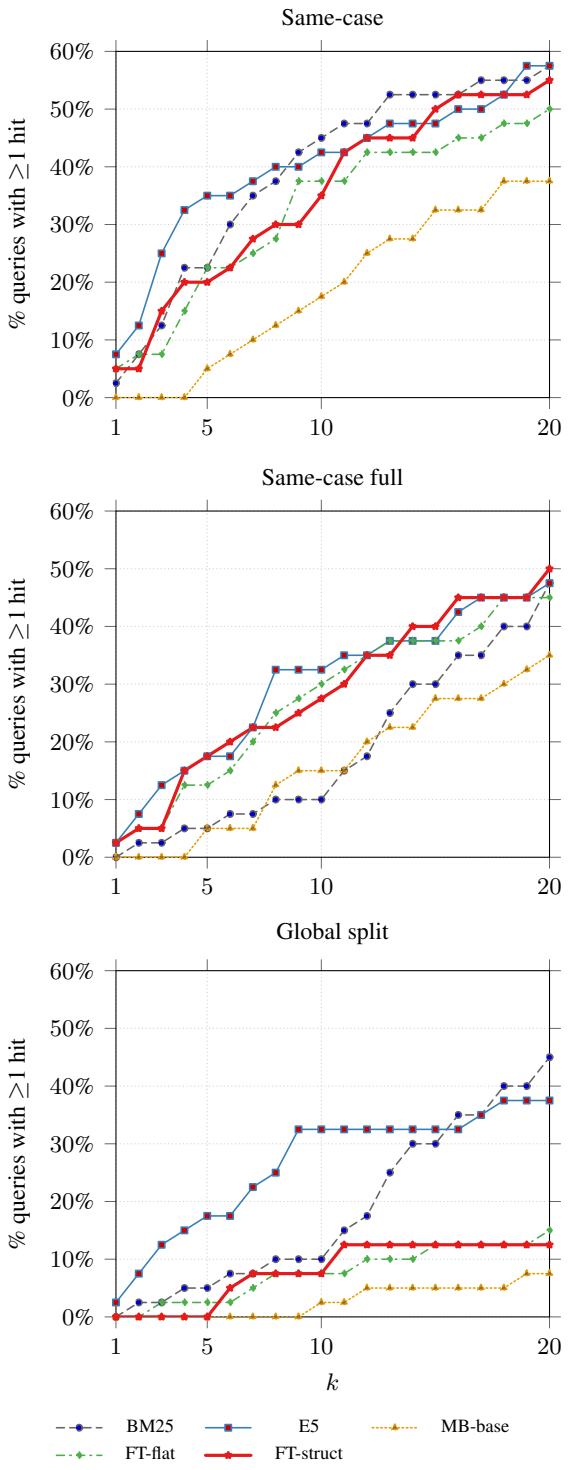

Figure 3: Hit rate@k across the three retrieval candidate pools. The retrieval methods tested are MB-base = ModernBERT-base flat-query retriever; FT-flat = fine-tuned flat-query retriever; FT-struct = fine-tuned structured-query retriever.

overpredicts *Rule* for gold *Analysis* passages, indicating a tendency to interpret case-specific reasoning as a general legal standard. LegalBERT demonstrates a more balanced error distribution, while ModernBERT more frequently makes the reverse mistake. Overall, the task is learnable, but the *Rule*/*Analysis* distinction remains the key challenge.

Retrieval. The retrieval results indicate a limited structure effect. Explicit structure helps most in the *same-case full* regime, where FT-struct gives the best balance of hit rate, recall, and exact set match. This is the clearest evidence that argument structure helps within-case completion.

However, this benefit does not extend to the global split regime. In this setting, BM25 and E5 outperform the fine-tuned retrievers. These results suggest that FT-flat and FT-struct capture patterns are effective within individual cases but fail to generalize effectively to previously unseen cases.

The *same-case* regime is also informative. BM25 and E5 are already strong in this easier pool, so it is hard to isolate a large structure effect there. Taken together, the results suggest that structure is helpful for local completion in a controlled candidate set, but not enough on its own for broader cross-case retrieval.

