Chain-of-Syllogisms: Unifying Analysis & Conclusions Boosts Argument Mining

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

[1]*Johns Hopkins University, Baltimore, MD, USA*
[2]*University of Maryland School of Law, Baltimore, MD, USA*

Abstract

We propose a simple, syllogism-centric scheme for argument mining that builds full arguments by chaining "atomic" syllogisms. Unlike taxonomies that distinguish "claim" vs. "premise," we collapse local (intermediate) conclusions into *analysis* units and reserve the "conclusion" label only for the document's final outcome. This choice aligns with polysyllogistic reasoning and reduces label ambiguity. We formalize the scheme with concise guidelines for marking atomic links and test it on U.S. corporate reorganization cases under I.R.C. §368. In passage classification with a Linear SVC over several embeddings (TF–IDF, SBERT, Legal-BERT, ModernBERT) and an LLM classifier (GPT-5-mini), collapsing intermediate conclusions into *analysis* (4-class variant) consistently improves macro-F1 by 7–15 points over the 5-class setup across embeddings (Table 1).

Keywords

Legal argumentation, argument mining

Motivation and scheme. Recent work has pushed legal argument mining beyond sentence-level tags [1, 2] toward structured, logic-aware predictions [3, 4, 5]. Our goal is to bridge single-passage classification and functional role labeling by *explicitly modeling polysyllogisms*: each atomic syllogism yields a local conclusion that immediately serves as a premise one level up, treating the latent argument as a proof tree with the overall conclusion at its root [6]. To reduce ambiguity during supervised learning (especially when passages are classified in isolation), we treat those local conclusions as *analyses* or *rules* and keep *conclusion* labels only for the document's final disposition. This collapses intermediate conclusions into their functional analytic role and matches how chained inference is actually used. This mapping aligns with widely taught drafting heuristics such as IRAC/CRAC [7].

Data and annotation. We collected 40 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. This report includes human expert annotations on 26 documents, containing a total of 333 valid classified passages. We define a passage as a span of text that does not necessarily align with sentence boundaries, which can be subjective and detached from logical units. This ensures that each passage is annotated according to its distinct role within the argument. The annotator, a second-year law student, worked independently using a modified version of the annotation software Label Studio. Their task was to select spans of text consisting of atomic claims, and connect them according to a syllogistic grammar drawn from Gardner and Bartholomew [8]. They were instructed to revise their annotations until they conformed to proper grammar and considered valid. No specific corrections or references to the passages were given.

Passages are labeled as **analysis**, **rule**, **conclusion** (final only), **background facts** (BF), and **procedural history** (PH). Links connect premises (rules/analyses) to their immediate conclusions, forming chains. Several similar argument chain approaches have been proposed for annotating cases and tackling argument mining tasks [9, 1, 10, 11]. A valid annotation connects and classifies passages according to our guidelines. Key points include: each argument tree should have one conclusion, trees must be directed acyclic graphs (DAGs), and BF and PH passages should not link to argument trees, as

CMNA'25: The 25th International Workshop on Computational Models of Natural Argument, December 12, 2025, Online
*Corresponding author.
lbrenap1@jhu.edu (L. Brena); wjurayj1@jhu.edu (W. Jurayj); vandurme@jhu.edu (B. Van Durme)

	5 classes						4 classes				
Embedding/LM	Avg	Analysis	BF	Conclusion	PH	Rule	Avg	Analysis	BF	PH	Rule
TF-IDF	0.61	0.70	0.70	0.23	0.66	0.78	0.76	0.78	0.77	0.73	0.76
SBERT	0.63	0.71	0.65	0.30	0.70	0.78	0.73	0.79	0.66	0.69	0.77
Legal-BERT	**0.73**	**0.77**	**0.81**	**0.45**	**0.79**	**0.83**	**0.82**	**0.85**	**0.81**	0.78	**0.83**
Modern-BERT	0.64	0.68	0.78	0.39	0.63	0.71	0.71	0.77	**0.81**	0.53	0.73
GPT-5-mini	0.65	0.57	0.64	**0.45**	0.76	0.82	0.75	0.80	0.59	**0.79**	0.81
Random	0.21	0.31	0.14	0.21	0.20	0.21	0.20	0.29	0.15	0.14	0.24

Table 1
Linear SVC results across embeddings and GPT-5-mini classification results (F1-score; US Corporate Reorganizations with five classes and with four classes; Avg = macro average).

they serve as supporting text, not active components of the argument. An example of a valid syllogism annotation is provided in Figure 1.

