Mining Legal Arguments in U.S. Corporate Case Law

Anonymous ACL submission

Abstract

The application of argument mining in the legal domain has the potential to influence various real-world tasks, including dense passage retrieval, argument completion, generation, and review. To address these applications, we introduce a simple, syllogism-centric scheme that constructs full arguments by chaining "atomic" syllogisms. In contrast to taxonomies that distinguish between "claim" and "premise," we draw inspiration from the IRAC annotation method by consolidating local (intermediate) *conclusions* into *analysis* or *rule* units, and reserving the "conclusion" label exclusively for the document's final outcome. We formalize this scheme with concise guidelines for marking atomic links and evaluate it by manually annotating 42 U.S. corporate reorganization cases under I.R.C. §368. Schema-based argument mining improves the predictive and retrieval performance of language models. Experimental results demonstrate that our scheme provides strong semantic separation (0.81 Macro-F1 for 4-class classification) and a structure that supports logical completion (60% Recall@20) in a computationally efficient and straightforward setting. The dataset of cases with extracted arguments is released as a resource for future research.

1 Introduction

Legal court decisions contain dense, structured reasoning. Although decisions are written in plain text, the reasoning occurs through a sequence of local inferences, where intermediate conclusions become premises for later steps (Gardner and Bartholomew, 2020). This structure matters for downstream legal NLP, supporting tasks such as finding the passages that justify an outcome, completing missing steps, and retrieving support for a claim (Hou et al., 2025; Šavelka and Ashley, 2016). However, many models still operate at the level of isolated sentences or spans, and lose the proof-like dependencies that make legal analysis coherent.

To address this challenge, we introduce the first syllogism-centric annotation scheme for U.S. federal case law, focusing on reorganization opinions under I.R.C. §368. The scheme labels spans by function (*Rule*, *Analysis*, *Conclusion*, *Background Facts*, *Procedural History*) and links spans to form argument trees. The design is motivated by how legal analysis is taught and drafted: rules are stated, applied to case facts, and iterated until the final disposition (Gardner and Bartholomew, 2020; Romantz and Vinson, 2020). To reduce ambiguity in single-span settings, we additionally surface intermediate conclusions as *Analysis* or *Rule* to yield complete and structurally consistent proof trees for each case. We release expert annotations for 42 decisions, including a double-annotated subset for agreement measurement. We report classification results for predicting functional labels, and retrieval results for argument component infilling, highlighting the versatility of our representation supporting deductive closure in chained reasoning, and highlight potential benefits in automated argument completion.

2 Related Work

Argument mining in legal text has often been framed as identifying premises and conclusions and predicting their relations. We outline the annotation schemes most similar to ours. The Legal Corpus for Argument Mining (Poudyal et al., 2020) annotates premise/conclusion clauses and their links, and provides baselines for recognition and relation prediction. However, their approach treats arguments as one type among many, each consisting of a single conclusion and a set of unstructured premises, and thus directs insufficient attention to the argument as the core payload of a case's text. Also closely related to our approach is Santin et al. (2023), who annotate a corpus of fiscal state aid decision with a richer, hierarchical scheme that distinguishes argumentative elements, their types, and argument schemes. However, their annotation approach treats premises as yielding multiple conclusions, inverting the core syllogistic structure of legal argumentation that our schema extracts.

More broadly, annotation efforts for argument mining have been conducted in various jurisdictions such as the ECHR (Poudyal et al., 2020), CJEU (Grundler et al., 2022), Malaysia (Kang et al., 2024), Canada (Xu et al., 2021), in individual states like Illinois (Blass and Forbus, 2022) and Texas (Chen et al., 2022), or in U.S. federal agency decision making processes (Walker et al., 2017). However, our work is the first to attempt the task of argument mining on United States federal case law, which is especially challenging because of the emphasis on overlapping precedents with differing degrees of compulsion and the system's parallel state and federal court systems.

