Mining Legal Arguments in U.S. Corporate Case Law

Anonymous ACL submission

Abstract

Legal argument mining supports passage classification, retrieval, and argument completion. This work introduces an expert-annotated dataset of 42 U.S. federal tax opinions on corporate reorganizations under I.R.C. §368. To our knowledge, it is the first expert-annotated, tree-structured corpus of arguments for this domain. The framework labels spans by legal function: *Rule*, *Analysis*, *Conclusion*, *Background Facts*, and *Procedural History*, and provides span-based, sentence-based, flat, and tree-structured representations. Rather than proposing a new IRAC label inventory, the dataset offers a domain-specific, expert-adjudicated representation of chained support relations in U.S. federal tax reorganization case law. Agreement analysis shows that functional labels are more reliable than directed support edges and implicit intermediate conclusions, so the tree layer should be interpreted more conservatively than the node-label layer. Classification experiments show that functional labels are learnable under case-disjoint evaluation. Retrieval experiments find a limited within-case benefit for structured queries in one candidate-pool setting, but weak cross-case generalization. The dataset supports legal passage classification and provides a conservative benchmark for structured argument mining in U.S. federal tax case law.

1 Introduction

Legal argument mining has produced resources across jurisdictions and tasks, including clause-level mining in European Court of Human Rights (ECHR) decisions (Poudyal et al., 2020; Habernal et al., 2024), argument-scheme annotation in Court of Justice of the European Union (CJEU) fiscal state-aid decisions (Grundler et al., 2022; Santin et al., 2023), functional and rhetorical-role labeling in U.S. judicial opinions (Walker et al., 2017; Savelka and Ashley, 2018; Belfathi et al., 2026), U.S. civil-procedure argument reasoning (Bongard et al., 2022), and U.S. case-law retrieval (Hou et al., 2025). Together, these resources show that legal reasoning can be modeled at several granularities, from clause and sentence roles to premise–conclusion links, legal-reasoning tasks, and retrieval targets. Yet U.S. federal tax opinions addressing corporate reorganizations under I.R.C. §368 remain underrepresented in legal NLP, and existing U.S. resources do not provide expert-adjudicated, document-level support trees for this tax-law setting.

This study examines U.S. federal tax opinions on corporate reorganizations under I.R.C. §368. The corpus includes opinions involving §368(a)(1)(A), (B), (C), (D), and (F), covering statutory mergers and consolidations, stock acquisitions, asset acquisitions, certain asset transfers, and changes in corporate identity, form, or place of organization. Recapitalizations under §368(a)(1)(E) and bankruptcy reorganizations under §368(a)(1)(G) are excluded to limit statutory variety (Office of the Law Revision Counsel, U.S. House of Representatives, 2026). Although narrow, the domain is computationally useful because courts must connect statutory categories, continuity-of-business-enterprise and continuity-of-interest requirements, transaction-specific facts, and intermediate legal conclusions (Electronic Code of Federal Regulations, 2025). Because U.S. courts operate in a precedent-based system, these opinions also support analysis of how statutory rules, case facts, and prior authorities interact in written legal reasoning (Valvoda et al., 2021). The domain therefore provides a testbed for legal NLP systems that model how rules and facts support downstream conclusions, rather than merely labeling isolated sentences.

Our annotation design follows legal-writing accounts of rule-based analysis, in which rules are stated, applied to facts, and developed through intermediate conclusions toward a final disposition

(Gardner and Bartholomew, 2020; Romantz and Vinson, 2020). Spans are labeled as *Rule*, *Analysis*, *Conclusion*, *Background Facts*, or *Procedural History*, and argumentative spans are organized into support trees. Rather than proposing a new IRAC inventory, this approach adapts established legal-reasoning categories to a domain-specific corpus, providing reliable functional-role annotations and an exploratory structural layer for studying chained support relations. A condensed annotation example appears in Appendix A.

Contributions. This paper presents four primary contributions:

1. We release an expert-annotated corpus comprising 42 U.S. federal tax opinions on corporate reorganizations under I.R.C. §368.

2. We provide span-based, sentence-based, flat, and tree-structured representations of functional legal-role annotations.

3. We report inter-annotator agreement for both reliable node labels and substantially less reliable structural annotations, thereby explicitly delineating the reliability limits of the tree layer.

4. We evaluate passage classification and argument-completion retrieval using case-disjoint splits, demonstrating that functional labels are learnable, whereas structure-aware retrieval exhibits limited generalization across cases.

2 Related Work

Argument mining is commonly defined as the automatic identification and extraction of inference and reasoning structures in natural-language arguments (Lawrence and Reed, 2019). In legal NLP, this has led to work on corpora with explicit argument units and relations, rhetorical-role and functional segmentation, legal-reasoning benchmarks, and retrieval resources.

2.1 Legal argument-structure corpora

Prior work has annotated legal argument structures across jurisdictions. Yamada et al. (2019) developed an annotation scheme for Japanese judgments aimed at structure-based summarization. Poudyal et al. (2020) released an ECHR corpus with clause-level premise, conclusion, and non-argument labels, plus relations among argumentative clauses. The Demosthenes project is especially related: Grundler et al. (2022) annotated CJEU fiscal state-aid decisions for argumentative elements, element types, and argument schemes, and Santin et al. (2023) extended this work to argument-structure prediction. Habernal et al. (2024) further expanded ECHR argument mining with a span-level actor–argument-type scheme grounded in legal argumentation theory and ECHR practice.

Other jurisdictions provide additional models. Annotation guidelines for Chinese judicial decisions formalize proposition types and relations such as support, attack, joint, match, and identity (Chen et al., 2026). In North American work, Walker et al. (2017) annotated U.S. veterans' claims decisions with sentence roles and propositional connective types, while Xu et al. (2020); Xu and Ashley (2022) studied Issue–Reason–Conclusion structures and granularity choices for argument mining and summarization. Together, these studies show that legal argument units and relations have been used for related tasks, but not for expert-adjudicated support trees in U.S. federal tax reorganization opinions.

2.2 Rhetorical-role and functional segmentation

A related line of work labels the discourse function of legal text without explicit support relations. Bhattacharya et al. (2019) introduced sentence-level rhetorical-role labeling for Indian Supreme Court judgments, and later Indian legal-document resources expanded this approach with larger corpora and more detailed role classes (Kalamkar et al., 2022; Malik et al., 2022; Nigam et al., 2025). In U.S. case law, Savelka and Ashley (2018) segmented trade-secret and cyber-crime opinions into consecutive, non-overlapping parts using seven functional and issue-specific labels, with a CRF-based system identifying part boundaries rather than arbitrary argumentative spans or support links.

Rhetorical-role labeling has also been extended to new jurisdictions and models. Csányi et al. (2025) developed a sentence-level rhetorical-role classifier for Hungarian judicial decisions, while Bambroo et al. (2025) proposed MARRO, a multi-task, multi-headed-attention model evaluated on Indian and U.K. Supreme Court datasets. Belfathi et al. (2026) introduced SCOTUS-Law, a U.S. Supreme Court corpus annotated at three levels of granularity for rhetorical sentence function. These resources are important precursors for labels such

as facts, rules, analysis, and conclusions, but they generally do not encode document-level support-tree structure.

2.3 IRAC-style legal reasoning benchmarks

IRAC-style reasoning is increasingly common in legal NLP benchmarks and prompting methods. Guha et al. (2023) introduced LegalBench, a broad legal-reasoning benchmark developed with legal professionals. More specialized benchmarks use IRAC or IRAC-aligned structures for scenario analysis, patent decisions, and tax-penalty questions (Kang et al., 2024; Jang et al., 2025; Choi et al., 2026). Related prompting work shows that IRAC-derived prompts can improve legal entailment (Yu et al., 2022), and Chain of Logic uses an IRAC-inspired method for rule-based reasoning (Servantez et al., 2024). Bongard et al. (2022) introduced a U.S. civil-procedure legal-argument reasoning task with case introductions, questions, proposed arguments, and explanatory analyses. This literature supports the use of legal-reasoning categories, while also highlighting the lack of support-tree annotations for U.S. federal tax judicial opinions.

2.4 Legal retrieval

Legal retrieval research studies how systems identify authorities or passages relevant to legal analysis. CLERC is especially relevant: it is a U.S. case-law retrieval and retrieval-augmented generation corpus for retrieving citations that support specific legal analyses and for generating analysis from retrieved authorities (Hou et al., 2025). By contrast, our retrieval task focuses on within-case argument completion and global multi-case passage retrieval using sentences derived from annotated support trees, making it narrower and more directly tied to argument structure.

3 Corpus Creation

We collected 42 U.S. corporate reorganization cases, each between 1k and 10k words, focusing on I.R.C. §368(a)(1)(A), (B), (C), (D), and (F), while excluding §368(a)(1)(E) and (G) to limit statutory variety. The opinions range in citation year from 1935 to 1987; Appendix M reports the distribution by court, year, and main topics.

