Mining Legal Arguments in U.S. Corporate Case Law

Luis Brena[1], William Jurayj[1], Gregory Deyesu[2], Zaid Al-Huneidi[2], Andrew Blair-Stanek[1,2], Benjamin Van Durme[1]

[1]Johns Hopkins University
[2]University of Maryland School of Law
{lbrenap1,wjurayj1,vandurme}@jhu.edu
{gdeyesu,zaidal-huneidi}@umaryland.edu
ablair-stanek@law.umaryland.edu

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 argument corpus for this domain. Explicit spans receive one of five functional labels: *Rule*, *Analysis*, *Conclusion*, *Background Facts*, and *Procedural History*. Rule, Analysis, and Conclusion spans can be linked into directed support trees, while Background Facts and Procedural History serve a contextual function. The corpus provides span-based, sentence-based, flat, and tree-structured representations. Agreement analysis shows that functional node labels are more reliable than directed support edges and implicit intermediate conclusions. Directed-path agreement is stronger than direct-edge agreement, which indicates that broad reachability is more stable than exact local decomposition. Classification experiments show that functional labels are learnable under case-disjoint evaluation. Retrieval experiments show that supervised fine-tuning improves within-case retrieval. However, cross-case generalization remains weak. 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. case law remains underrepresented in legal NLP, and existing resources do not provide expert-adjudicated, document-level support trees for this domain.

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). This narrow 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 allows annotations beyond merely labeling isolated sentences, providing a testbed for legal NLP systems that model how rules and facts support downstream conclusions.

Contributions. This paper presents three primary contributions:

1. We release an expert-annotated corpus comprising 42 U.S. federal tax opinions on corpo-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 35359–35386
October 24-29, 2026 ©2026 Association for Computational Linguistics

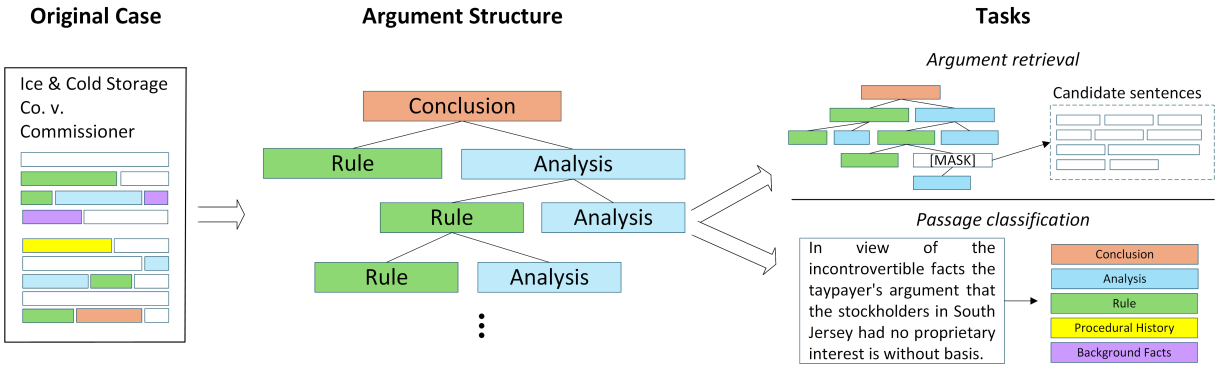

Figure 1: Overview of the annotation method and the associated tasks. In the annotation process, annotators selected spans of text without length constraints and labeled them as *Conclusion*, *Rule*, *Analysis*, *Background Facts*, or *Procedural History*. Annotators connected these spans based on their roles within the argument structure. The annotation enabled two primary tasks: passage classification and masked-slot argument retrieval.

rate reorganizations under I.R.C. §368.[1]

2. We provide span-based, sentence-based, flat, and tree-structured representations of functional legal-role annotations, alongside inter-annotator statistics that highlight reliability limits of the tree layer.

3. 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 (Figure 1).

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 (Appendix O).

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. Jurayj et al. (2026) take a different approach in U.S. Tax Law, using language models instead to parse rules and fact into digestible formalizations that offload argumentation to a symbolic prover (Wielemaker et al., 2012). 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

[1] https://huggingface.co/datasets/lbrenap1/mining-legal-arguments-us-corporate-case-law

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

The Issue, Rule, Analysis, Conclusion (IRAC) reasoning framework is increasingly prevalent 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., 2024, 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.

	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 on explicit functional-node labels for 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 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 N 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 (Tkachenko et al., 2020-2025) supporting span labeling, directed links, and tree visualization. Appendix C summarizes corpus scale, label distributions, and tree-level structural properties.

(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.

3.1 Annotation procedure

Annotators selected free spans expressing atomic units of reasoning, assigned each a functional-node 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.
- *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 when the missing component could not be found explicitly. Appendix A shows a condensed example.

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: explicit functional-node labels, implicit intermediate-conclusion insertion, directed support edges, and directed-path reachability.

3.4.1 Functional-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. Functional-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 directed-edge agreement

Implicit-node and directed-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. Directed-edge and directed-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. Directed-edge analysis treated each ordered pair of paired units as a binary decision indicating whether a directed 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.

Directed-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 directed edges should not be treated as a high-confidence layer without adjudication (Appendix K).

3.4.3 Directed-path agreement

Directed-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 aligned comparison 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. Directed-path agreement exceeded directed-edge agreement. Full-transitive agreement was strongest under relaxed same-label pairing, with P_+ approximately 0.64 and κ approximately 0.59 (Table 11). 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 J).

4 Experiments and Results

4.1 Classification Experiments

Explicit passages are classified into functional roles using two settings: a five-label scheme and a four-class scheme that maps *Conclusion* to *Analysis*. Each instance has one adjudicator-selected explicit span as its target and one functional-role label as its output. We report both settings because conclusion passages represent the final step of the court's analysis and can be difficult for a classifier without case context to distinguish from *Analysis*. Implicit intermediate conclusions are excluded, yielding 718 explicit passages from 42 cases.

Non-GPT classifiers employ five-fold case-disjoint `StratifiedGroupKFold` cross-validation (Stone, 1974; Kohavi, 1995; Roberts et al., 2017). Passages are grouped by case identifier to ensure that no case appears in both training and test data within a fold. The five-class and four-class folds were generated independently. For each setting, predictions from the five held-out folds are concatenated prior to computing Macro-F1 and per-class F1 scores.

TF-IDF and the three dense representations are classified using a linear support-vector classifier (Boser et al., 1992; Cortes and Vapnik, 1995). To account for class imbalance, the classifier assigns higher weights to samples from less frequent classes. A fixed random seed ensures reproducibility, and up to 5,000 optimization iterations are allowed. TF-IDF (Salton and Buckley, 1988) utilizes word unigrams and bigrams, including features present in at least two training passages and excluding those found in more than 90% of passages. The dense representations are SBERT (Reimers and Gurevych, 2019), LegalBERT (Chalkidis et al., 2020), and ModernBERT (Warner et al., 2025). Each dense embedding is scaled to unit Euclidean length before classification.

We also report zero-shot GPT-5-mini with high reasoning effort and GPT-5.4 with medium reasoning effort. Both receive label definitions. The with-context condition also receives the full case text without annotation labels, while the no-context condition receives only the target passage. The exact prompt appears in Appendix P. These runs are context ablations and points of comparison, not strict baselines.