3 Corpus Creation

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

3.1 Annotation Procedure

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

3.2 Argumentation Scheme

Labels. The scheme uses five labels. *Rule* marks generally applicable statements, including legal rules, tests, and other abstract criteria. *Analysis* marks case-specific reasoning that applies rules to the facts and often captures intermediate conclusions by function. *Conclusion* is reserved for the final outcome of an argument tree. *Background Facts* and *Procedural History* mark contextual spans that provide narrative support for the opinion but do not participate in the reasoning.

Label	α_u	**Soft-F1**
Analysis	0.57	0.46
Background Facts	0.88	0.70
Conclusion	0.72	0.70
Procedural History	0.45	0.47
Rule	0.61	0.51

Table 1: Inter-annotator agreement on 10 double-annotated cases.

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

3.3 Inter-Annotator Agreement

We measured agreement on the 10 double-annotated cases. We compute Krippendorff's unitized α_u separately for each label by treating each label as a binary segmentation task over character offsets and using a length-weighted coincidence matrix that includes background (unlabeled) text. We also report a span-level soft-F1 that matches span strings between annotators with maximum-weight 1–1 matching. Table 1 summarizes the results.

4 Experiments and Results

4.1 Classification Experiments

Passage classification into functional roles was evaluated using two classification experiments. The first experiment employed the five classes defined in the scheme, while the second collapsed the *Conclusion* labels into *Analysis*. Implicit intermediate conclusions were excluded from all experiments. The rationale for collapsing the classes was to assess both the impact of reducing the number of categories and the alignment of revised definitions with

	5 classes						4 classes				
Embedding/LM	Avg	Analysis	BF	Conclusion	PH	Rule	Avg	Analysis	BF	PH	Rule
TF-IDF	0.69	0.75	0.77	0.42	0.82	0.69	0.78	0.81	0.80	0.79	0.70
SBERT	0.65	0.74	0.72	0.37	0.70	0.73	0.74	0.81	0.70	0.72	0.74
Legal-BERT	0.71	**0.77**	**0.82**	0.45	0.79	0.74	**0.81**	**0.83**	**0.81**	0.83	0.75
Modern-BERT	0.65	0.73	0.78	0.39	0.65	0.70	0.71	0.79	0.76	0.60	0.69
GPT-5-mini	**0.76**	0.74	0.71	**0.69**	**0.85**	**0.80**	0.78	0.81	0.68	**0.84**	**0.80**
Random	0.17	0.29	0.13	0.06	0.12	0.24	0.20	0.29	0.15	0.14	0.24

Table 2: Linear SVC results across embeddings and GPT-5-mini classification results (F1-score; Classification experiments with five and four classes; Avg = macro average; BF=*Background Facts*; PH=*Procedural History*).

actual annotations. When considered separately, a *Conclusion* shares the same role and semantic characteristics as an *Analysis*. This similarity may decrease classifier performance and increase confusion between these passages when the labels are distinct.

After filtering for valid annotations, the tests were conducted on 719 valid passages. A Linear SVC classifier (Cortes and Vapnik, 1995) was compared, trained on TF–IDF features and three embedding families (SBERT (Reimers and Gurevych, 2019), LegalBERT (Chalkidis et al., 2020), and ModernBERT (Warner et al., 2025)), alongside a zero-shot GPT-5-mini baseline. Stratified 5-fold cross-validation was applied for each embedding, and for GPT-5-mini, a prompt containing the labels' descriptions and the full case text as context was used.

Table 2 presents macro-F1 scores for all embeddings and the LLM classifier. Reducing the number of classes from five to four by merging final conclusions into analysis consistently improved performance: +0.10 for LegalBERT and +0.02 for GPT-5-mini.

4.2 Dense Passage Retrieval Experiments

Motivation. Argument completion is framed as a passage retrieval task. Given an incomplete argument, the objective is to retrieve the missing supporting text required to complete the syllogistic reasoning. Potential applications include human drafting support and retrieval-augmented generation.

Task construction. Each query is an argument-structured string containing exactly one masked gap token, `[MASK]`. The query encodes the known components of an argument tree and masks one atomic unit. The retriever is required to return the missing span(s) from a pool of candidate passages derived from the same case. Trees are derived from span nodes and directed support links (premise→conclusion) and treated each *Conclusion* node as a separate tree root.