Two law students with backgrounds in tax and corporate law annotated all 42 documents. Before full annotation, they completed three calibration cases and six hours of guideline training, followed by iterative review and continued consultation of the guidelines and platform. Ten documents were independently double-annotated for inter-annotator agreement, and final annotations were adjudicated by a professor of law with expertise in tax law research and teaching. The annotators and adjudicator are listed as authors. Annotation was conducted in a customized version of Label Studio supporting span labeling, directed links, and tree visualization. Appendix C summarizes corpus scale, label distributions, and tree-level structural properties.

	α_u		F1	
Label	Span	Sentence	Edit	Semantic
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**

Table 1: Inter-annotator agreement for explicit, text-anchored labels on the 10 double-annotated cases. Span-level α_u is computed over annotated text spans, while sentence-level α_u projects span annotations to sentence units. Edit and Semantic denote F1 under edit-distance-based and semantic span pairing.

3.1 Annotation Procedure

Annotators selected free spans expressing atomic units of reasoning, assigned each a functional label, and added directed support links to form argument trees following a syllogistic pattern (Gardner and Bartholomew, 2020). Each linked step represents a local inference supporting a downstream claim. When a premise was implicit, annotators could mark enthymemes and, when needed, insert implicit intermediate conclusions to preserve structural consistency. Appendix B illustrates the workflow.

3.2 Argumentation Scheme

Labels. The scheme uses five labels.

- *Rule* marks generally applicable statements, including legal rules, tests, and abstract criteria.
- *Analysis* marks case-specific reasoning applying rules to facts and often functioning as intermediate conclusions.
- *Conclusion* marks the final outcome of an argument tree.

(a) Implicit Nodes Contingency Counts

Pairing	Label match	Ctx	YY	YN	NY	NN
Edit	Blind	64	2	7	14	41
Semantic	Blind	62	2	7	14	39
Edit	Same label	63	2	7	14	40
Semantic	Same label	64	2	7	15	40

(b) Implicit Nodes Agreement Summary

Pairing	Label match	P_o	P_+	P_-	κ	$R_{A1/A2}$	$E_{A1/A2}$
Edit	Blind	0.67	0.16	0.80	-0.02	0.14 / 0.25	9 / 24
Semantic	Blind	0.66	0.16	0.79	-0.03	0.15 / 0.26	9 / 24
Edit	Same label	0.67	0.16	0.79	-0.03	0.14 / 0.25	9 / 24
Semantic	Same label	0.66	0.15	0.78	-0.04	0.14 / 0.27	9 / 26

Table 2: Agreement on implicit intermediate-conclusion (IC) insertion. Ctx is the number of evaluable comparison contexts after explicit spans were aligned. Pairing methods are edit distance (Edit) and semantic similarity (Semantic). Label match indicates whether explicit nodes could be paired across classes (Blind) or only within the same class (Same label). YY, YN, NY, and NN indicate whether annotator 1 and annotator 2 inserted an implicit IC in the same context. P_o is observed agreement; P_+ and P_- are positive and negative agreement for insertion vs. non-insertion; $R_{A1/A2}$ gives per-annotator insertion rates; $E_{A1/A2}$ gives the number of inserted implicit nodes that could be mapped to an evaluable context for agreement scoring.

- *Background Facts* marks case facts and transaction details that explain the factual setting but do not contribute directly to argumentative reasoning, such as corporate structure, stock ownership, asset transfers, trust arrangements, or business operations.

- *Procedural History* marks litigation posture and procedural events, such as refund claims, deficiency notices, appeals, remands, and the court whose decision is under review.

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

Intermediate implicit conclusions and enthymemes. Enthymemes are arguments that depend on one or more implicit premises (Feng and Hirst, 2011). In our syllogistic scheme, an enthymeme is an abbreviated syllogism in which an unstated *Analysis* or *Rule* must be inferred from shared background knowledge, often through a causal relation. Annotators were instructed to use this label only when strictly necessary and only when the missing component could not be found explicitly.

3.3 Dataset Curation

Adjudication. For the ten double-annotated cases, the adjudicator reviewed complete annotation files, including spans, labels, implicit nodes, and support links, and selected the version that best followed the guidelines. Adjudication therefore produced one final annotation file per case rather than discarding cases with disagreement or merging spans and edges across annotators. Disagreements were used for reliability analysis, while the released dataset reflects the adjudicated case-level files.

Span-to-sentence mapping. We generated a sentence-level version of the dataset for retrieval experiments and as an alternative annotation format. Character-interval overlap linked each sentence to any overlapping span node, with the sentence inheriting the node attributes and the passage text set to the full sentence. When a sentence overlapped multiple annotations, it was linked to all overlapping nodes for positive lookup, but the corpus record retained one original annotation ID and one label. Sentence identity was determined by maximum overlap, with ties resolved by earlier span start, earlier span end, and then lexicographic node ID. Before mapping, sentence-final periods and spaces were removed, and abbreviations, section names, and other non-sentence elements were incorporated into complete sentences.

3.4 Inter-Annotator Agreement

Agreement was evaluated on ten independently double-annotated cases across four layers: text-anchored functional labels, implicit intermediate-conclusion insertion, directed support edges, and directed path reachability.

3.4.1 Node-label Agreement

Because annotators freely chose span boundaries, labels, and graph structure, there was no predefined set of shared spans. We therefore used two complementary label-agreement measures. First, Krippendorff's unitized α_u was computed separately for each label over character offsets (Krippendorff, 1995). Each label was treated as a binary segmentation task, with characters inside spans of that

label marked positive and all others as background; the coincidence matrix was weighted by segment length.

Second, span-level soft-F1 measured sensitivity to boundary variation (Johansson and Moschitti, 2010; Tammewar et al., 2020). Explicit spans were aligned using maximum-weight bipartite matching. The edit-distance condition used the Yujian-Bo edit-distance metric (Li and Liu, 2007), while the semantic condition used cosine similarity between Cohere Embed v4 span embeddings with 512 output dimensions (Cohere Team, 2025). The Hungarian algorithm selected the one-to-one alignment with maximum total similarity (Kuhn, 1955). Soft-F1 complements, but does not replace, α_u: it scores matched span pairs after alignment, whereas α_u measures segmentation agreement over the full character sequence.

Results. Node-label agreement was the strongest annotation layer. In the span view, agreement was high for *Background Facts* ($\alpha_u = 0.879$) and *Conclusion* ($\alpha_u = 0.724$), and moderate for *Rule* ($\alpha_u = 0.613$), *Analysis* ($\alpha_u = 0.567$), and *Procedural History* ($\alpha_u = 0.454$). The sentence view produced nearly identical α_u values, showing that span-to-sentence conversion had little effect on unitized agreement. Soft-F1 was higher in the sentence view for *Analysis*, *Background Facts*, and *Rule*, indicating that some disagreement reflects span-boundary variation rather than label choice alone (Table 1).

3.4.2 Implicit Nodes and Direct Edge Agreement

Implicit-node and direct-edge agreement required pairing explicit nodes across annotators before comparing structural decisions. One-to-one regimes used maximum-weight bipartite matching based on either edit-distance or semantic similarity. Implicit-node insertion agreement is reported only for these one-to-one regimes. Direct-edge and path agreement also use a relaxed offset-overlap regime, in which explicit spans were connected when their overlap covered at least a fixed fraction of the shorter span; connected components of these links formed non-overlapping comparison groups. Both label-blind and same-label pairing are reported to test sensitivity to label disagreement and span-granularity differences.

For structural scoring, platform relations were normalized to a parent-to-child direction. Direct-edge analysis treated each ordered pair of paired units as a binary decision indicating whether a direct support edge existed. The all-pairs context evaluates every ordered source-target pair and includes true negatives, while the edge-union context evaluates only pairs where at least one annotator proposed an edge. Because edge-union contains no true negatives, P_-, $\bar{P}_\pm$, and κ are not reported there; κ should be interpreted only for all-pairs results.

Implicit-node agreement was measured as insertion agreement. For each matched explicit parent that served as a structured parent in at least one annotation, we recorded whether each annotator inserted at least one implicit intermediate conclusion under that parent. This measures agreement on the decision to insert an implicit reasoning step, not agreement on the exact content of the implicit node.

Results. Implicit-node insertion showed high observed agreement but weak positive agreement. Observed agreement was approximately 0.66–0.67, while P_+ was only about 0.15–0.16 and κ was slightly negative (Table 2). Most agreement therefore came from shared non-insertion decisions rather than consistent insertion of implicit intermediate conclusions. Implicit nodes are useful for adjudicated structural representation, but should be treated as a low-reliability layer.

Direct-edge agreement was also weak under strict one-to-one pairing. In the all-pairs setting, one-to-one edit and semantic pairings yielded low positive agreement ($P_+ = 0.06$ to 0.16) and low κ (0.02 to 0.12). Relaxed same-label pairing improved positive agreement ($P_+ = 0.39$) and κ (0.36), suggesting that span granularity explains some structural disagreement. Even so, individual direct edges should not be treated as a high-confidence layer without adjudication (Appendix G).