Table 3 presents Macro-F1 and per-class F1 scores. Merging Conclusion into *Analysis* increases Macro-F1 for all models. Among embedding-based models, LegalBERT achieves the highest performance, with Macro-F1 scores of 0.71 for five classes and 0.80 for four classes. With case context, GPT-5.4 achieves the highest overall Macro-F1, with scores of 0.79 and 0.84, respectively; without context, its scores decrease to 0.75 and 0.80. GPT-5-mini's performance declines from 0.76 and 0.79 with context to 0.70 and 0.74 without context. The

Model	5 classes						4 classes				
	Macro Avg	Analysis	BF	Conclusion	PH	Rule	Macro Avg	Analysis	BF	PH	Rule
TF-IDF	0.69	0.75	0.76	0.42	**0.82**	0.69	0.78	0.81	0.80	0.79	0.70
SBERT	0.65	0.74	0.72	0.35	0.69	0.73	0.74	0.81	0.70	0.72	0.74
Legal-BERT	**0.71**	**0.77**	**0.81**	**0.45**	0.79	**0.75**	**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† with context	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‡ with context	0.79	0.79	0.81	0.67	0.89	0.80	0.84	0.84	0.81	0.90	0.80
GPT-5-mini† no context	0.70	0.67	0.58	0.59	0.83	0.82	0.74	0.75	0.55	0.85	0.80
GPT-5.4‡ no context	0.75	0.74	0.70	0.59	0.88	0.81	0.80	0.83	0.68	0.87	0.82
Random	0.17	0.29	0.10	0.09	0.10	0.27	0.21	0.34	0.12	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. Non-GPT classifiers use five-fold case-disjoint `StratifiedGroupKFold` evaluation. 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. The GPT rows are zero-shot evaluations and all use label definitions. The *with context* condition additionally provides the full case text without annotation labels, whereas the *no context* condition provides only the target passage. BF=*Background Facts*; PH=*Procedural History*.

largest context effects are observed for *Background Facts*, while *Rule* exhibits minimal change. TF-IDF remains competitive at 0.69 and 0.78 Macro-F1 and achieves 0.82 F1 on five-class *Procedural History*, reflecting stable lexical cues for that role.

4.2 Retrieval Experiments

Task. We evaluate masked-slot 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 dataset comprises 490 queries and 5,286 passages from 42 cases; each query has between 1 and 10 gold sentences, with an average of 2.41.

We employ five case-disjoint outer folds. In each rotation, one fold serves as the test set, the next as the validation set, and the remaining three as the training set. Each case is evaluated once as held-out test data (Appendix H). Each query masks a single slot in the argument tree. For internal slots, the builder constructs a maximal consecutive same-label block over the ordered direct premises of a conclusion. This block may include both explicit and implicit nodes. All eligible explicit nodes within the block are hidden. Consecutive refers to the order in the direct-premise list, not necessarily to adjacent text in the case. Any sentence with nonzero character overlap with a hidden explicit node is included in the gold set. The same masking and projection mechanism is applied when the hidden node is a terminal *Conclusion*. Implicit nodes may appear in the visible query context but are not considered retrievable candidates.

Candidate pools. All well-formed sentences in the relevant cases are retrievable candidates, not only those overlapping annotated spans. Three pools were evaluated: *Same-case filtered*, which retains the current query's gold sentences but removes same-case passages that are gold only for other queries; *Same-case full*, which ranks all same-case passages; and *Fold-global*, which ranks all passages in the held-out fold. Their mean candidate-pool sizes are 109.4, 134.8, and 1,057.2, respectively.

Query structure. Three query views are employed, although not all are fully crossed with every retriever. The *structured* view preserves tree markup, including the argument root, tree context, focused reasoning step, functional-role tags, and one `[MASK]` token (Appendix L). The *flat-masked* view linearizes the same content while preserving `[MASK]` (Appendix M), whereas the flat-plain view replaces the mask with the textual placeholder `missing span`. BM25 and E5-base-v2 use *flat-plain*. The fixed ModernBERT-base baseline and the fine-tuned flat retriever use

flat-masked, while the fine-tuned structured retriever uses the *structured* view. BM25 uses Pyserini/Lucene with (k_1=0.9) and (b=0.4). E5 uses `intfloat/e5-base-v2` with the standard `query:` and `passage:` prefixes and a deterministic procedure that prioritizes the focused reasoning step when fitting queries within its 512-position limit.

Training. The fine-tuned retrievers are ModernBERT dual encoders initialized from `answerdotai/ModernBERT-base`. The query encoder uses the `[MASK]` hiddenstate, while the passage encoder uses normalized mean pooling over non-padding tokens, excluding the first special token. Both encoders 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 is conducted for 20 epochs with a learning rate of 10^{-5}, temperature 0.07, maximum query length 4096, and maximum passage length 500. Each example uses up to four positives and one of two negative samplers. *Case-focused* training samples 40 unique negatives from the query's case and 20 from other training cases; *pool-uniform* training samples 60 unique negatives passage-uniformly from the training folds. Positive selection remains constant across query views and samplers. Both query views are trained with both samplers and seeds 17, 29, and 43 in every fold. Checkpoints are selected to maximize validation case-macro Recall@20, then untruncated MRR, with the earlier epoch breaking a tie (Appendix F).

Evaluation. We use case-macro Fold-global Hit@20 as the primary evaluation metric. Query-level metrics are first averaged within each held-out case. For fine-tuned systems, the three seed-specific case means are then averaged before weighting all 42 cases equally. Hit@K measures whether at least one gold sentence appears in the top K; Recall@K measures the fraction of each query's gold set recovered; Complete recovery@K requires all gold sentences to appear in the top K; and untruncated MRR uses the rank of the first gold sentence in the complete candidate-pool ranking. Table 7 reports the metrics at $K = 20$, and Figure 2 plots Hit@K at intervals of two. The random baseline is the case-macro analytic probability that a uniformly random top-K sample contains at least one gold sentence. To determine the sensitivity of retriever comparisons to the composition of the 42-case test collection, paired case-resampling intervals are calculated using 10,000 bootstrap samples of the fixed per-case results.

Results. Results vary by candidate pool. In the *Same-case filtered* regime, the fine-tuned flat retriever with case-focused sampling performs best on the main multi-positive metrics, reaching 80.45% Hit@20, 57.17% Recall@20, and 36.04% Complete recovery@20. Both fine-tuned representations improve substantially over ModernBERT-base in this regime.

In the *Same-case full* regime, the fine-tuned flat retriever with case-focused sampling again performs best on the main multi-positive metrics, with 67.98% Hit@20, 43.98% Recall@20, and 26.07% Complete recovery@20. The fine-tuned structured retriever with case-focused sampling has the highest MRR at 18.87%, indicating a slightly earlier first hit, but recovers less of the complete gold set.

In the *Fold-global* regime, BM25 leads all four metrics, with 42.49% Hit@20, 23.99% Recall@20, 12.82% Complete recovery@20, and 7.84% MRR. E5-base-v2 follows with 35.65% Hit@20. The fine-tuned ModernBERT retrievers are weaker: the flat models reach 15.67% with case-focused sampling and 15.58% with pool-uniform sampling. Overall, fine-tuning helps in the within-case settings but does not show robust cross-case generalization, and tree structure does not provide a consistent advantage over the flat representation (Appendix G).

All 95% paired case-bootstrap intervals for the prespecified Fold-global Hit@20 comparisons included zero, providing no reliable evidence of an advantage from structured queries or pool-uniform negative sampling (Appendix I).