Candidate pool. Each case was divided into well-formed sentences. All case sentences were used as candidates, not only annotated spans. To align span annotations to sentences, each sentence was assigned a single label based on maximum character overlap with any annotated span. Sentences without overlap were treated as unlabeled. For retrieval, Background Facts and Procedural History are treated as unlabeled.

Query linearization and positives. Each case is linearized into one or more argument trees rooted at the *Conclusion*. We use structure tokens to represent the tree, a focused step, and derived conclusions (Figure 1). The query is wrapped in `[ARG]`. We mark the root with `[ROOT]` and encapsulate each tree in `[TREE]`. We represent steps with `[STEP]` and derive conclusions with `[CONCL]`. Finally, we included a `[FOCUS]` block that wraps the target step and contains the single `[MASK]`.

Premises with the same label within a step were grouped into contiguous blocks. We removed one block and inserted `[MASK]`. The positives are the set of candidate sentences that overlap the removed block, and only *Rule*, *Analysis*, and *Conclusion* blocks are eligible positives. Candidate sentences that appear in the query itself were excluded.

Retriever fine-tuning. We fine-tuned a dual-encoder retriever initialized from ModernBERT-base (Warner et al., 2025). We represent the query by the final-layer hidden state at the `[MASK]` position and apply L2 normalization. Each candidate sentence is encoded with the same encoder, and we compute its passage vector by mean pooling over non-padding token embeddings followed by L2 normalization. Candidates are scored by dot-product similarity.

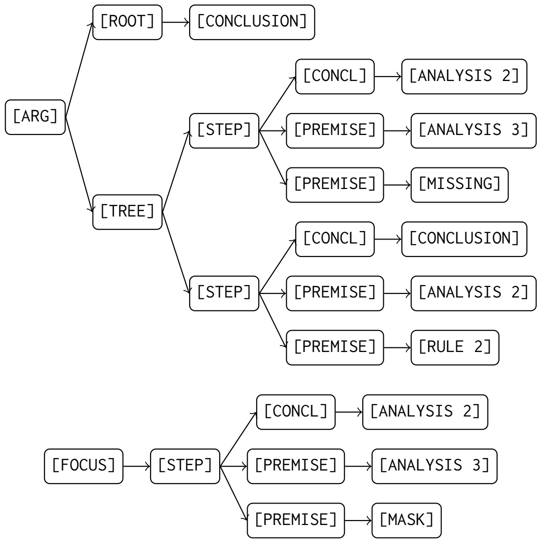

Figure 1: Syllogistic argument tree and its linearized query representation.

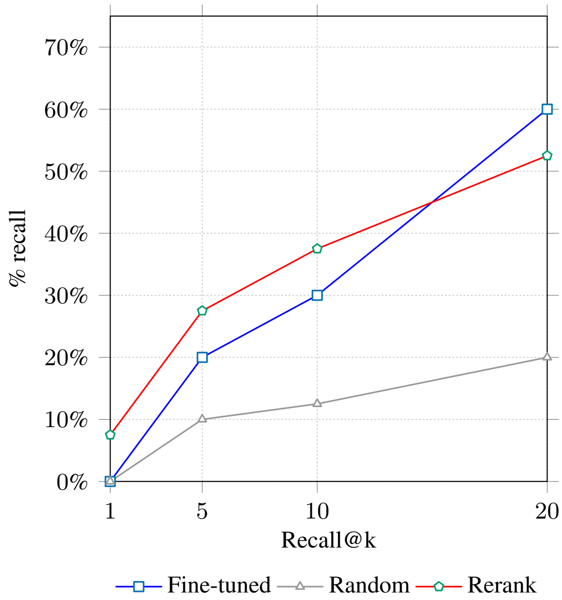

Figure 2: Retrieval vs. Reranking curves.