3.4.3 Path Agreement

Path agreement evaluates whether annotators preserved directed reachability between paired explicit units, even when they chose different local edge decompositions. We report two criteria. The full-transitive criterion counts a target as reachable from a source if any directed path exists between paired units. The stricter implicit-bridge criterion permits traversal through implicit nodes and explicit nodes within the same source group, but stops at a

different paired explicit unit. This captures source-target recovery through implicit mediation without allowing arbitrary explicit nodes to serve as intermediaries.

Results. Path agreement exceeded direct-edge agreement. Full-transitive agreement was strongest under relaxed same-label pairing, with P_+ approximately 0.64 and κ approximately 0.59 (Table 7). Implicit-bridge agreement was lower, with the strongest same-label relaxed condition yielding $P_+ = 0.42$ and $\kappa = 0.37$. These results show that annotators agreed more often on broad support reachability than on precise local edge placement. The path layer is therefore useful as a structural diagnostic, but individual edges and implicit insertions should still be interpreted cautiously (Appendix F).

4 Experiments and Results

4.1 Classification Experiments

Explicit passages are classified into functional roles under two settings: the original five-label scheme and a four-class version that merges *Conclusion* into *Analysis*. We report both because conclusion passages are the final step of the court's analysis and may be difficult for non-contextual classifiers to distinguish. Implicit intermediate conclusions are excluded, leaving 719 explicit passages from 42 cases.

We use five-fold case-disjoint `StratifiedGroupKFold` cross-validation (Stone, 1974; Kohavi, 1995; Roberts et al., 2017), grouping passages by case ID so that no case appears in both training and test folds. Non-GPT models use a linear support-vector classifier (Boser et al., 1992; Cortes and Vapnik, 1995). TF-IDF (Salton and Buckley, 1988) is compared with SBERT (Reimers and Gurevych, 2019), LegalBERT (Chalkidis et al., 2020), and ModernBERT (Warner et al., 2025). We also report GPT-5-mini with high reasoning effort and GPT-5.4 with medium reasoning effort. The GPT runs use full case context and label descriptions in a zero-shot prompt (Appendix N) and serve as context-rich comparisons rather than strict baselines.

Table 3 reports Macro-F1 and per-class F1. Merging *Conclusion* into *Analysis* improves Macro-F1 for all models. Among embedding-based models, LegalBERT performs best, with Macro-F1 scores of 0.71 for five classes and 0.80 for four classes. GPT-5.4 achieves the highest overall Macro-F1, with 0.79 and 0.84, respectively. GPT-5-mini trails GPT-5.4 but exceeds the embedding models in the five-class setting, suggesting that model capacity and contextual reasoning affect performance.

4.2 Retrieval Experiments

Task. We evaluate masked-slot legal argument retrieval: given an incomplete argument, the retriever must rank candidate sentences so that the missing supporting sentence or sentences appear near the top. The task is multi-positive because each query has one masked slot that may correspond to multiple gold sentences. The processed dataset contains 403 training queries, 28 validation queries, and 40 test queries under case-disjoint splits. The test split contains four held-out cases, and each test query has 1–6 gold sentences, averaging 2.15.

Queries are derived from span nodes and directed support links in the adjudicated argument trees. Each terminal *Conclusion* node is treated as a motion root, and each query is built from the subtree under that root. For internal reasoning steps, the masked slot is a contiguous group of premise nodes supporting a conclusion; for root-conclusion queries, the final conclusion is masked. Implicit nodes may appear in the query context but are not retrievable candidates.

Candidate pools. All well-formed sentences in the relevant cases are retrievable candidates, not only sentences overlapping annotated spans. We evaluate three pools: *Same-case*, which ranks same-case sentences after removing gold sentences for other queries; *Same-case full*, which ranks all same-case sentences; and *Global split*, which ranks all 581 sentences from the four held-out test cases. Average pool sizes are 138.8, 157.0, and 581.0.

Query structure. We compare three query views. The *structured* view preserves tree markup, including the argument root, tree context, focused reasoning step, functional-role tags, and one `[MASK]` token. The *flat-masked* view flattens the same content while preserving `[MASK]`, and the *flat-plain* view replaces the mask with a textual missing-span placeholder for BM25 and E5. We compare BM25, E5-base-v2, ModernBERT-base without task fine-tuning, and fine-tuned ModernBERT retrievers using flat and structured queries. BM25 uses Pyserini/Lucene with $k_1 = 0.9$ and $b = 0.4$; E5 uses `intfloat/e5-base-v2` with standard `query:` and `passage:` prefixes.

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	0.79	0.81	0.69	0.85	0.81
GPT-5.4‡	0.79	0.79	0.81	0.67	0.89	0.80	0.84	0.84	0.81	0.90	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 3: Classification experiments with five and four classes (F1-score). Macro Avg denotes macro-averaged F1. Experiments use case-disjoint `StratifiedGroupKFold` cross-validation. Bold values mark the strongest non-GPT score in each column. †GPT-5-mini uses high reasoning effort. ‡GPT-5.4 uses medium reasoning effort. Both GPT rows use label descriptions and additional case context, so they are reported as context-rich comparisons rather than strict baselines. BF=*Background Facts*; PH=*Procedural History*.

Training. The fine-tuned retrievers are ModernBERT dual encoders initialized from `answerdotai/ModernBERT-base`. The query encoder uses the `[MASK]` hidden state, while the passage encoder uses normalized mean pooling over non-padding tokens, excluding the first special token. Both are trained with a multi-positive contrastive objective using log-sum-exp over positives in the numerator and over all candidates in the denominator.

Flat and structured retrievers share the same setup and differ only in query view. Training uses 20 epochs, learning rate 10^{-5}, temperature 0.07, maximum query length 4096, and maximum passage length 500. Each example samples up to four positives, 56 same-case negatives, and four cross-case *Background Facts* distractors; if fewer positives are available, additional same-case negatives fill the batch. Other passages serve as in-batch negatives. This sampling scheme most closely matches the same-case retrieval setting.

Evaluation. We report hit rate@K, MRR@K, mean set recall@K, and exact set match@K for $K \in \{1, 5, 10, 20\}$. Evaluation uses maximum query length 4096, maximum passage length 600, query batch size 64, and passage batch size 256; E5 test queries are left-truncated because of its 512-token limit. Hit rate@K measures whether at least one gold sentence appears in the top K, mean set recall@K measures the fraction of each gold set recovered, exact set match@K requires all gold sentences to appear in the top K, and MRR@K is the reciprocal rank of the first gold sentence. Figure 1 plots hit rate@K, with a random baseline computed as the mean analytic probability that a uniformly random top-K sample contains at least one gold sentence.

Results. Results vary by candidate pool. In the *same-case* regime, BM25 and E5-base-v2 are strongest on hit rate@20, both reaching 57.5%; E5-base-v2 also leads on mean set recall@20, exact set match@20, and MRR@20. The structured fine-tuned retriever improves over ModernBERT-base, reaching 55.0% hit rate@20 and 35.3% mean set recall@20, but does not surpass BM25 or E5-base-v2.

In the *same-case full* regime, the structured fine-tuned retriever performs best on the main multi-positive metrics, with 50.0% hit rate@20, 31.3% mean set recall@20, and 20.0% exact set match@20. E5-base-v2 has the highest MRR@20, indicating earlier first hits, but recovers less of the complete gold set. This is the clearest setting in which tree-structured queries appear useful.

In the *global split* regime, BM25 has the highest hit rate@20 at 45.0%, while E5-base-v2 leads on mean set recall@20, exact set match@20, and MRR@20. The fine-tuned ModernBERT retrievers are much weaker, with 15.0% hit rate@20 for the flat model and 12.5% for the structured model. Overall, structure-aware fine-tuning helps in some within-case settings but does not show robust cross-case generalization; given the 40-query test set, close differences should be interpreted descriptively.