The primary improvement of our approach lies in focusing on the functional role within the argument, rather than isolating individual roles, and employing a recursive method to build these structures. Our dataset aims to create a consistent and closely deductive structure that mimics human legal reasoning and provides an indication of logical deduction at each step of the argument construction.

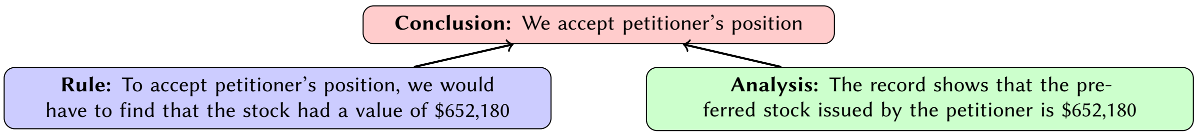

Figure 1: Annotation example of an atomic syllogism

Models. We compare (i) a Linear SVC [12] over TF–IDF, (ii) three transformer encoder-based masked language models [13]: SBERT [14], Legal-BERT [15], and ModernBERT [16], and (iii) a GPT-5-mini classifier prompted with brief label definitions (single-passage setting). Span classification experiments using the Linear SVC employed stratified 5-fold cross-validation for each embedding. The GPT-5-mini model was configured with low reasoning effort. ModernBERT provides an efficient long-context encoder with strong classification performance [16], while Legal-BERT offers domain-specific priors for legal text [15].

Results and Discussion. Table 1 shows macro F1 across embeddings and the LLM classifier. Moving from 5 classes to the 4-class variant (collapsing final *conclusions* into *analysis*) yields consistent gains: +0.15 (TF–IDF), +0.10 (SBERT), +0.09 (Legal-BERT), +0.07 (ModernBERT), and +0.10 (GPT-5-mini). Legal-BERT attains the highest average macro F1 among embeddings in the 4-class setup (0.82) as well as in the 5-class setting (0.73). We include a random baseline for reference. We hypothesize that removing the ambiguous "final conclusion" label reduces overlap with *analysis*, improving separability when passages are judged out of context. As a secondary contribution, the *background facts* and *procedural history* provide span-level section annotations of U.S. case law, which are formally structured in ECHR [1] and CJEU [9] cases but only informally organized in U.S. cases, limiting previous work to heuristics for identifying section boundaries [17, 18].

Limitations and Future Work. Our evaluation uses a single annotator over 26 opinions and classifies passages independently; richer context (e.g., graph models) may recover information lost at sentence scope by modeling relations between passages and allow visual reasoning about entailment and counterfactuals. Future work includes structured prediction over chains, explicit entailment links, and information retrieval to test argument completion.

References

[1] P. Poudyal, J. Savelka, A. Ieven, M. F. Moens, T. Goncalves, P. Quaresma, ECHR: Legal corpus for argument mining, in: E. Cabrio, S. Villata (Eds.), Proceedings of the 7th Workshop on Argument Mining, Association for Computational Linguistics, Online, 2020, pp. 67–75. URL: https://aclanthology.org/2020.argmining-1.8/.

[2] H. Xu, J. Šavelka, K. D. Ashley, Using Argument Mining for Legal Text Summarization, in: S. Villata, J. Harašta, P. Křemen (Eds.), Proceedings of the 33rd International Conference on Legal Knowledge and Information Systems (JURIX 2020), IOS Press, Virtual event, 2020, pp. 184–193. doi:10.3233/FAIA200862.

[3] N. Holzenberger, B. Van Durme, Factoring statutory reasoning as language understanding challenges, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 2742–2758. URL: https://aclanthology.org/2021.acl-long.213/. doi:10.18653/v1/2021.acl-long.213.