5 Discussion

Explicit functional node labels represent the most reliable component of the resource. Span-level and sentence-level agreement are nearly identical, while implicit intermediate conclusions and direct support edges exhibit less stability. Path agreement exceeds direct-edge agreement, indicating that annotators frequently maintain similar broad reachability relations even when decomposing local support differently. Annotator A1 inserted 11 implicit nodes, whereas A2 inserted 30, suggesting that the annotators employed different thresholds for determining when an implicit step was strictly necessary (Appendix D). Consequently, the func-

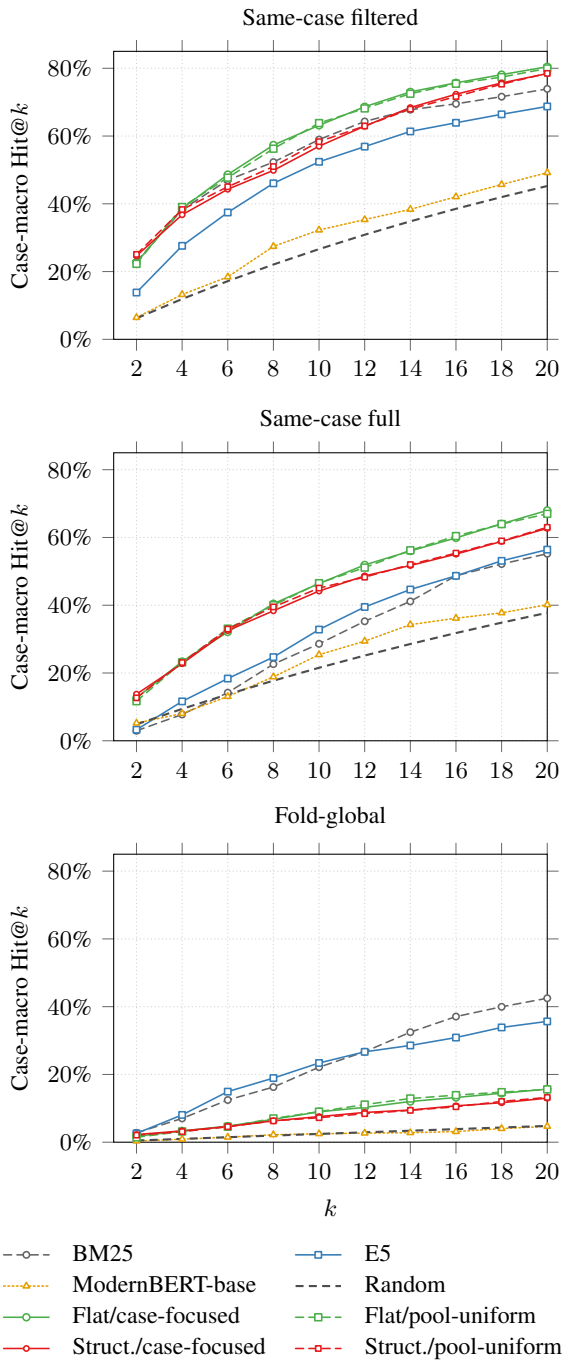

Figure 2: Case-macro Hit@k (%) across three candidate-pool regimes using five-fold case-disjoint evaluation. For each fine-tuned model, the three seed-specific case means are averaged, then averaged equally across all cases. Same-case filtered excludes gold passages for other queries; Same-case full includes all passages from the query's case; Fold-global ranks all passages in the held-out fold. *Case-focused* refers to 40 same-case negatives and 20 from other cases, while the *pool-uniform* curve corresponds to 60 negatives uniformly sampled from the complete training fold.

tional node-label layer should be regarded as the primary reliable annotation layer. The tree layer is best considered an exploratory representation, with directed paths offering a more stable structural benchmark than individual edges.

The classification results show that functional labels are learnable under case-disjoint evaluation, although the five-class functional-role scheme remains challenging. The main confusion occurs between *Rule* and *Analysis*. The four-class setting yields higher Macro-F1 scores across all models. The GPT context experiments demonstrate that context is particularly important for Background Facts, while performance on Rule is largely unaffected by the removal of case context.

The retrieval experiments show a within-case benefit from supervised fine-tuning, but no consistent advantage for structured queries over flat queries. The flat retriever achieves the highest performance on the main set-recovery metrics in both same-case settings, while the structured retriever attains a slightly higher MRR. In the *Fold-global* setting, BM25 performs best, and both fine-tuned representations generalize weakly across cases. The paired case-bootstrap intervals do not provide reliable evidence that structured queries or pool-uniform negative sampling improve Fold-global Hit@20.

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 function: *Rule*, *Analysis*, *Conclusion*, *Background Facts*, and *Procedural History*, and provides span-based, sentence-based, flat, and tree-structured representations of chained support relations.

Under case-disjoint evaluation, functional labels are learned reliably, while structural annotations are less reproducible, suggesting that argumentative reachability is more stable than precise local decomposition. The functional layer should therefore be treated as the corpus's primary representation, with the tree layer serving as an exploratory structural benchmark.

Retrieval fine-tuning improves within-case ranking over ModernBERT-base, but BM25 remains stronger under fold-global evaluation. Together, these results position the dataset as a benchmark for both legal passage classification and the study of how reliably structured legal arguments can be annotated and generalized across cases.

Limitations

The corpus comprises 42 cases from a specialized area of English-language U.S. federal tax law. Due to the high cost of expert annotation, both the corpus size and the quantity of supervised data available for model training are limited. The results may not generalize to other tax doctrines, legal domains, jurisdictions, or non-legal texts.

The opinions span from 1935 to 1987. As the corpus lacks post-1990 opinions, it is not temporally comprehensive and does not reflect recent case language or developments in corporate reorganization practice. Future work should incorporate modern opinions and directly assess temporal transfer.

The reliability of the annotation layers varies. For each case, the adjudicator selected a single complete annotation file without editing or merging individual components. As a result, the released trees represent coherent, expert-selected interpretations, but should not be regarded as the sole valid local decompositions.

The retrieval task employs the selected trees to define hidden slots and gold passages, measuring recoverability under the released representation. Fold-global results show weak cross-case transfer for fine-tuned retrievers, and the tested tree markup does not solve that problem. Close differences between query representations and negative samplers are considered descriptive, as their paired case-bootstrap intervals include zero.

Acknowledgments

ChatGPT and Grammarly were utilized to assist with language refinement and LaTeX formatting. All generated content was reviewed and revised by the authors, who assume full responsibility for the final manuscript.

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, Laura Monceaux, 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. ArXiv v2, 16 June 2026.

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*, volume 36, pages 44123–44279. Curran Associates, Inc.

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 Hou, William Jurayj, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2024. Gaps or hallucinations? scrutinizing machine-generated legal analysis for fine-grained text evaluations. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 280–302.

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 7913–7928, 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.

William Jurayj, Nils Holzenberger, and Benjamin Van Durme. 2026. Language models and logic programs for trustworthy tax reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(45):38688–38698.

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. International Joint Conferences on Artificial Intelligence, Inc. Distributed by Morgan Kaufmann Publishers; ISBN 1-55860-363-8.

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.

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.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

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, Online. Association for Computational Linguistics.

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, Nathan Cooper, Griffin 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.

Jan Wielemaker, Tom Schrijvers, Markus Triska, and Torbjörn Lager. 2012. SWI-Prolog. *Theory and Practice of Logic Programming*, 12(1-2):67–96.

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. ArXiv v2.