5 Discussion

Explicit functional-role labels are the most reliable part of the resource. Span- and sentence-level agreement are nearly identical, while implicit in-

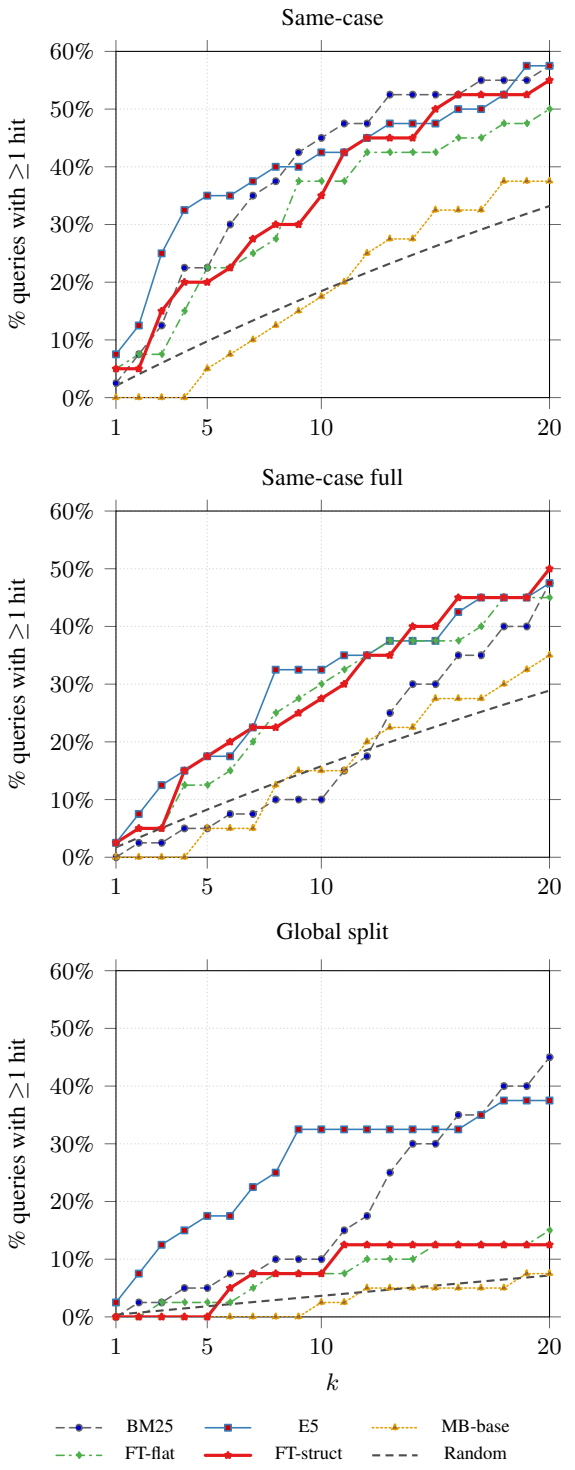

Figure 1: Hit rate@k across the three retrieval candidate pools. The retrieval methods tested are MB-base = ModernBERT-base flat-query retriever; FT-flat = fine-tuned flat-query retriever; FT-struct = fine-tuned structured-query retriever. The Random baseline is the analytic probability that a uniformly random top-k sample from the candidate pool contains at least one gold passage.

termediate conclusions and direct support edges are less stable. Path agreement is stronger than direct-edge agreement, suggesting that annotators often preserve similar broad reachability relations even when they decompose local support differently. The node-label layer should therefore be treated as the primary reliable annotation layer, while the tree layer is best viewed as an exploratory representation of chained support relations.

The classification results show that functional labels are learnable under case-disjoint evaluation, but the inventory remains challenging. Merging *Conclusion* into *Analysis* improves Macro-F1 for all models, and the main remaining confusion is between *Rule* and *Analysis*. The retrieval experiments provide limited evidence for explicit structure: the structured retriever performs best in the *same-case full* setting, but BM25 and E5-base-v2 remain strong in other settings, especially the *global split*. Overall, the dataset is best understood as a resource for legal passage classification and a conservative benchmark for structured argument mining in U.S. federal tax case law.

6 Conclusion

We introduced an expert-annotated dataset of 42 U.S. federal tax opinions on corporate reorganizations under I.R.C. §368. The corpus labels spans by legal function: *Rule*, *Analysis*, *Conclusion*, *Background Facts*, and *Procedural History*, and provides span-based, sentence-based, flat, and tree-structured representations of chained support relations.

The results show that functional labels are learnable under case-disjoint evaluation, while structural annotations are less reliable. Node-label agreement is stronger than agreement on directed support edges and implicit intermediate conclusions, so the tree layer should be interpreted more conservatively than the functional labels. Retrieval results show a limited within-case benefit for structured queries in the same-case full setting, but weak cross-case generalization.

The dataset is best understood as a resource for legal passage classification and a conservative benchmark for structured argument mining in U.S. federal tax case law. It provides a foundation for studying how legal rules, case facts, and intermediate reasoning steps support downstream conclusions in a narrowly defined doctrinal setting.

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

Purbid Bambroo, Subinay Adhikary, Paheli Bhattacharya, Abhijnan Chakraborty, Saptarshi Ghosh, and Kripabandhu Ghosh. 2025. MARRO: Multi-headed attention for rhetorical role labeling in legal documents. *Artificial Intelligence and Law*.

Anas Belfathi, Nicolas Hernandez, Monceaux Laura, Warren Bonnard, Mary Catherine Lavissière, Christine Jacquin, and Richard Dufour. 2026. Coupling local context and global semantic prototypes via a hierarchical architecture for rhetorical roles labeling. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2986–3004, Rabat, Morocco. Association for Computational Linguistics.

Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2019. Identification of rhetorical roles of sentences in indian legal judgments. In Michał Araszkiewicz and Víctor Rodríguez-Doncel, editors, *Legal Knowledge and Information Systems*, volume 322 of *Frontiers in Artificial Intelligence and Applications*, pages 3–12. IOS Press.

Leonard Bongard, Lena Held, and Ivan Habernal. 2022. The legal argument reasoning task in civil procedure. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 194–207, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, Pennsylvania, USA. ACM.

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.

Kun Chen, Xianglei Liao, Kaixue Fei, Yi Xing, and Xinrui Li. 2026. Guidelines for the annotation and visualization of legal argumentation structures in chinese judicial decisions. *Preprint*, arXiv:2603.05171.

Eunkyung Choi, Young Jin Suh, Siun Lee, Hongseok Oh, Juheon Kang, Won Hur, Hun Park, and Wonseok Hwang. 2026. Taxation perspectives from large language models: A case study on additional tax penalties. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4702–4726, Rabat, Morocco. Association for Computational Linguistics.

Cohere Team. 2025. Introducing embed 4: Multimodal search for business. Cohere Blog. Accessed: 2026-04-30.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

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).

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.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. LEGALBENCH: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems 36 - 37th Conference on Neural Information Processing Systems, NeurIPS 2023*, volume 36 of *Advances in Neural Information Processing Systems*. Neural Information Processing Systems Foundation.

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.

Yehoon Jang, Chaewon Lee, Hyun-seok Min, and Sungchul Choi. 2025. PILOT-bench: A benchmark for legal reasoning in the patent domain with IRAC-aligned classification tasks. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 240–280, Suzhou, China. Association for Computational Linguistics.

Richard Johansson and Alessandro Moschitti. 2010. Syntactic and semantic structure for opinion expression detection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 67–76, Uppsala, Sweden. Association for Computational Linguistics.

Prathamesh Kalamkar, Aman Tiwari, Astha Agarwal, Saurabh Karn, Smita Gupta, Vivek Raghavan, and Ashutosh Modi. 2022. Corpus for automatic structuring of legal documents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4420–4429, Marseille, France. European Language Resources Association.

Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Zhuang Li, and Adnan Trakic. 2024. Bridging law and data: Augmenting reasoning via a semi-structured dataset with irac methodology. *ArXiv*, abs/2406.13217.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, volume 2, pages 1137–1143, Montreal, Quebec, Canada. Morgan Kaufmann.

Klaus Krippendorff. 1995. On the reliability of unitizing continuous data. *Sociological Methodology*, 25:47–76.

Harold W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97.

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

Yujian Li and Bo Liu. 2007. A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095.

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.

Shubham Kumar Nigam, Tanmay Dubey, Govind Sharma, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2025. LegalSeg: Unlocking the structure of indian legal judgments through rhetorical role classification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1129–1144, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris

Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.

David S. Romantz and Kathleen Elliott Vinson. 2020. *Legal Analysis: The Fundamental Skill*, 3 edition. Carolina Academic Press, Durham, NC. EISBN: 978-1-5310-1198-7.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

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 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.

Sergio Servantez, Joe Barrow, Kristian Hammond, and Rajiv Jain. 2024. Chain of logic: Rule-based reasoning with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2721–2733, Bangkok, Thailand. Association for Computational Linguistics.

M. Stone. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133.

Aniruddha Tammewar, Alessandra Cervone, Eva-Maria Messner, and Giuseppe Riccardi. 2020. Annotation of emotion carriers in personal narratives. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1517–1525, Marseille, France. European Language Resources Association.

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.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Huihui Xu and Kevin D. Ashley. 2022. Multi-granularity argument mining in legal texts. In Enrico Francesconi, Georg Borges, and Christoph Sorge, editors, *Legal Knowledge and Information Systems*, volume 362 of *Frontiers in Artificial Intelligence and Applications*, pages 261–266. IOS Press.