6 Conclusion

We presented an expert-annotated dataset of 42 U.S. corporate reorganization opinions under I.R.C. §368 and a syllogism-centered annotation framework for representing chained legal reasoning. The corpus labels spans by legal function: Rule, Analysis, Conclusion, Background Facts, and Procedural History. It organizes argumentative spans into support trees to capture intermediate reasoning steps in judicial opinions.

Our results suggest three main conclusions. First, the functional labels are reasonably learnable under case-disjoint evaluation, though the distinction between Rule and Analysis remains the main source of classification error. Second, annotation reliability is stronger for node labels than for structural features. Third, explicit structure is most useful for within-case argument completion. The structured retriever performs best in the same-case full setting, but this advantage does not transfer to the harder global split regime.

These findings position the contribution primarily as a resource and benchmark for studying legal argument structure in a narrowly defined doctrinal setting. The dataset provides a useful foundation for future work on legal passage classification, structured retrieval, and argument completion.

Limitations

The complexity of these cases' subject matters makes their large-scale annotation prohibitively expensive, constraining our dataset to only 42 high-quality annotated examples.

This study should not be regarded as a broadly generalizable result. It is based on a small, specialized corpus of English-language U.S. federal tax cases concerning corporate reorganizations under I.R.C. §368; therefore, its findings may not extend to other legal domains, legal systems, or non-legal texts. The analysis indicates that certain components of the annotation scheme are more reliable than others.

Another limitation of this study is that some of the proposed baselines needed the same cases to be annotated with different schemes to ensure reliable comparisons. Also, to make the results more general, the annotation process should also be repeated on datasets from other domains. For these reasons, we narrow the claims about this paper's contribution.

References

Administrative Office of the U.S. Courts. 2022. Understanding the federal courts. https://www.uscourts.gov/sites/default/files/understanding-federal-courts.pdf. Accessed 2026-03-16.

Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2019. Identification of rhetorical roles of sentences in indian legal judgments. In *Legal Knowledge and Information Systems – JURIX 2019: The Thirty-second Annual Conference*, pages 3–12. IOS Press.

Gergely Márk Csányi, István Üveges, Dorina Lakatos, Dóra Ripszám, Kornélia Kozák, Dániel Nagy, and János Pál Vadász. 2025. Sentence-level rhetorical role labeling in judicial decisions. *Big Data and Cognitive Computing*, 9(12).

René David and John E. C. Brierley. 1985. *Major legal systems in the world today : an introduction to the comparative study of law*, 3rd ed. edition. Stevens, London.

Electronic Code of Federal Regulations. 2025. 26 c.f.r. s 1.368-1 — purpose and scope of exception of reorganization exchanges. https://www.ecfr.gov/current/title-26/chapter-I/subchapter-A/part-1/subject-group-ECFR8273f1baff2c569/section-1.368-1. Current e-CFR text accessed 2026-03-16.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996, Portland, Oregon, USA. Association for Computational Linguistics.

James A. Gardner and Christine P. Bartholomew. 2020. *Legal Argument: The Structure and Language of Effective Advocacy*, 3rd edition. Carolina Academic Press, Durham, North Carolina.

Giulia Grundler, Piera Santin, Andrea Galassi, Federico Galli, Francesco Godano, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. 2022. Detecting arguments in CJEU decisions on fiscal state aid. In *Proceedings of the 9th Workshop on Argument Mining*, pages 143–157, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2024. Mining legal arguments in court decisions. *Artificial Intelligence and Law*, 32:1–38.

Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2025. CLERC: A dataset for U. S. legal case retrieval and retrieval-augmented analysis generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7898–7913, Albuquerque, New Mexico. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Carl Emilio Lewis. 2023. The european court of human rights and its search for common values. *European Convention on Human Rights Law Review*, 4(2):179 – 220.

Vijit Malik, Rishabh Sanjay, Shouvik Kumar Guha, Angshuman Hazarika, Shubham Kumar Nigam, Arnab Bhattacharya, and Ashutosh Modi. 2022. Semantic segmentation of legal documents via rhetorical roles. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 153–171, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

John Henry Merryman and Rogelio Pérez-Perdomo. 2019. *The civil law tradition : an introduction to the legal systems of Europe and Latin America*, fourth edition. edition. Stanford University Press, Stanford, California.