[4] W. Jurayj, N. Holzenberger, B. V. Durme, Language models and logic programs for trustworthy tax reasoning, 2025. URL: https://arxiv.org/abs/2508.21051. arXiv:2508.21051.

[5] P. Santin, G. Grundler, A. Galassi, F. Galli, F. Lagioia, E. Palmieri, F. Ruggeri, G. Sartor, P. Torroni, Argumentation structure prediction in cjeu decisions on fiscal state aid, in: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 247–256. URL: https://doi.org/10.1145/3594536.3595174. doi:10.1145/3594536.3595174.

[6] N. Weir, P. Clark, B. Van Durme, Nellie: A neuro-symbolic inference engine for grounded, compositional, and explainable reasoning, in: K. Larson (Ed.), Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, International Joint Conferences on Artificial Intelligence Organization, 2024, pp. 3602–3612. URL: https://doi.org/10.24963/ijcai.2024/399. doi:10.24963/ijcai.2024/399, main Track.

[7] D. S. Romantz, K. E. Vinson, Legal Analysis: The Fundamental Skill, 3 ed., Carolina Academic Press, Durham, NC, 2020. EISBN: 978-1-5310-1198-7.

[8] J. A. Gardner, C. P. Bartholomew, Legal Argument: The Structure and Language of Effective Advocacy, 3rd ed., Carolina Academic Press, Durham, North Carolina, 2020.

[9] G. Grundler, P. Santin, A. Galassi, F. Galli, F. Godano, F. Lagioia, E. Palmieri, F. Ruggeri, G. Sartor, P. Torroni, Detecting arguments in CJEU decisions on fiscal state aid, in: G. Lapesa, J. Schneider, Y. Jo, S. Saha (Eds.), Proceedings of the 9th Workshop on Argument Mining, International Conference on Computational Linguistics, Online and in Gyeongju, Republic of Korea, 2022, pp. 143–157. URL: https://aclanthology.org/2022.argmining-1.14/.

[10] I. Habernal, D. Faber, N. Recchia, S. Bretthauer, I. Gurevych, I. Spiecker genannt Döhmann, C. Burchard, Mining Legal Arguments in Court Decisions, Artificial Intelligence and Law (2023).

[11] F. Weber, T. Wambsganss, S. P. Neshaei, M. Soellner, Structured persuasive writing support in legal education: A model and tool for German legal case solutions, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2296–2313. URL: https://aclanthology.org/2023.findings-acl.145/. doi:10.18653/v1/2023.findings-acl.145.

[12] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297. URL: https://doi.org/10.1023/A:1022627411411. doi:10.1023/A:1022627411411.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[14] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL: https://arxiv.org/abs/1908.10084. arXiv:1908.10084.

[15] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2898–2904. URL: https://aclanthology.org/2020.findings-emnlp.261/. doi:10.18653/v1/2020.findings-emnlp.261.

[16] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, G. T. Adams, J. Howard, I. Poli, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 2526–2547. URL: https://aclanthology.org/2025.acl-long.127/. doi:10.18653/v1/2025.acl-long.127.

[17] A. Hou, W. Jurayj, N. Holzenberger, A. Blair-Stanek, B. Van Durme, Gaps or hallucinations? scrutinizing machine-generated legal analysis for fine-grained text evaluations, in: N. Aletras, I. Chalkidis, L. Barrett, C. Goanță, D. Preoțiuc-Pietro, G. Spanakis (Eds.), Proceedings of the Natural Legal Language Processing Workshop 2024, Association for Computational Linguistics, Miami, FL, USA, 2024, pp. 280–302. URL: https://aclanthology.org/2024.nllp-1.24/. doi:10.18653/v1/2024.nllp-1.24.

[18] A. B. Hou, O. Weller, G. Qin, E. Yang, D. Lawrie, N. Holzenberger, A. Blair-Stanek, B. Van Durme, CLERC: A dataset for U. S. legal case retrieval and retrieval-augmented analysis generation, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Findings of the Association for Computational Linguistics: NAACL 2025, Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 7898–7913. URL: https://aclanthology.org/2025.findings-naacl.441/. doi:10.18653/v1/2025.findings-naacl.441.