A Annotation Structure Example

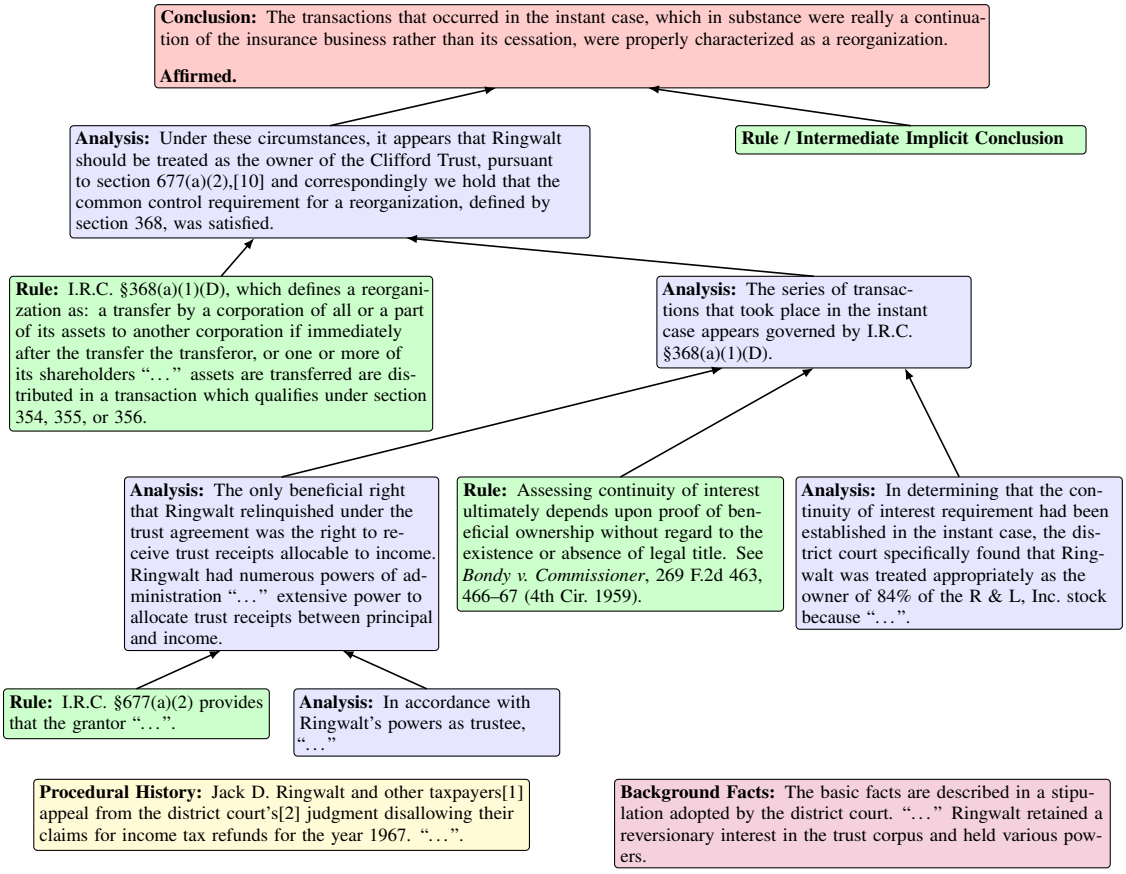

Figure 3: Condensed view of syllogistic argument tree annotation of the case *Ringwalt v. U.S.*, 549 F.2d 89 (8th Cir. 1977). *Background Facts* and *Procedural History* are included in the annotation but are not considered part of the argument structure, as they are defined as contextual spans rather than as support for the argument's claims.

B Annotation Workflow

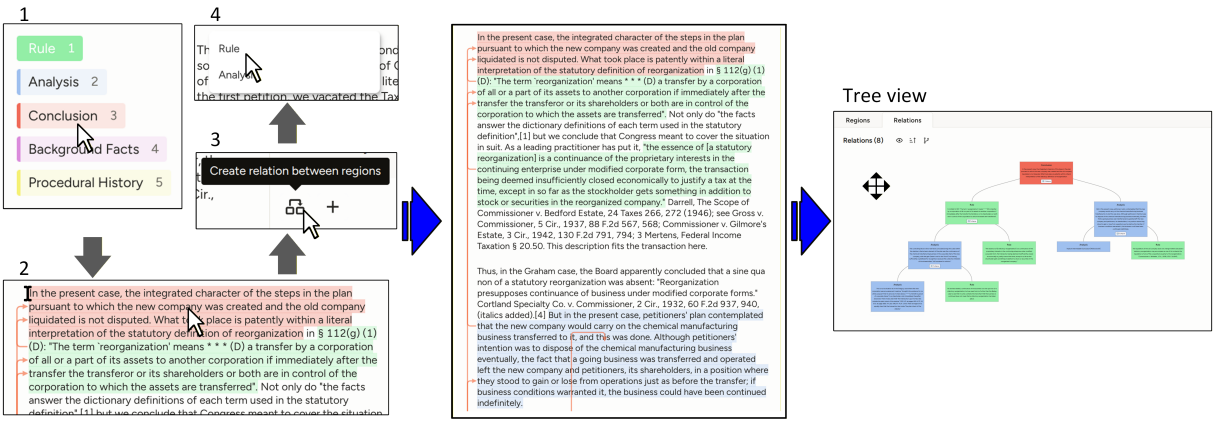

Figure 4: Annotation workflow, visualization of the annotation interface, structure tree visualization, and an example of an annotated case data format. The workflow: (1) selecting the label; (2) highlighting the text, which automatically applies the label's color and saves the annotation; (3) creating 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.

C Dataset Statistics

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	2.41
Max depth	10
Average branching	2.17

	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	7.80	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*.

D Implicit nodes annotation 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 counts. AvgA and AvgR denote average Analysis and Rule implicit nodes insertion.

E Classification task confusion matrices

Gold	A	BF	C	PH	R
Legal-BERT (Macro-F1 = 0.71)					
Analysis	0.75	0.03	0.04	0.02	0.17
BF	0.05	0.83	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	0.21	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 with context (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 6: 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.

F Retrieval Models Training Setup

Dual-encoder scoring. The retriever is initialized from ModernBERT-base and uses one shared encoder for queries and passages, with representation-specific pooling. For a query q, the query representation is the final-layer hidden state at its single `[MASK]` token. For a candidate passage c, the passage representation is obtained by mean pooling the non-padding hidden states after excluding sequence position zero. Both representations are L^2-normalized:

$$\mathbf{z}_q = \frac{g_\theta^q(q)}{\|g_\theta^q(q)\|_2}, \qquad \mathbf{z}_c = \frac{g_\theta^p(c)}{\|g_\theta^p(c)\|_2},$$

where g_θ^q and g_θ^p denote the query- and passage-pooling operations applied to the same encoder with parameters θ. Their similarity is

$$s_\theta(q, c) = \mathbf{z}_q^\top \mathbf{z}_c.$$

Because both vectors are normalized, this dot product is cosine similarity.

Multi-positive contrastive objective. Let $\mathcal{B}_m$ be the set of real queries in distributed microbatch m. For each query $q \in \mathcal{B}_m$, let $\mathcal{A}_q$ be its sampled candidate set, containing between one and four

explicitly selected positive passages and exactly 60 negative passages. The candidate set shared across the distributed microbatch is

$$\mathcal{C}_m = \bigcup_{q \in \mathcal{B}_m} \mathcal{A}_q,$$

where the union is taken over passage identifiers across all workers. Consequently, a passage proposed by multiple queries or workers occurs only once in $\mathcal{C}_m$.