Huihui Xu, Jaromír Šavelka, and Kevin D. Ashley. 2020. Using argument mining for legal text summarization. In Serena Villata, Jakub Harašta, and Petr Křemen, editors, *Proceedings of the 33rd International Conference on Legal Knowledge and Information Systems (JURIX 2020)*, volume 334 of *Frontiers in Artificial Intelligence and Applications*, pages 184–193. IOS Press, Virtual event.

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.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. Legal prompting: Teaching a language model to think like a lawyer. *Preprint*, arXiv:2212.01326.

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 4: 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*.

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.

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.

A Annotation Structure Example

B Annotation Workflow

C Dataset Statistics

D Implicit intermediate conclusions per annotator label count

E Retrieval expanded results

F Path Agreement

G Direct Edge Agreement

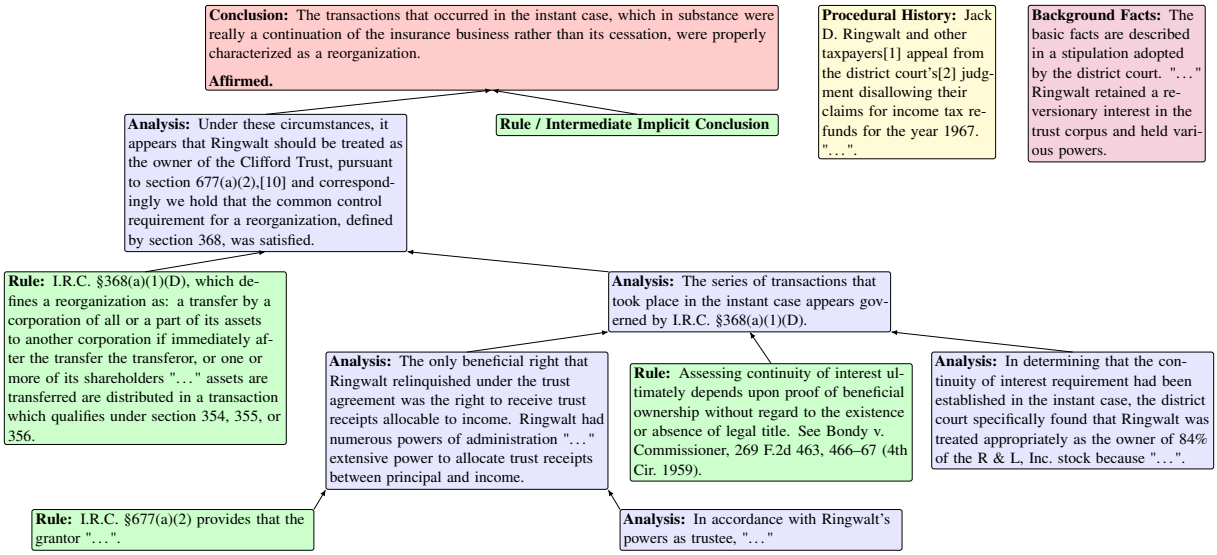

Figure 2: Condensed view of syllogistic argument tree annotation of the case Ringwalt v. U.S., 549 F.2d 89 (8th Cir. 1977). *Background Facts* and *Procedural History* are included in the annotation but are not considered part of the argument structure, as they are defined as contextual spans rather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

Path criterion	Pairing	Label match	P_o	P_+	P_-	$\bar{P}_\pm$	κ
Full transitive	1:1 Edit	Blind	0.85	0.40	0.91	0.66	0.31
Full transitive	1:1 Semantic	Blind	0.86	0.46	0.92	0.69	0.38
Full transitive	Relaxed ≥ 0.50	Blind	0.90	0.63	0.94	0.79	0.57
Full transitive	1:1 Edit	Same label	0.88	0.58	0.93	0.76	0.52
Full transitive	1:1 Semantic	Same label	0.87	0.55	0.93	0.74	0.47
Full transitive	Relaxed ≥ 0.50	Same label	0.91	0.64	0.95	0.80	0.59
Implicit bridge	1:1 Edit	Blind	0.90	0.10	0.95	0.52	0.05
Implicit bridge	1:1 Semantic	Blind	0.91	0.16	0.95	0.55	0.11
Implicit bridge	Relaxed ≥ 0.50	Blind	0.91	0.38	0.95	0.67	0.34
Implicit bridge	1:1 Edit	Same label	0.91	0.20	0.95	0.57	0.15
Implicit bridge	1:1 Semantic	Same label	0.90	0.19	0.95	0.57	0.14
Implicit bridge	Relaxed ≥ 0.50	Same label	0.92	0.42	0.96	0.69	0.37

Table 7: Path agreement over all aligned node pairs. The full-transitive condition treats two aligned nodes as agreeing when each annotator makes the target reachable from the source, allowing paths of any length. The implicit-bridge condition counts only paths recovered through an implicit intermediate node. Path agreement is higher than direct-edge agreement because annotators often preserve reachability even when they choose different local edge decompositions. Relaxed pairing uses a minimum of 50% overlap between passages to be paired.

H Case-disjoint fold sizes

I Classification task confusion matrices

J Structured Query and Example

K Flat query structure and example

L Cases Description

M Related Work Comparison

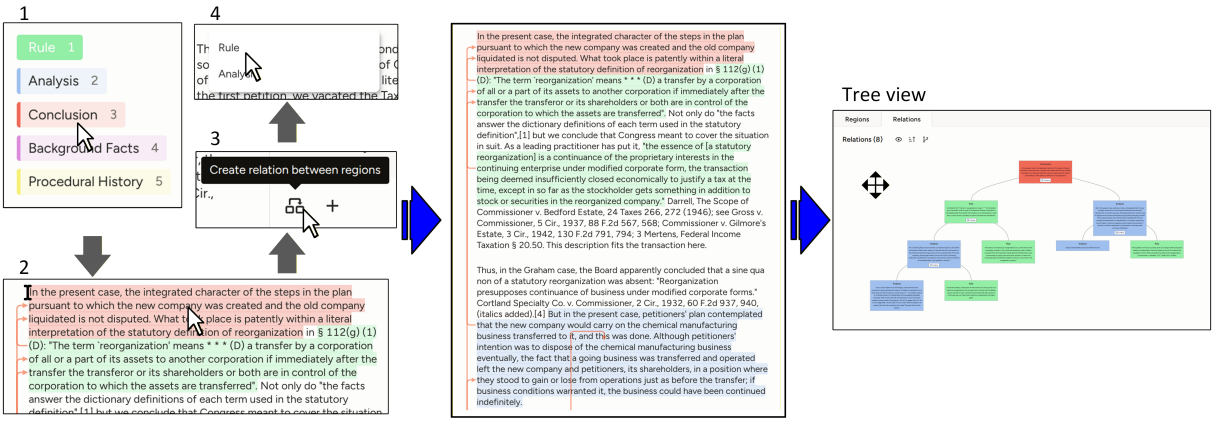

Figure 3: Annotation workflow, visualization of the annotation interface, structure tree visualization, and an example of an annotated case data format. The workflow: (1) selecting the label; (2) highlighting the text, which automatically applies the label's color and saves the annotation; (3) creating 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.

Pairing	Label match	Context type	Contexts	P_o	P_+	P_-	$\bar{P}_\pm$	κ
1:1 Edit	Blind	All pairs	1120	0.91	0.06	0.96	0.51	0.02
1:1 Edit	Blind	Edge union	99	0.03	0.06	–	–	–
1:1 Semantic	Blind	All pairs	1120	0.92	0.10	0.96	0.53	0.06
1:1 Semantic	Blind	Edge union	99	0.05	0.10	–	–	–
Relaxed ≥ 0.50	Blind	All pairs	712	0.92	0.31	0.96	0.63	0.27
Relaxed ≥ 0.50	Blind	Edge union	66	0.18	0.31	–	–	–
1:1 Edit	Same label	All pairs	1054	0.92	0.16	0.96	0.56	0.12
1:1 Edit	Same label	Edge union	94	0.09	0.16	–	–	–
1:1 Semantic	Same label	All pairs	1054	0.92	0.15	0.96	0.55	0.12
1:1 Semantic	Same label	Edge union	96	0.08	0.15	–	–	–
Relaxed ≥ 0.50	Same label	All pairs	576	0.93	0.39	0.97	0.68	0.36
Relaxed ≥ 0.50	Same label	Edge union	50	0.24	0.39	–	–	–

Table 8: Direct-edge agreement on matched explicit-span contexts. *Contexts* is the number of aligned source–target span pairs evaluated. We measure agreement as a binary decision over each context: whether a direct support edge exists from the source span to the target span. Edge-union rows evaluate only source–target pairs where at least one annotator proposed an edge; because $NN = 0$ by construction, P_-, $\bar{P}_\pm$, and κ are not reported for those rows. Relaxed pairing uses a minimum of 50% overlap between passages to be paired.