Office of the Law Revision Counsel, U.S. House of Representatives. 2026. 26 u.s.c. s 368 — definitions relating to corporate reorganizations. https://uscode.house.gov/view.xhtml?edition=prelim&num=0&req=granuleid:USC-prelim-title26-section368. Current through the preliminary 2026 U.S. Code display; accessed 2026-03-16.

Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. ECHR: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75, Online. Association for Computational Linguistics.

Piera Santin, Giulia Grundler, Andrea Galassi, Federico Galli, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. 2023. Argumentation structure prediction in cjeu decisions on fiscal state aid. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 247–256, New York, NY, USA. Association for Computing Machinery.

Jaromír Šavelka and Kevin D. Ashley. 2016. Extracting case law sentences for argumentation about the meaning of statutory terms. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 50–59, Berlin, Germany. Association for Computational Linguistics.

Jaromír Savelka and Kevin D. Ashley. 2018. Segmenting u.s. court decisions into functional and issue specific parts. In *Legal Knowledge and Information Systems: JURIX 2018*, pages 111–120. IOS Press.

Josef Valvoda, Tiago Pimentel, Niklas Stoehr, Ryan Cotterell, and Simone Teufel. 2021. What about the precedent: An information-theoretic analysis of common law. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2275–2288.

Vern R. Walker, Ji Hae Han, Xiang Ni, and Kaneyasu Yoseda. 2017. Semantic types for computational legal reasoning: Propositional connectives and sentence roles in the veterans' claims dataset. In *Proceedings of the 16th International Conference on Artificial Intelligence and Law*, pages 217–226, London, United Kingdom. Association for Computing Machinery.

Huihui Xu, Jaromír Šavelka, and Kevin D. Ashley. 2020. Using argument mining for legal text summarization. In *Legal Knowledge and Information Systems*. Open Access by IOS Press.

Hiroaki Yamada, Simone Teufel, and Takenobu Tokunaga. 2019. Building a corpus of legal argumentation in japanese judgement documents: Towards structure-based summarisation. *Artificial Intelligence and Law*, 27(2):141–170.

A Implicit intermediate conclusions per annotator label count

Ref	File	Annot.	Anal.	Rule	Tot.	AvgA	AvgR	AvgT
4	35	A1	0	0	0	–	–	–
4	85	A2	7	0	7	–	–	–
5	64	A1	1	0	1	–	–	–
5	63	A2	1	1	2	–	–	–
6	57	A1	1	1	2	–	–	–
6	90	A2	2	2	4	–	–	–
7	72	A1	1	0	1	–	–	–
7	39	A2	1	1	2	–	–	–
8	41	A1	0	0	0	–	–	–
8	84	A2	3	1	4	–	–	–
9	73	A1	1	0	1	–	–	–
9	93	A2	1	3	4	–	–	–
10	36	A1	2	1	3	–	–	–
10	95	A2	1	1	2	–	–	–
11	37	A1	0	1	1	–	–	–
11	52	A2	2	0	2	–	–	–
12	45	A1	2	0	2	–	–	–
12	51	A2	2	0	2	–	–	–
13	69	A1	0	0	0	–	–	–
13	43	A2	1	0	1	–	–	–
Summary		A1	8	3	11	0.8	0.3	1.1
Summary		A2	21	9	30	2.1	0.9	3.0

Table 5: Annotation statistics by reference ID and annotator.

B Retrieval expanded results

System	Hit rate @20	Recall @20	Exact @20	MRR @20
Same-case				
BM25	**57.5**	36.8	20.0	13.1
E5-base-v2	**57.5**	**37.0**	**22.5**	**18.4**
ModernBERT-base	37.5	22.2	12.5	4.1
Fine-tuned (flat)	50.0	30.2	15.0	12.2
Fine-tuned (structured)	55.0	35.3	20.0	13.0
Same-case full				
BM25	47.5	23.0	7.5	5.2
E5-base-v2	47.5	25.7	12.5	**10.8**
ModernBERT-base	35.0	21.1	12.5	3.6
Fine-tuned (flat)	45.0	26.4	12.5	9.0
Fine-tuned (structured)	**50.0**	**31.3**	**20.0**	9.6
Global split				
BM25	**45.0**	22.4	7.5	5.0
E5-base-v2	37.5	**22.5**	**12.5**	**9.9**
ModernBERT-base	7.5	4.2	2.5	0.6
Fine-tuned (flat)	15.0	7.1	0.0	2.0
Fine-tuned (structured)	12.5	7.7	5.0	1.6

Table 6: Aggregate retrieval results on 40 test queries (2.15 gold passages/query on average). Recall@20 is the mean fraction of gold passages recovered in the top 20. Exact@20 is exact set match over the gold set.