Let $\mathcal{P}_q$ denote the complete set of gold passage identifiers for query q, and let $\tau > 0$ be the temperature. The gold passages that contribute to the loss are those present in the current microbatch candidate set, $\mathcal{P}_q \cap \mathcal{C}_m$. This intersection is guaranteed to be nonempty because every query explicitly contributes at least one positive candidate. The loss for query q is

$$\ell_{q,m} = -\log \frac{\sum_{c \in \mathcal{P}_q \cap \mathcal{C}_m} \exp\left(s_\theta(q,c)/\tau\right)}{\sum_{c \in \mathcal{C}_m} \exp\left(s_\theta(q,c)/\tau\right)}.$$

Training configuration. We did not perform a systematic hyperparameter search. One manually selected configuration was shared across folds, seeds, samplers, and query representations. For matched fold–seed–sampler runs, the flat and structured conditions differed only in their query representation.

- **Initialization:** a shared-weight dual encoder initialized from ModernBERT-base. Query embeddings use the hidden state at the single `[MASK]` token; passage embeddings use mean pooling over attended positions after excluding sequence position zero. Both are L^2-normalized.
- **Optimization:** AdamW with learning rate 10^{-5}, weight decay 0.01, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. We used a linear learning-rate schedule, a warmup ratio of 0.1, and maximum gradient norm 1.0.
- **Schedule:** 20 epochs and three optimizer updates per epoch, for 60 updates in total.
- **Temperature:** $\tau = 0.07$.
- **Positive sampling:** all gold passages are selected when a query has at most four; otherwise, four are selected without replacement. Selection is deterministic given the experiment seed, epoch, query identifier, and passage identifier, and is matched across query views and negative samplers.
- **Case-focused sampler:** 40 unique same-case negatives and 20 unique other-case negatives. The other-case sample is passage-uniform over the eligible passages from all other training cases.
- **Pool-uniform sampler:** 60 unique negatives sampled passage-uniformly from all eligible passages in the training folds, including passages from the query's own case.
- **Sampling exclusions:** both samplers exclude every gold passage of the current query and sample without replacement. Each query therefore proposes 61–64 candidates: one to four selected positives and 60 negatives.
- **Cross-validation and seeds:** five case-disjoint rotations, with three folds used for training, one for validation, and one for testing. Each training rotation contains 294 queries from 24–26 cases. We used seeds 17, 29, and 43, two query representations, and two negative samplers, resulting in $5 \times 3 \times 2 \times 2 = 60$ controlled runs.
- **Checkpoint selection:** all 20 epochs were trained and evaluated. Checkpoints were selected lexicographically by (i) highest validation case-macro set Recall@20, (ii) highest validation case-macro full-ranking reciprocal rank of the first gold passage, and (iii) earliest epoch in the event of a tie.

G Retrieval expanded results

Candidate Pool	Hit rate @20	Recall @20	Complete recovery @20	MRR
Same-case filtered				
BM25	73.92	51.39	32.49	27.72
E5-base-v2	68.74	46.27	29.00	19.93
ModernBERT-base	49.19	29.93	15.66	11.83
Flat (case-focused)	**80.45**	**57.17**	**36.04**	28.98
Flat (pool-uniform)	79.94	57.01	36.03	28.11
Structured (case-focused)	78.43	53.81	33.17	**29.39**
Structured (pool-uniform)	78.50	54.23	33.30	29.27
Same-case full				
BM25	55.21	33.33	18.67	9.35
E5-base-v2	56.41	34.53	20.00	10.78
ModernBERT-base	40.15	22.03	10.36	9.46
Flat (case-focused)	**67.98**	**43.98**	**26.07**	18.68
Flat (pool-uniform)	66.98	43.20	25.34	17.93
Structured (case-focused)	62.70	39.60	22.17	**18.87**
Structured (pool-uniform)	63.00	39.68	22.21	18.47
Fold-global				
BM25	**42.49**	**23.99**	**12.82**	**7.84**
E5-base-v2	35.65	20.04	11.05	7.64
ModernBERT-base	4.68	1.60	0.46	1.34
Flat (case-focused)	15.67	8.25	4.63	3.91
Flat (pool-uniform)	15.58	8.38	4.74	3.84
Structured (case-focused)	13.00	6.64	2.94	4.03
Structured (pool-uniform)	13.22	6.72	2.88	4.00

Table 7: Five-fold case-macro retrieval results are reported as percentages. Queries are averaged within held-out cases. For fine-tuned models, results are further averaged across seeds 17, 29, and 43 within each case before averaging across all 42 cases. Flat and Structured refer to the representations used to fine-tune ModernBERT, while *case-focused* and *pool-uniform* denote the positive and negative sample strategies used during training. Recall@20 represents the mean fraction of each query's gold passages recovered in the top 20. Complete recovery@20 indicates the proportion of queries for which all gold passages are recovered in the top 20. MRR is untruncated and calculated using the rank of the first gold passage in the complete candidate-pool ranking.

H Case-disjoint fold sizes

Evaluation fold	Training cases	Test cases	Test spans
5 classes			
1	34	8	121
2	34	8	161
3	34	8	131
4	33	9	140
5	33	9	165
4 classes			
1	33	9	175
2	33	9	168
3	34	8	105
4	35	7	129
5	33	9	141

Table 8: Case-disjoint fold sizes for the classification experiments. The five- and four-class splits were generated independently using five-fold `StratifiedGroupKFold`. Training and test sets never overlap within any fold.

Evaluation fold	Training cases	Validation cases	Test cases	Test queries
1	24	9	9	98
2	25	8	9	98
3	26	8	8	98
4	26	8	8	98
5	25	9	8	98

Table 9: Case-disjoint fold sizes for the retrieval experiments. In each evaluation fold, one group of cases is used for testing, one for validation, and the remaining three groups are used for training. Each fold has 294 training, 98 validation, and 98 test queries.

I Retrieval Uncertainty Analysis

Negative sampler: pool-uniform − case-focused		
Query representation	Estimate	95% interval
Flat	-0.09	$[-1.47, 1.29]$
Structured	0.23	$[-1.03, 1.62]$
Query representation: Structured − Flat		
Negative sampler	Estimate	95% interval
Case-focused	-2.68	$[-7.11, 1.62]$
Pool-uniform	-2.36	$[-6.08, 1.24]$

Table 10: Paired case-bootstrap contrasts for Fold-global Hit@20 using 10,000 resamples. Values are percentage points; all intervals include zero.

J Path Agreement

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 11: 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.

K Direct Edge Agreement

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 12: 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.

L Structured Query and Example

Structured retrieval schema

[ARG]
[ROOT] [MISSING]
[TREE]

[STEP]
  [CONCL] [ANALYSIS] *analysis span 1*
  [PREMISE] [RULE] *rule span 1*
  [PREMISE] [ANALYSIS] *analysis span 2*
[/STEP]

[STEP]
  [CONCL] [ANALYSIS] *analysis span 3*
  [PREMISE] [ANALYSIS] *analysis span 4*
  [PREMISE] [RULE] *rule span 2*
  [PREMISE] [ANALYSIS] *analysis span 1*
[/STEP]

[STEP]
  [CONCL] [ANALYSIS] *analysis span 5*
  [PREMISE] [ANALYSIS] *analysis span 3*
  [PREMISE] [RULE] *rule span 3*
[/STEP]

[STEP]
  [CONCL] [MISSING]
  [PREMISE] [ANALYSIS] *analysis span 5*
  [PREMISE] [IMPLICIT] [RULE]
[/STEP]
[/TREE]

[FOCUS]
[STEP]
  [CONCL] [MASK]
  [PREMISE] [ANALYSIS] *analysis span 5*
  [PREMISE] [IMPLICIT] [RULE]
[/STEP]
[/FOCUS]
[/ARG]

Condensed example 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 5: Structured retrieval schema and condensed example. Bracketed strings are literal model-input tokens; italicized spans are variables, and . . . marks omitted text. The unique [MASK] is the retrieval slot, whereas [MISSING] marks non-focus occurrences of the withheld node.