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

Structured retrieval schema

```
argument
root
  conclusion: [CONCLUSION]
context
node 1
  conclusion: analysis: [ANALYSIS 2]
  premises
    analysis: [ANALYSIS 3]
    missing
node 2
  conclusion: conclusion: [CONCLUSION]
  premises
    analysis: [ANALYSIS 2]
    rule: [RULE 2]
focus
query
  conclusion: analysis: [ANALYSIS 2]
  premises
    analysis: [ANALYSIS 3]
    [MASK]
```

Concrete generated query

```
[ARG]
[ROOT] [MISSING]
[TREE]
[STEP]
  [CONCL] [ANALYSIS] The only beneficial right that Ringwalt relinquished
...
  [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 ...
[/STEP]
[STEP]
  [CONCL] [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 Ringwalt relinquished
...
[/STEP]
[STEP]
  [CONCL] [ANALYSIS] Under these circumstances, it appears that Ringwalt
should be treated as the owner of the Clifford Trust ...
  [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
...
[/STEP]
[STEP]
  [CONCL] [MISSING]
  [PREMISE] [ANALYSIS] Under these circumstances, it appears ... defined by
section 368, was satisfied.
  [PREMISE] [IMPLICIT] [RULE]
[/STEP]
[/TREE]
[FOCUS]
[STEP]
  [CONCL] [MASK]
  [PREMISE] [ANALYSIS] Under these circumstances, it appears ... defined by
section 368, was satisfied.
  [PREMISE] [IMPLICIT] [RULE]
[/STEP]
[/FOCUS]
[/ARG]
```

Figure 4: Side-by-side view of the structured retrieval schema and one concrete generated query. The concrete example uses the actual tags consumed by the structured retriever: [ARG], [ROOT], [TREE], [FOCUS], [STEP], [CONCL], [PREMISE], label tags, [MISSING], [IMPLICIT], and [MASK].

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 10: 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.

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 query

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

Case	Court	Year	Cited statutes	Main topics
Helvering v. Minnesota Tea Co.	U.S.	1935	Revenue Act 1928 §112(i)(1)(A); §203(h)(1)(A)	Asset transfer for stock and cash; substantially all property; continuing proprietary interest versus sale.
Britt v. Commissioner of Internal Revenue	C.C.A. 4	1940	Revenue Act §§203(b)(3), 203(b)(4), 203(h)(1)(A); §§113(a)(6), 113(a)(12)	Basis for redeemed preferred stock; asset transfer for stock and cash; statutory reorganization and nonrecognition exchange.
Le Tulle v. Scofield	U.S.	1940	Revenue Act §112(i); §112(b)(4)	Asset transfer for cash and bonds; creditor status versus proprietary interest; sale or exchange not qualifying as reorganization.
Commissioner of Internal Revenue v. Segall	C.C.A. 6	1940	Revenue Act §§112(b)(4), 112(d)(1), 311(a)(1), 311(f)	Asset transfer and dissolution; sale versus tax-free reorganization; cash and debenture consideration; shareholder transferee liability.
National Rubber Machinery Co. v. U.S.	Ct. Cl.	1941	Revenue Act 1928 §§112(b)(4), 112(d)(1), 113, 114	Depreciation basis for acquired assets; reorganization exchange; cost basis versus transferor basis.
Lyon, Inc., v. Commissioner of Internal Revenue	C.C.A. 6	1942	Revenue Acts/IRC §§112(b)(5), 112(g), 113(a)(6), 113(a)(7), 113(a)(12)	Patent depreciation basis; stepped-up versus carryover basis; temporary corporation and reorganization; business purpose.
Glenn v. Courier-Journal Job Printing Co.	C.C.A. 6 (Ky.)	1942	Revenue Act 1934 §§23, 112, 112(g)	Loss and bad-debt deductions; no stock exchange/control; liquidation rather than reorganization.
Cushman Motor Works v. Commissioner of Internal Revenue	C.C.A. 8	1942	Revenue Act 1934 §§112(b)(3), 112(b)(4), 112(g), 112(g)(1)(C), 112(g)(2)	Depreciation and inventory basis; dissolved corporation asset acquisition; liquidation not reorganization.
Roebling v. Commissioner of Internal Revenue	C.C.A. 3	1944	Revenue Act 1938 §112(g)(1)(A); §112(b)(3); Treasury Regulation 101	Stock-for-bonds statutory merger; continuity of interest; lessor-lessee utility merger; taxable gain recognition.
Westfir Lumber Co. v. C.I.R.	Tax Ct.	1946	§112(g)(1)(B); Revenue Act 1936; Revenue Act 1939 §213(g)	Foreclosure/bondholder reorganization; solely-for-voting-stock test; nonassenting bondholder cash; carryover basis for depreciation and invested capital.
Survaunt v. C.I.R.	C.C.A. 8	1947	IRC §§112(g)(1)(D), 112(g), 112(b)(3), 112(h); Reg. §19.112(g)-1	Recapitalization versus liquidation; integrated transaction; business continuity and business purpose.
Lewis v. C.I.R.	C.C.A. 1	1947	IRC §§112(g)(1)(D), 112(b)(3), 112(c), 115(c), 117; Reg. §112(g)-2	Partial liquidation versus reorganization; business-purpose challenge; dividend versus capital-gain treatment.
Nelson v. U.S.	Ct. Cl.	1947	Revenue Act §§203(b)(3), 203(b)(4), 203(c), 203(e), 203(h)(1)(A)	United Carbon consolidation; basis for later stock sale; stock received under reorganization plan versus taxable transfer.
Lewis v. C.I.R.	C.A.1	1949	IRC §§112(g)(1)(D), 112(c)(1), 112(c)(2), 112(b)(3), 115(c), 117	Old-to-new company transfer of continuing business; boot/liquid-assets distribution; Gregory/business-purpose argument; dividend treatment.
H. Grady Manning Trust v. C.I.R.	Tax Ct.	1950	IRC §§112(g)(1), 112(b)(3), 115(a), 115(g)	Holding company merged into operating company; stock/preferred exchanged for stock and debentures; business-purpose reorganization; debentures not dividend.
U.S. v. Arcade Co.	C.A.6 (Tenn.)	1953	IRC §§112(g), 112(g)(1), 112(g)(2), 112(b)(4); Reg. §19.112(g)-6	Dissolution and transfer through trustees; no corporate party/transfer; reorganization rejected; professional fee deduction.
Forest Hotel Corp. v. Fly	S.D. Miss.	1953	IRC §§112(b)(3), 112(g), 112(g)(1)(B), 113(a)(7), 23(a)(1)	Hotel corporation transaction with Maybar; sale/exchange versus reorganization or merger; leasehold amortization and refund claims.
Becher v. C.I.R.	Tax Ct.	1954	IRC §§112(b)(3), 112(c)(1)–(2), 112(g), 115(c), 115(g), 115(i), 117	War-business shutdown; assets to new furniture corporation; liquidation versus reorganization; cash as partial liquidation/dividend equivalent.
Pebble Springs Distilling Co. v. C.I.R.	C.A.7	1956	IRC §§112(g)(1)(D), 112(b)(3), 112(h), 112(c), 112(e)	Claimed loss on sale of distillery assets; related transferee corporation; control test; D reorganization and loss nonrecognition.
National Bank of Commerce of Norfolk v. U.S.	E.D. Va.	1958	IRC 1939 §113(a)(7); Revenue Act §§202–204, 203(h)(1)	Bank consolidation; goodwill/deposits basis; invested capital for excess profits tax; reorganization basis consequences.
Grede Foundries, Inc. v. U.S.	E.D. Wis.	1962	IRC §§331(a)(1), 361(a), 368(a)(1)(C), 368(a)(1)(A), 311	Taxable gain on Liberty asset acquisition; C reorganization solely-for-voting-stock issue; refund denied.
Grubbs v. C.I.R.	Tax Ct.	1962	IRC §§368(a)(1)(D), 354, 356, 356(a)(2), 316, 302, 355	Ford dealership assets/franchise to new corporation; cash distribution to old shareholders; D reorganization; ordinary income versus capital gain.