Training uses a multi-positive contrastive (InfoNCE-style) objective (Zimmermann et al., 2024). Because a masked block can align to multiple sentences, we aggregated all positive candidates with a log-sum-exp term rather than selecting a single positive. In our final setup, each query uses 56 same-case negatives and 4 cross-case negatives (only *Background Facts*). To reduce false negatives, we removed from the negative pool any sentence that is a positive for a different query from the same case. We trained for 20 epochs.

Evaluation and reranking. We evaluated on 40 queries with an average candidate pool of 138.8 sentences. We report Recall@K for $K \in \{1, 5, 10, 20\}$. We reranked the dense retriever's full candidate lists with a cross-encoder reranker (rerank-v4.0-pro) and recompute metrics on the reranked order. Figure 2 shows the recall curves. The fine-tuned retriever reaches 60% Recall@20, compared to 20% for random ranking. Reranking improves Recall@1 (0.0% to 7.5%) and Recall@10 (30% to 37.5%), but reduces Recall@20 (60% to 52.5%). MRR for the fine-tuned model is 11.77% and 15.82% for Rerank.

5 Discussion

Classification results indicate that, both functionally and semantically, a *Conclusion* is equivalent to an *Analysis* that does not introduce an additional premise or is part of another syllogism, as it represents the final decision in the case. We also hypothesize that the language model did not achieve a substantial improvement when moving from five to four classes because of its already strong baseline performance when provided with contextual and task-specific information. In contrast, the embeddings demonstrated greater benefit from a clearer semantic distinction.

Experimental retrieval results indicate that the proposed scheme and dataset are effective for fine-tuning and retrieval tasks. The masked model encodes both structural and semantic information from passages into improved vector representations for the intended application. While the dataset size presents a limitation, the scheme and related fine-tuning and annotation approaches may facilitate applications including argument completion, rule retrieval for document drafting, and the analysis of logical and deductive validity.

6 Conclusion

We introduce a novel annotation scheme for legal argument mining that demonstrates recursion, deductive closure, and structural logic mapping properties. This scheme improves upon previous approaches by incorporating argument-theoretic notions of logical syllogisms. We demonstrate this argument extraction schema on the first dataset of argument mining annotations on United States federal decisions, showing high annotator agreement across a diverse set of span labels. The results indicate significant potential for automated analysis of arguments in unstructured case texts using language models and retrieval systems.

Limitations

Our approach has several limitations. The complexity of these cases' subject matters makes their large-scale annotation prohibitively expensive, constraining our dataset to only 42 high-quality annotated examples. Furthermore, we focus primarily on a narrow subset of United States federal case law, which is specific to the domain of corporate reorganizations and exclusively written in English. Although we believe that our argument extraction scheme should apply to any persuasive document in any language, we recognize the limitation of this evaluation procedure.

References

Joseph Blass and Kenneth Forbus. 2022. The Illinois Intentional Tort Qualitative Dataset. In Enrico Francesconi, Georg Borges, and Christoph Sorge, editors, *Frontiers in Artificial Intelligence and Applications*. IOS Press.

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.

Haihua Chen, Lavinia F. Pieptea, and Junhua Ding. 2022. Construction and Evaluation of a High-Quality Corpus for Legal Intelligence Using Semi-automated Approaches. *IEEE Transactions on Reliability*, 71(2):657–673. Conference Name: IEEE Transactions on Reliability.

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

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.

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

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.

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. *Preprint*, arXiv:1908.10084.

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

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

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

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 edition of the International Conference on Articial Intelligence and Law*, pages 217–226, London United Kingdom. ACM.

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

Huihui Xu, Jaromir Savelka, and Kevin D. Ashley. 2021. Toward summarizing case decisions via extracting argument issues, reasons, and conclusions. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, pages

250–254, New York, NY, USA. Association for Computing Machinery.

Roland S. Zimmermann, Evgenia Rusak, Wieland Brendel, Attila Juhos, Patrik Reizinger, and Oliver Bringmann. 2024. InfoNCE: Identifying the gap between theory and practice. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*.