M Flat-masked query schema and example

Flat-masked schema

argument
root: missing

context
conclusion: analysis: *analysis span 1*
premise: rule: *rule span 1*
premise: analysis: *analysis span 2*

conclusion: analysis: *analysis span 3*
premise: analysis: *analysis span 4*
premise: rule: *rule span 2*
premise: analysis: *analysis span 1*

conclusion: analysis: *analysis span 5*
premise: analysis: *analysis span 3*
premise: rule: *rule span 3*

conclusion: missing
premise: analysis: *analysis span 5*
premise: implicit rule

focus
conclusion: [MASK]
premise: analysis: *analysis span 5*
premise: implicit rule

Condensed 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 6: Schematic structure and condensed example of the flat-masked query representation. The unique [MASK] marks the retrieval slot; missing marks non-focus occurrences of the withheld node.

N Cases Description

Case	Court	Year	Key provisions	Main topics
Helvering v. Minnesota Tea Co.	U.S.	1935	Revenue Act of 1928 § 112(i)(1)(A)–(B)	Substantially-all-assets transfer for stock and cash; material proprietary continuity; reorganization without dissolution.
Britt v. C.I.R.	4th Cir.	1940	Revenue Act of 1926 § 203(b)(3)–(4), 203(h)(1)(A); Revenue Act of 1932 § 113(a)(6); Revenue Act of 1934 § 113(a)(12)	Substantially-all-assets reorganization; §203(b)(3) nonrecognition; redeemed-stock carryover basis; prior decision not preclusive.
Le Tulle v. Scofield	U.S.	1940	Revenue Act § 112(b)(4), 112(g), 112(i)(1)(A)	Asset transfer for cash and bonds; creditor rather than proprietary status; taxable sale rather than reorganization.
C.I.R. v. Segall	6th Cir.	1940	Revenue Act of 1928 or 1932 § 112(b)(4), 112(d)(1)	Cash-and-debenture asset transfer; taxable sale rather than reorganization; creditor status regardless of debenture term; 1932 transaction timing.
National Rubber Machinery Co. v. U.S.	Ct. Cl.	1941	Revenue Act of 1928 §§ 112(b)(4), 112(d)(1), 113(a)(6)–(8), 114	Depreciation basis after asset acquisition; 80-percent control; binding stock-resale commitments; cost versus carryover basis.
Lyon, Inc. v. C.I.R.	6th Cir.	1942	Revenue Act of 1928 §§ 112(b)(5), 112(i)(1), 113(a)(6)–(7); Revenue Act of 1932 § 113(a)(7); Revenue Act of 1934 § 113(a)(12)	Patent depreciation basis; carryover basis; business-purpose reorganization; separate tax-planning steps.
Glenn v. Courier-Journal Job Printing Co.	6th Cir.	1942	Revenue Act of 1934 §§ 23(f), 23(k), 112, 112(g)	Stock-loss and bad-debt deductions; liquidation rather than reorganization; no continuity of ownership.
Cushman Motor Works v. C.I.R.	8th Cir.	1942	Revenue Act of 1934 § 112(g)(1)(C), 112(g)(1)(E), 112(g)(2), 112(h); Neb. Comp. Stat. §§ 24-107, 24-220	Dissolution and sheriff sale; no intercorporate transfer; ownership shift; no reorganization.
Roebling v. C.I.R.	3d Cir.	1944	Revenue Act of 1938 § 112(b)(3), 112(g)(1)(A); Treas. Reg. 101	Stock-for-bonds statutory merger; continuity of interest; creditor rather than proprietary status; taxable gain.
Westfir Lumber Co. v. C.I.R.	Tax Ct.	1946	Revenue Act of 1936 § 112(g)(1)(B); Revenue Act of 1939 § 213(g)	Substantially-all-assets acquisition for voting stock; nonassenting bondholder cash; integrated transfers; carryover basis.
Survaunt v. C.I.R.	8th Cir.	1947	I.R.C. § 112(g)(1)(D), 112(h); Reg. 103 §§ 19.112(g)-1 and -2	Reorganization rather than liquidation; integrated transaction; continuity and business purpose; carryover basis.
Lewis v. C.I.R.	1st Cir.	1947	I.R.C. §§ 112(b)(3), 112(c)(1)–(2), 112(g)(1)(D), 115(c), 117; Reg. 103 §§ 19.112(g)-1 and -2	Partial liquidation versus reorganization; business-purpose requirement; integrated transaction; remand for missing finding.
Nelson v. U.S.	Ct. Cl.	1947	Revenue Act of 1926 § 203(b)(3)–(4), 203(c), 203(e)(1), 203(h)(1)(A)	Substantially-all-assets reorganization; shareholder stock distribution under plan; later stock-sale basis; delayed liquidation distributions.
Lewis v. C.I.R.	1st Cir.	1949	I.R.C. §§ 112(b)(3), 112(c)(1)–(2), 112(g)(1)(D), 115(c), 117	Type D reorganization; business-purpose doctrine; integrated transferor liquidation; boot and dividend treatment.
H. Grady Manning Trust v. C.I.R.	Tax Ct.	1950	I.R.C. §§ 112(b)(3), 112(g)(1)(A), 112(g)(1)(D), 115(a), 115(g)	Holding-company merger; stock-and-debenture exchange; business-purpose reorganization; debentures not dividends.

Table 13: Cases in the corpus. Key provisions are selected for material relevance.