Table 11: Cases in the corpus, Part 1. List of cited statutes is non-exhaustive

Case	Court	Year	Cited statutes	Main topics
Mills v. C.I.R.	C.A.5 (Ga.)	1964	IRC §368(a)(1)(B); former §112(g)(1)(B)	Stock-for-stock reorganization; small cash paid for fractional shares; solely-for-voting-stock requirement.
Davant v. C.I.R.	C.A.5 (Tex.)	1966	IRC §§368(a)(1)(D)/(F), 331, 337, 354, 356, 301, 482	Stock sale through intermediary; Warehouse-to-Water asset transfer; liquidation-reincorporation; dividend/boot versus capital-gain treatment.
Sharp v. U.S.	S.D. Tex.	1966	IRC §§331, 354, 356, 368(a)(1)(D), 368(a)(1)(F), 368(c)	Liquidation versus reorganization income character; substantially all assets; water district bond gain issue.
Holliman v. U.S.	S.D. Ala.	1967	IRC §§368(a)(1)(F), 381(a)	F reorganization; same business/assets/shareholders; successor refund and loss carryback.
Stauffer's Estate v. C.I.R.	C.A.9 (Cal.)	1968	IRC §§381, 381(b), 368(a)(1)(A), 368(a)(1)(F), 172(b), 382(b)	Multi-corporation merger into New Mexico entity; F reorganization versus statutory merger; post-merger NOL carryback to premerger years.
King Enterprises, Inc. v. U.S.	Ct. Cl.	1969	IRC §§356(a)(2), 243(a)(1), 334(b)(2), 368(a)(1)(A), 354(a)(1), 302	Tenco stock transfer to Minute Maid plus later merger; unified reorganization versus sale; stock nonrecognition; cash/notes dividend treatment.
Calcote v. U.S.	D.N.J.	1971	28 U.S.C. §1346(a)(1); IRC §§368(a)(1)(B), 354(a)(1), 368(b)	Type B stock-for-stock exchange; Chem-Atom/Aero Chem/Pfaudler triangular structure; continuity and party-to-reorganization issue; refund claim.
Yoc Heating Corp. v. Commissioner of Internal Revenue	U.S. Tax Ct.	1973	IRC §§334(b)(2), 332(b), 368(a)(1)(D)/(F), 362(b), 381(b)(3)	Asset basis step-up versus carryover; stock purchase followed by asset transfer; integrated transaction doctrine; NOL carryback/carryover.
Swanson v. U.S.	C.A.9 (Cal.)	1973	IRC §§368(a)(1)(D), 354(b)(1)(A), 354(a)(1), 331, 337	Old Stockton liquidation and New Stockton formation; D reorganization challenge; no plan of reorganization; substantially-all-assets requirement.
Performance Systems, Inc. v. U.S.	M.D. Tenn.	1973	IRC §§368(a)(1)(F), 381(b), 332, 368(a)(1)(A), 172, 334(b)(2)	Wholly owned subsidiary merged into parent; F reorganization qualification; NOL carryback to subsidiary separate taxable year.
West Side Federal Sav. and Loan Ass'n of Fairview Park v. U.S.	C.A.6 (Ohio)	1974	IRC §368(a)(1)(A); Treas. Reg. §1.368-1(b); IRC §7701; Revenue Act 1926 §203	Savings-and-loan statutory merger; savings accounts as proprietary interests; continuity of interest; tax-free merger status.
Movielab, Inc. v. U.S.	Ct. Cl.	1974	IRC §§381(b), 332, 368(a)(1)(F), 368(a)(1)(A), 381(a), 172, 355, 361	Parent-subsidiary statutory merger; F reorganization and §332 liquidation overlap; NOL carryback to subsidiary premerger year.
Stanton v. U.S.	C.A.3 (Pa.)	1975	IRC §§1361, 331, 356, 368(a)(1)(D), 368(c), 354	Subchapter R termination; liquidation-reincorporation; control after spouse stock issuance; boot/earned surplus.
Aetna Cas. & Sur. Co. v. U.S.	C.A.2 (Conn.)	1976	IRC §§172, 381(b)(3), 368(a)(1)(F), 368(a)(1)(C), 368(a)(1)(B)	F reorganization; post-reorganization NOL carrybacks; shell subsidiary merger; shift in proprietary interests.
Ringwalt v. U.S.	C.A.8 (Neb.)	1977	IRC §§368(a)(1)(D), 331, 337, 354, 677(a)(2)	Liquidation versus reincorporation; insurance agency continuity; Clifford trust ownership/control; reorganization treatment.
Atlas Tool Co., Inc. v. C.I.R.	C.A.3	1980	IRC §§356(a)(2), 368(a)(1)(D), 337, 331, 354, 355, 533, 6901(a)(1)(A)	Fletcher liquidation into Atlas; capital gain versus dividend; D reorganization; accumulated earnings tax; transferee liability.
General Housewares Corp. v. U.S.	C.A.5 (Ala.)	1980	IRC §§337, 368(a)(1)(C), 354, 356, 356(a)(2), 331, 361(a)	C reorganization plus liquidation; corporate §337 treatment; shareholder stock/cash distribution; boot as dividend versus capital gain.
Simon v. C.I.R.	C.A.5	1981	IRC §§331, 356, 368(a)(1)(D), 354, 354(b)(1)	Related-company asset/franchise transfer followed by dissolution; liquidation versus D reorganization; shareholder distribution ordinary income versus capital gain.
Rose v. U.S.	C.A.9 (Wash.)	1981	IRC §§368(a)(1)(D), 337, 331, 354, 355, 356(a)(2), 301, 361	Dissolution of Port Dry Kiln with assets to Exeter; D reorganization despite liquidation form; no tax-avoidance motive requirement.
Russell v. C.I.R.	C.A.6	1987	IRC §§332, 334(b)(2), 368(a)(1), 368(a)(1)(B), 351, 381(b)(3)	Liquidation versus reorganization; carryback denial; continuity/step transaction; tax-free contribution claim.

Table 12: Cases in the corpus, Part 2. List of cited statutes is non-exhaustive

Table 13: Comparison with related work. The table emphasizes how the present corpus differs in domain, annotation schema, structural representation, and evaluation setup.

Work	Domain / jurisdiction	Schema and granularity	Goal / task	Structure layer	How this paper differs
This paper	U.S. federal tax opinions on corporate reorganizations under I.R.C. §368	Free spans plus sentence, flat, and tree views; *Rule*, *Analysis*, *Conclusion*, *Background Facts*, *Procedural History*	Corpus release, functional-role classification, and argument-completion retrieval	Directed support trees with implicit intermediate conclusions; reliability separately reported for labels and structure	Narrow but underrepresented U.S. tax-law domain; expert-adjudicated support trees; explicit reliability limits for structural annotations; case-disjoint classification and retrieval experiments.
Lawrence and Reed (2019)	General argument mining	Survey-level taxonomy, not a legal annotation corpus	Defines and surveys argument mining	Not a corpus-specific structural layer	Provides background framing; this paper operationalizes argument mining in a concrete U.S. tax-law corpus with annotated roles and support trees.
Yamada et al. (2019)	Japanese judgment documents	Legal argumentation corpus for structure-based summarization	Structure-based legal summarization	Argument structure for summarization	Different jurisdiction and downstream goal; this paper targets U.S. federal tax opinions and evaluates both functional-role learning and argument-completion retrieval.
Poudyal et al. (2020)	European Court of Human Rights decisions	Clause-level premise, conclusion, and non-argument labels	Legal argument-mining corpus	Relations among argumentative clauses	ECHR human-rights domain and clause-level argumentative labels; this paper uses functional legal-role spans/sentences in U.S. tax opinions and organizes argumentative spans into support trees.
Grundler et al. (2022)	CJEU fiscal state-aid decisions	Argumentative elements, element types, and argument schemes	Argument detection in EU fiscal state-aid law	Argument-scheme structure	Closest in fiscal/legal subject matter, but EU state-aid rather than U.S. corporate-reorganization tax law; this paper uses a smaller functional-role inventory and support-tree representation.
Santin et al. (2023)	CJEU fiscal state-aid decisions	Argument-structure representation extending the Demosthenes setting	Argumentation-structure prediction	Predicted argument relations / structures	Focuses on relation prediction in CJEU state-aid decisions; this paper contributes a U.S. tax corpus and evaluates within-case and split-global argument-completion retrieval from support trees.
Habernal et al. (2024)	European Court of Human Rights decisions	Span-level actor–argument-type scheme with 16 legally grounded argument types	Mining legal arguments in court decisions	Legal argument types rather than a small IRAC-style role inventory	Rich ECHR-specific argument-type taxonomy; this paper instead uses compact functional legal roles and studies support-tree reliability in U.S. tax case law.
Chen et al. (2026)	Chinese judicial decisions	Proposition types plus relations such as support, attack, joint, match, and identity	Annotation and visualization guidelines	Multi-relation argumentation structures	Broader relation taxonomy and Chinese-law setting; this paper constrains relations to directed support trees for U.S. tax opinions and reports empirical agreement and downstream experiments.
Walker et al. (2017)	U.S. veterans' claims decisions	Sentence roles and propositional connective types	Computational legal reasoning resource	Connective/propositional relations	U.S. legal setting, but different administrative domain and sentence-role/connective focus; this paper targets corporate tax opinions with span-level support trees and sentence-level derived views.
Xu et al. (2020)	Legal texts for summarization	Issue–Reason–Conclusion-oriented argument-mining representation	Legal text summarization	Argument-mining structure used for summarization	Uses argument mining primarily for summarization; this paper releases a domain-specific corpus and tests classification plus argument-completion retrieval.
Xu and Ashley (2022)	Legal texts	Multi-granularity Issue–Reason–Conclusion structures	Argument mining and summarization	Compares granularity choices	Emphasizes granularity for legal AM; this paper provides adjudicated U.S. tax annotations in multiple released views and explicitly separates reliable node labels from lower-confidence edges.