C Case-disjoint fold sizes

Setting	Held-out cases/fold	Held-out passages/fold
5 classes	8, 8, 8, 9, 9	121, 162, 131, 140, 165
4 classes	9, 9, 8, 7, 9	175, 169, 105, 129, 141

Table 7: Case-disjoint fold sizes for the classification experiments. Both settings use five-fold `StratifiedGroupKFold` grouped by case identifier.

D Classification task confusion matrices

Gold	A	BF	C	PH	R
Legal-BERT (Macro-F1 = 0.71)					
Analysis	0.74	0.03	0.04	0.02	0.17
BF	0.03	0.85	0.00	0.12	0.00
Conclusion	0.41	0.00	0.41	0.16	0.02
PH	0.02	0.07	0.02	0.89	0.00
Rule	0.22	0.00	0.01	0.00	0.77
Modern-BERT (Macro-F1 = 0.65)					
Analysis	0.73	0.02	0.07	0.05	0.13
BF	0.08	0.80	0.00	0.10	0.02
Conclusion	0.43	0.00	0.41	0.11	0.05
PH	0.11	0.07	0.05	0.77	0.00
Rule	0.30	0.01	0.02	0.01	0.66
GPT-5-mini (Macro-F1 = 0.76)					
Analysis	0.65	0.08	0.03	0.00	0.24
BF	0.08	0.86	0.00	0.03	0.02
Conclusion	0.34	0.00	0.66	0.00	0.00
PH	0.05	0.11	0.04	0.79	0.02
Rule	0.05	0.00	0.00	0.00	0.95

Table 8: Row-normalized confusion matrices for the five-class case-disjoint experiment. Rows are gold labels and columns are predicted labels. A=Analysis, BF=*Background Facts*, C=Conclusion, PH=*Procedural History*, and R=Rule.

E Flat query structure and example

Flat structure

argument
root: conclusion: [CONCLUSION]

context
conclusion: analysis: [ANALYSIS 2]
premise: analysis: [ANALYSIS 3]
premise: missing

conclusion: conclusion: [CONCLUSION]
premise: analysis: [ANALYSIS 2]
premise: rule: [RULE 2]

focus
conclusion: analysis: [ANALYSIS 2]
premise: analysis: [ANALYSIS 3]
premise: [MASK]

Example

argument
root: missing

context
conclusion: analysis: The only beneficial right . . .
premise: rule: I.R.C. § 677(a)(2) provides that the grantor . . .
premise: analysis: In accordance with Ringwalt's powers as trustee, the 1967 liquidating distribution received from dissolution of R & L, Inc. was allocated to principal, held by the trust for future distribution and actually distributed to Ringwalt upon termination of the trust in 1969.

conclusion: analysis: The series of transactions that took place in the instant case appears governed by I.R.C. § 368(a)(1)(D),
premise: analysis: In determining that the continuity of interest . . . adequately supported by the record.
premise: rule: Assessing continuity of interest . . . Bondy v. Commissioner, 269 F.2d 463, 466–67 (4th Cir. 1959);
premise: analysis: The only beneficial right that . . . entitled to total control after ten years.

conclusion: analysis: Under these circumstances, it . . . by section 368, was satisfied.
premise: analysis: The series of transactions that took place in the instant case appears governed by I.R.C. § 368(a)(1)(D),
premise: rule: I.R.C. § 368(a)(1)(D), which defines a reorganization as: . . . under section 354, 355, or 356

conclusion: missing
premise: analysis: Under these circumstances, it appears . . . defined by section 368, was satisfied.
premise: implicit rule

focus
conclusion: [MASK]
premise: analysis: Under these circumstances, . . . section 368, was satisfied.
premise: implicit rule

Figure 4: Side-by-side view of the linearized flat structure and one example.