Case	Court	Year	Key provisions	Main topics
U.S. v. Arcade Co.	6th Cir.	1953	I.R.C. § 112(b)(4), 112(g)(1)(D), 112(g)(2); Reg. 103 § 19.112(g)-6	Dissolution and trustee transfer; no intercorporate transfer; invalid reorganization plan; professional-fee deduction.
Forest Hotel Corp. v. Fly	S.D. Miss.	1953	I.R.C. §§ 23(a)(1), 111, 112(b)(3), 112(g)(1)(B); Reg. 101 art. 112(g)(1); Reg. 45 art. 109	Integrated tax-free merger; continuity and solely-voting-stock test; leasehold amortization; estoppel rejected.
Becher v. C.I.R.	Tax Ct.	1954	I.R.C. §§ 112(b)(3), 112(c)(1)–(2), 112(g)(1)(D), 115(c), 115(g), 115(i), 117	Business-purpose reorganization; tax-free stock exchange; separate partial liquidation; dividend-equivalent cash distribution.
Pebble Springs Distilling Co. v. C.I.R.	7th Cir.	1956	I.R.C. § 112(b)(3)–(4), 112(c), 112(e), 112(g)(1)(D), 112(h)	Controlled asset transfer; 80-percent control; Type D reorganization; claimed-loss nonrecognition.
National Bank of Commerce of Norfolk v. U.S.	E.D. Va.	1958	I.R.C. § 113(a)(7) (1939); Revenue Act of 1926 § 203(b)(3), 203(h)(1); Reg. 130 § 40.458-2(b)	Bank asset purchase; no substantially-all-assets reorganization; deposit goodwill basis; excess-profits invested capital.
Grede Foundries, Inc. v. U.S.	E.D. Wis.	1962	I.R.C. §§ 311(a), 331(a)(1), 361(a), 368(a)(1)(C); Treas. Reg. §§ 1.311-1, 1.368-1 and -2	Integrated Liberty asset transfer and liquidation; Type C solely-voting-stock failure; refund denied.
Grubbs v. C.I.R.	Tax Ct.	1962	I.R.C. §§ 301, 302(b)(2), 316, 354, 356(a)(2), 368(a)(1)(D), 368(c)	Integrated Type D reorganization; stock-plus-cash exchange; dividend-equivalent distribution; control under §368(c).
Mills v. C.I.R.	5th Cir.	1964	I.R.C. § 368(a)(1)(B); former § 112(g)(1)(B)	Type B stock acquisition; fractional-share cash; solely-voting-stock requirement; substance over form.
Davant v. C.I.R.	5th Cir.	1966	I.R.C. §§ 301, 316, 331, 337, 354, 356(a)(2), 368(a)(1)(D), 368(a)(1)(F), 482	Stock sale through conduit; integrated liquidation-reincorporation; Type D and F reorganizations; boot and dividend treatment.
Sharp v. U.S.	S.D. Tex.	1966	I.R.C. §§ 331(a)(1), 356, 368(a)(1)(D), 368(c)	Complete liquidation versus Type D reorganization; plan and substantially-all-assets requirements; capital gain versus dividend; bond-gain dispute.
Holliman v. U.S.	S.D. Ala.	1967	I.R.C. §§ 368(a)(1)(F), 381(a)	Type F reorganization; enterprise continuity; reduced unsecured debt; loss-carryback refund.
Stauffer's Estate v. C.I.R.	9th Cir.	1968	I.R.C. §§ 172(b), 368(a)(1)(F), 381(b)(3), 381(c)(1)(A); Treas. Reg. § 1.381(b)-1(a)(2)	Multi-corporation merger; Type F reorganization; postmerger NOL carryback; allocation to transferor operations.
King Enterprises, Inc. v. U.S.	Ct. Cl.	1969	I.R.C. §§ 243(a)(1), 354(a)(1), 356(a)(1)–(2), 368(a)(1)(A)	Stock acquisition and later merger as Type A reorganization; step transaction; boot as dividend; dividends-received deduction.
Calcote v. U.S.	D.N.J.	1971	I.R.C. §§ 354(a)(1), 368(a)(1)(B), 368(b)	Pre-1964 triangular Type B exchange; integrated stock transfers; party-to-reorganization requirement; continuity of interest.
Yoc Heating Corp. v. C.I.R.	Tax Ct.	1973	I.R.C. §§ 332(b), 334(b)(2), 334(b)(4), 362(b), 368(a)(1)(D), 368(a)(1)(F), 368(c), 381(b)(3)	Integrated stock purchase and asset transfer; Type D and F failures; stepped-up basis; NOL carryback denial.

Table 13: Cases in the corpus (continued).

Case	Court	Year	Key provisions	Main topics
Swanson v. U.S.	9th Cir.	1973	I.R.C. §§ 331, 337, 354(a)(1), 354(b)(1)(A), 368(a)(1)(D)	Old Stockton liquidation and New Stockton formation; no reorganization plan; substantially-all-assets failure; §§331 and 337 treatment.
Performance Systems, Inc. v. U.S.	M.D. Tenn.	1973	I.R.C. §§ 172, 332, 368(a)(1)(A), 368(a)(1)(F), 381(b); Treas. Reg. § 1.381(b)-1(a)(2)	Upstream parent-subsidiary merger; concurrent Type A, Type F, and §332 treatment; §381(b) NOL carryback.
West Side Federal Sav. & Loan Ass'n of Fairview Park v. U.S.	6th Cir.	1974	I.R.C. §§ 368(a)(1)(A), 7701; Treas. Reg. § 1.368-1(b)	Savings-and-loan statutory merger; savings accounts as proprietary interests; continuity of interest; tax-free Type A reorganization.
Movielab, Inc. v. U.S.	Ct. Cl.	1974	I.R.C. §§ 172(b), 332, 361, 368(a)(1)(F), 381(a), 381(b)(3)	Parent-subsidiary statutory merger; concurrent Type F and §332 treatment; postmerger NOL carryback to subsidiary.
Stanton v. U.S.	3d Cir.	1975	I.R.C. §§ 1361(l)–(m), 331, 356, 368(a)(1)(D), 368(c)	Subchapter R termination; liquidation-reincorporation; control despite spouse stock; retained-property boot.
Aetna Cas. & Sur. Co. v. U.S.	2d Cir.	1976	I.R.C. §§ 172, 381(b)(3), 368(a)(1)(F)	Shell-subsidiary merger; Type F reorganization; postreorganization NOL carryback; minority-shareholder redemption.
Ringwalt v. U.S.	8th Cir.	1977	I.R.C. §§ 331, 337, 354–356, 368(a)(1)(D), 677(a)(2)	Liquidation-reincorporation; Type D reorganization; continuity and common control; grantor-trust ownership.
Atlas Tool Co., Inc. v. C.I.R.	3d Cir.	1980	I.R.C. §§ 331(a)(1), 337, 354, 356(a)(2), 368(a)(1)(D), 531–533, 6901(a)(1)(A); Treas. Reg. § 1.368-2(g)	Type D reorganization; boot dividend limit; accumulated-earnings tax; transferee liability.
General Housewares Corp. v. U.S.	5th Cir.	1980	I.R.C. §§ 301(c), 316(a), 331(a)(1), 337, 354, 356(a), 358(a)(1), 361(a), 368(a)(1)(C)	Concurrent Type C reorganization and §337 liquidation; shareholder stock nonrecognition; cash boot as dividend.
Simon v. C.I.R.	5th Cir.	1981	I.R.C. §§ 331, 354, 354(b)(1), 356(a)(2), 368(a)(1)(D)	Integrated asset and franchise transfer and dissolution; Type D reorganization; substantially-all-assets and plan requirements; distribution character.
Rose v. U.S.	9th Cir.	1981	I.R.C. §§ 331, 337, 354(a)(1), 354(b)(1), 356(a)(2), 368(a)(1)(D), 368(c)	Type D reorganization despite liquidation; no tax-avoidance motive required; cash boot as dividend; summary judgment.
Russell v. C.I.R.	6th Cir.	1987	I.R.C. §§ 332, 334(b)(2), 351(a), 354(a)(1), 368(a)(1)(B), 368(a)(1)(F), 368(c), 381(b)(3); Treas. Reg. § 1.368-2(a)	§§332 and 334(b)(2) liquidation; stepped-up basis; step transaction; NOL carryback and §351 denial.

Table 13: Cases in the corpus (continued).

O Related Work Comparison

Work	Domain	Schema and granularity	Goal	Structure layer	How our 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; our 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; our 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 and judgments	Clause-level premise, conclusion, and non-argument labels	Legal argument-mining corpus	Premise–conclusion argument groupings; no typed support or attack labels	ECHR human-rights domain and clause-level argumentative labels; our paper uses functional legal-role spans and 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 labels on legal premises; no explicit support or attack edges	Fiscal-law subject matter, but EU state-aid rather than U.S. corporate-reorganization tax law; our 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	Directed link prediction between known argumentative components	Five typed inferential-link relations among sentence-level components	Focuses on relation prediction in CJEU state-aid decisions; our 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	Flat argument-type and actor labels; no relation graph or explicit support or attack edges	Rich ECHR-specific argument-type taxonomy; our paper instead uses compact functional legal roles and studies support-tree reliability in U.S. tax case law.
Chen et al. (2026) (arXiv preprint)	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; our paper constrains relations to directed support trees for U.S. tax opinions and reports empirical agreement and downstream experiments.