Continued on next page

Work	Domain / jurisdiction	Schema and granularity	Goal / task	Structure layer	How this paper differs
Bhattacharya et al. (2019)	Indian Supreme Court judgments	Sentence-level rhetorical roles	Rhetorical-role identification	No explicit support-link layer	Rhetorical segmentation without argument-support trees; this paper links rules and analyses to conclusions in U.S. federal tax opinions.
Kalamkar et al. (2022)	Indian legal documents	Rhetorical-role / document-structure labels	Automatic legal-document structuring	No explicit support-link layer	Larger rhetorical-role style resource; this paper focuses on functional legal roles plus support-tree structure in a narrow U.S. tax domain.
Malik et al. (2022)	Indian legal documents	Rhetorical-role semantic segmentation	Semantic segmentation of legal documents	No explicit support-link layer	Labels discourse function but not premise-to-conclusion support; this paper annotates both role labels and directed support among argumentative spans.
Nigam et al. (2025)	Indian legal judgments	Rhetorical-role classification for judgment structure	Segmenting legal judgments	No explicit support-link layer	Focuses on rhetorical-role classification at scale; this paper adds explicit support-tree annotations and retrieval tasks tied to missing argumentative supports.
Savelka and Ashley (2018)	U.S. trade-secret and cyber-crime opinions	Consecutive, non-overlapping segments with seven functional and issue-specific labels	CRF-based segmentation into functional parts	No arbitrary span or support-relation annotation	U.S. case-law precedent, but different substantive domains and segmentation design; this paper permits arbitrary spans and directed support trees in corporate tax opinions.
Csányi et al. (2025)	Hungarian judicial decisions	Sentence-level rhetorical-role labels	Rhetorical-role classification	No explicit support-link layer	Different jurisdiction and no structural argument layer; this paper targets U.S. tax-law reasoning with functional roles and tree-structured support.
Bambroo et al. (2025)	Indian and U.K. Supreme Court datasets	Sentence-level rhetorical-role classification	Multi-task model for rhetorical roles	No explicit support-link layer	Primarily a modeling contribution over existing RRL datasets; this paper contributes a new U.S. tax corpus and evaluates whether annotated structure helps retrieval.
Belfathi et al. (2026)	U.S. Supreme Court opinions	Three-level rhetorical-role granularity; sentences labeled by rhetorical function	Hierarchical rhetorical-role labeling	No explicit support-tree layer in the related-work description	U.S. and multi-granular, but broad SCOTUS rhetorical roles; this paper is domain-specific to federal tax reorganizations and includes support-tree annotations.
Guha et al. (2023)	Broad legal-reasoning benchmark	Task-based benchmark built with legal professionals	Measuring legal reasoning in LLMs	Benchmark tasks, not document-level support trees	Broad evaluation suite rather than an annotated case-law corpus; this paper provides case-level annotations and domain-specific retrieval/classification tasks.
Kang et al. (2024)	Legal scenario analysis	Semi-structured IRAC methodology	Augmenting reasoning with IRAC-style data	IRAC-style structure, not opinion-level support trees	Uses IRAC to structure legal reasoning tasks; this paper adapts functional roles to actual U.S. tax opinions and annotates support paths among spans.
Jang et al. (2025)	Patent-law domain	IRAC-aligned classification tasks	Legal-reasoning benchmark	Task/classification structure	Patent benchmark rather than tax case-law corpus; this paper studies corporate-reorganization opinions and support-tree-based argument completion.
Choi et al. (2026)	Tax-penalty questions	IRAC-aligned tax-question setting	Evaluating LLMs on additional tax penalties	Task-level reasoning structure	Shares tax orientation, but focuses on penalty questions; this paper annotates judicial opinions about corporate reorganizations under I.R.C. §368.
Yu et al. (2022)	Legal entailment / prompting	IRAC-derived prompts	Improving model reasoning through prompting	Prompt structure, not annotated document relations	Methodological prompting work; this paper contributes data annotations and evaluates learnability and retrieval from legal-role/support-tree representations.
Servantez et al. (2024)	Rule-based legal reasoning with LLMs	Chain-of-Logic, IRAC-inspired prompting	Prompting method for rule-based reasoning	Prompt-level reasoning chain	Focuses on inference prompting; this paper creates an annotated corpus of real opinions with explicit support links and reliability analysis.

Continued on next page

Work	Domain / jurisdiction	Schema and granularity	Goal / task	Structure layer	How this paper differs
Bongard et al. (2022)	U.S. civil procedure	Case introductions, questions, proposed arguments, and explanatory analyses	Legal argument reasoning task	Task instances rather than document-level support trees	U.S. legal-reasoning task, but civil procedure and constructed task format; this paper annotates full tax opinions and derives retrieval queries from support trees.
Hou et al. (2025)	U.S. case-law retrieval and RAG	Retrieval targets are citations/authorities supporting legal analyses	Retrieve supporting citations and generate legal analysis	Retrieval over authorities, not within-opinion support trees	Closest retrieval-related U.S. case-law resource; this paper instead performs within-case and split-global passage retrieval for missing argument supports derived from annotated trees.

N Classification Experiment Prompt

Prompt N.1: Prompt construction for passage classification

```
if not has_conclusion:
    conclusion_fallback = (
        "IMPORTANT: 'Conclusion' is NOT an available label in this run. "
        "If the passage states the final outcome/disposition, label it as Analysis.\n\n"
    )

prompt = (
    "TASK\n"
    "You will be given (1) case text (context) and (2) one TARGET PASSAGE from that case.\n"
    f"Choose exactly ONE label from: {label_options}\n\n"
    "OUTPUT FORMAT (STRICT)\n"
    "Return exactly ONE line:\n"
    "Class:<label>\n"
    "Do not output anything else.\n\n"
    f"{conclusion_fallback}"
    "GUIDELINES (match the human annotators)\n"
    f"{guidelines}\n\n"
    "<<<CASE_TEXT>>>\n"
    f"{context}\n"
    "<<<END_CASE_TEXT>>>\n\n"
    "<<<TARGET_PASSAGE>>>\n"
    f"{passage}\n"
    "<<<END_TARGET_PASSAGE>>>\n"
)

_SYSTEM_PROMPT = (
    "You are a precise legal passage classifier for U.S. judicial opinions. "
    "Choose exactly ONE label from the allowed set and output ONLY one line in the format "
    "Class:<label>. No extra words, no punctuation, no explanation."
)

_GUIDELINES_CORE = (
    "Annotation scheme (match human annotators): spans are labeled by FUNCTIONAL ROLE in a "
    "chain of syllogisms (polysyllogistic IRAC). Rules and Analyses are argument nodes; "
    "Background Facts and Procedural History are contextual blocks and are not part of the reasoning chain.\n\n"
    "LABEL DEFINITIONS (use these meanings):\n"
    "- Background Facts: narrative context about what happened outside the courtroom (events, transactions, parties, dates). "
    "Includes IRS/agency administrative steps (audits, assessments, refund claims/denials). "
    "These spans inform the reader but do not themselves apply a rule or draw an inference toward the holding.\n"
    "- Procedural History: court/litigation process and posture (complaints, motions, hearings, judgments, appeals, remands, "
    "petitions for certiorari/grants). Focus is the procedural timeline in court. "
    "IRS administrative steps are NOT procedural history.\n"
    "- Rule: a generally applicable premise used to justify an inference statutes, regulations, precedent holdings, tests, "
    "definitions, and other reusable generalizations (including implicit/brute premises that license an inference). "
```

```
    "Summaries of precedent (facts/holdings) used as authority count as Rule. Citations often appear but are not required.\n"
    "- Analysis: case-specific reasoning that applies/interprets a Rule using this case's facts/record; evaluates evidence; "
    "accepts/rejects/distinguishes arguments; draws causal/logical inferences. "
    "IMPORTANT: intermediate/local conclusions in a reasoning chain are labeled Analysis (even if phrased 'we conclude...' "
    "or 'therefore...') when they support a later step.\n\n"
    "TIE-BREAKERS:\n"
    "1) Court procedure => Procedural History. IRS/admin steps => Background Facts.\n"
    "2) Stating a legal standard / definition / precedent holding => Rule. Applying or distinguishing it here => Analysis.\n"
    "3) If both appear, choose the dominant function: 'state the law/test' => Rule; 'apply to facts / infer' => Analysis.\n"
)

_CONCLUSION_GUIDELINE = (
    "CONCLUSION (only when this label exists in the allowed set):\n"
    "- Conclusion: ONLY the terminal outcome of an argument tree / issue, the courts ultimate holding or disposition "
    "(e.g., 'judgment affirmed/reversed', 'summary judgment granted', 'the deduction is allowed/denied'). "
    "Do NOT use Conclusion for intermediate steps that can function as premises for later reasoning; those are Analysis (or Rule).\n"
)
```