A Appendix

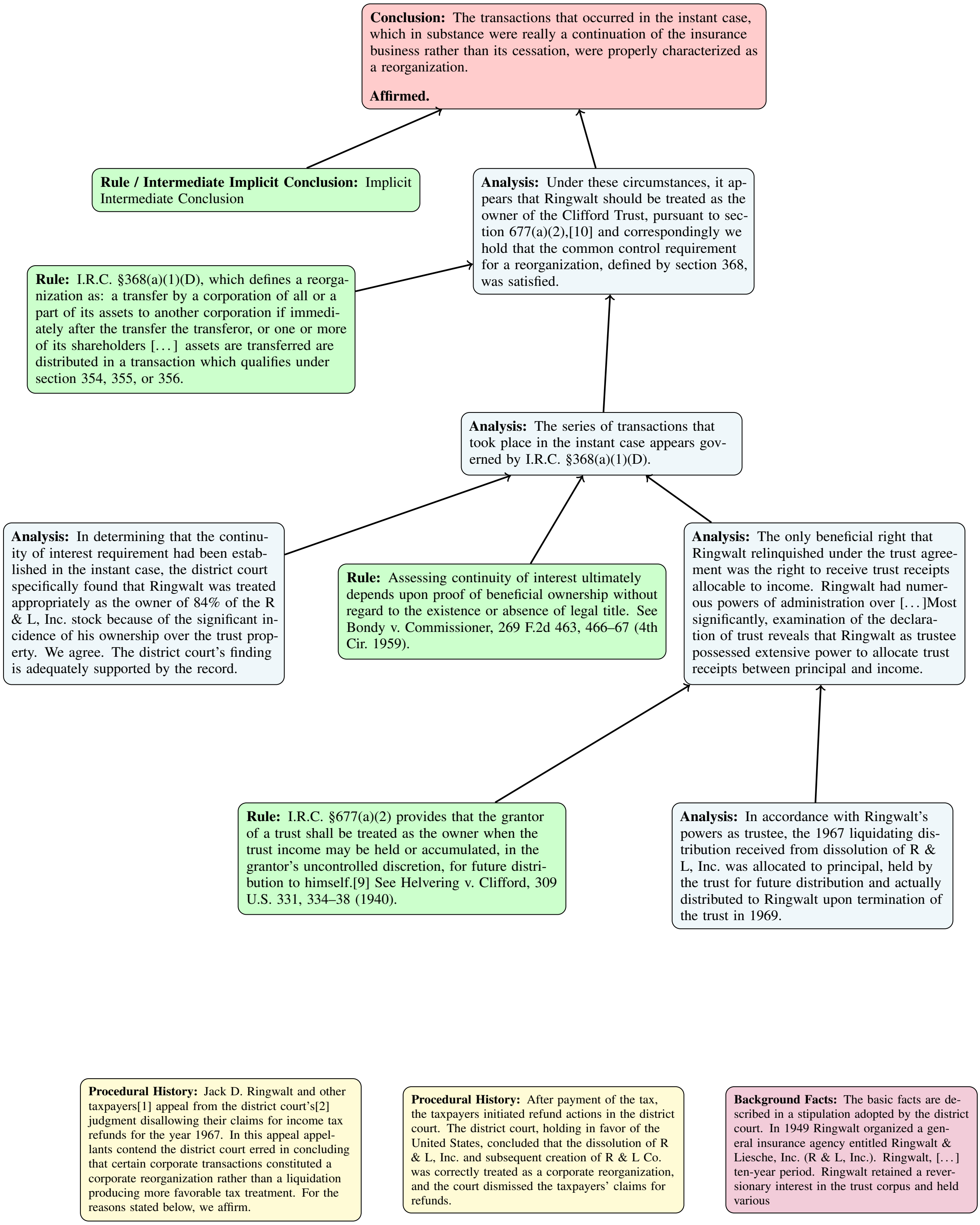

Figure 3: Syllogistic argument tree annotation of the case Ringwalt v. U.S., C.A.8 (Neb.) 1977, 549 F.2d 89