Table 14: Comparison with related work. The table emphasizes differences in domain, annotation schema, structural representation, and evaluation setup.

Work	Domain	Schema and granularity	Goal	Structure layer	How our paper differs
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 and connective focus; our paper targets corporate tax opinions with span-level support trees and sentence-level derived views.
Xu et al. (2020)	Canadian case decisions and expert-written summaries	Sentence-level *Issue*, *Reason*, *Conclusion*, and non-IRC labels	Sentence-role classification toward extractive summarization	No explicit links among Issue–Reason–Conclusion sentences	Uses role classification as a step toward summarization; our paper releases a domain-specific corpus and tests classification plus argument-completion retrieval.
Xu and Ashley (2022)	Canadian case decisions and expert-written summaries	Sentence-level Issue–Reason–Conclusion labels with BIO token labels derived from them	Token- and sentence-level rhetorical-role classification	Compares token and sentence representations; no explicit support or IRC-link layer	Emphasizes granularity for legal AM; our paper provides adjudicated U.S. tax annotations in multiple released views and explicitly separates reliable node labels from lower-confidence edges.
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; our paper links rules and analyses to conclusions in U.S. federal tax opinions.
Kalamkar et al. (2022)	Indian legal documents	Rhetorical-role and document-structure labels	Automatic legal-document structuring	No explicit support-relation layer	Multi-court rhetorical-role resource; our 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; our 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; our 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; our 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; our 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-relation layer	A modeling contribution that also adds 100 newly annotated Indian judgments; our paper contributes a new U.S. tax corpus and evaluates whether annotated structure helps retrieval.

Table 14: Comparison with related work (continued).

Work	Domain	Schema and granularity	Goal	Structure layer	How our paper differs
Belfathi et al. (2026)	U.S. Supreme Court opinions	Three-level rhetorical-role granularity; sentences labeled by rhetorical function	Hierarchical rhetorical-role labeling	Hierarchical label taxonomy and sentence sequence; no explicit support-relation layer	U.S. and multi-granular, but broad U.S. Supreme Court rhetorical roles; our 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; our paper provides case-level annotations and domain-specific retrieval and classification tasks.
Kang et al. (2024)	Malaysian contract-law scenario analysis	Semi-structured Issue–Rule–Application–Conclusion (IRAC) methodology	Augmenting reasoning with IRAC-style data	IRAC-stage annotations plus a semi-structured legal knowledge graph; not opinion-level support trees	Uses IRAC to structure legal reasoning tasks; our paper adapts functional roles to actual U.S. tax opinions and annotates support paths among spans.
Jang et al. (2025)	U.S. PTAB ex parte patent appeals	Case-level Issue, Rule, and Conclusion classification tasks (Application omitted)	Legal-reasoning benchmark	Independent case-level labels; no explicit support-relation layer	Patent benchmark rather than tax case-law corpus; our paper studies corporate-reorganization opinions and support-tree-based argument completion.
Choi et al. (2026)	Korean additional-tax-penalty questions	Case-level facts and claims rendered as binary, rationale-choice, and essay tasks; IRAC is used for analysis and rubrics	Evaluate LLM legal reasoning about penalty-exemption lawfulness	Parallel case-level task views; no document support-relation layer	Shares tax orientation, but focuses on Korean penalty-exemption questions; our paper annotates judicial opinions about corporate reorganizations under I.R.C. § 368.
Yu et al. (2022) (arXiv preprint)	Legal entailment and prompting	IRAC-derived prompts	Improving model reasoning through prompting	Prompt structure, not annotated document relations	Methodological prompting work; our paper contributes data annotations and evaluates learnability and retrieval from legal-role and 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; our paper creates an annotated corpus of real opinions with explicit support links and reliability analysis.
Bongard et al. (2022)	U.S. civil procedure	Topic introductions, civil-procedure fact patterns, answer candidates, and explanatory analyses	Binary answer-correctness classification for legal educational problems	Task instances rather than document-level support trees	U.S. legal-reasoning task, but civil procedure and constructed task format; our paper annotates full tax opinions and derives retrieval queries from support trees.
Hou et al. (2025)	U.S. federal case-law retrieval and retrieval-augmented generation (RAG)	Masked citation contexts linked to cited federal case documents and derived passages	Retrieve cited case authorities and generate citation-conditioned legal analysis	Directed citing-case-to-cited-authority edges; no proposition-level support labels	Retrieval-related U.S. case-law resource; our paper instead performs within-case and split-global passage retrieval for missing argument supports derived from annotated trees.

Table 14: Comparison with related work (continued).

P Classification Experiment Prompt

Passage-classification prompt: system instruction and guidelines (part 1 of 2)

```
# Passed separately as ‘instructions‘ in the Responses API call.
_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."
)

_DEFAULT_ALLOWED_LABELS = (
    "Analysis",
    "Background Facts",
    "Conclusion",
    "Procedural History",
    "Rule",
)

# The four-class setting removes Conclusion from this inventory.
def _get_allowed_labels(combine_analysis_conclusion):
    labels = list(_DEFAULT_ALLOWED_LABELS)
    if combine_analysis_conclusion:
        labels = [label for label in labels if label != "Conclusion"]
    return labels

_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\u2014statutes,
     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\u2019s
     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"
)
```

Passage-classification prompt: renderer and variants (part 2 of 2)

```
_CONCLUSION_GUIDELINE = (
    "CONCLUSION (only when this label exists in the allowed set):\n"
    "- Conclusion: ONLY the terminal outcome of an argument tree / issue\u2014i.e., the court\
    u2019s ultimate holding or disposition "
    "(e.g., 'judgment affirmed/reversed', 'summary judgment granted', 'the deduction is allowed/
    denied'). ")
def _format_label_options(labels):
    if not labels:
        raise ValueError("At least one label must be provided.")
    if len(labels) == 1:
        return labels[0]
    return ", ".join(labels[:-1]) + f", or {labels[-1]}"
def _render_prompt(
    *,
    context,
    passage,
    context_mode,
    allowed_labels,
    include_conclusion_guidance,
):
    guidelines = _GUIDELINES_CORE
    if include_conclusion_guidance:
        guidelines += _CONCLUSION_GUIDELINE
    label_options = _format_label_options(list(allowed_labels))
    has_conclusion = any(lbl.lower() == "conclusion" for lbl in allowed_labels)
    conclusion_fallback = ""
    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")

    task_intro = (
        "TASK\n"
        "You will be given one TARGET PASSAGE from a case.\n"
        "Use the label definitions to pick exactly one class.\n")
    if context_mode == "case_context":
        task_intro = (
            "TASK\n"
            "You will be given (1) case text (context) and (2) one TARGET PASSAGE from that case.\
    n")
    prompt = (
        task_intro
        + 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")
    if context_mode == "case_context":
        prompt += (
            "<<<CASE_TEXT>>>\n"
            f"{context}\n"
            "<<<END_CASE_TEXT>>>\n\n")
    prompt += (
        "<<<TARGET_PASSAGE>>>\n"
        f"{passage}\n"
        "<<<END_TARGET_PASSAGE>>>\n")
    return prompt
```