LawngNLI: a multigranular, long-premise NLI benchmark for evaluating models' in-domain generalization from short to long contexts

Anonymous ACL submission

Abstract

Natural language inference has trended with NLP toward studying reasoning over long contexts, with several datasets moving beyond the sentence level. However, short-sequence models typically perform best despite their sequence limits. Confounded by domain shifts between datasets, it has remained unclear whether long premises are truly needed at fine-tuning time to learn long-premise NLI. We construct LawngNLI,[1] with premises that skew much longer than in existing NLI benchmarks and are multigranular: all contain a short version. LawngNLI is constructed from U.S. legal opinions, with automatic labels with high human-validated accuracy. Evaluating on its long-premise NLI, we show top performance is achieved only with fine-tuning using these long premises. Models only fine-tuned on existing datasets and even our short premises (which derive from judge-selected relevant Entail excerpts in source documents) thus controlling for domain underperform considerably. Top performance is by short-sequence models prepended with a standard retrieval method filtering across each premise, but they underperform absent fine-tuning using long premises as inputs. LawngNLI also holds relevance for the legal community, as NLI is a principal cognitive task in developing cases and advice. Models performing well could double as retrieval or implication scoring systems for legal cases.

1 Introduction

In this work, we construct a new NLI benchmark LawngNLI and use it to demonstrate that models need long premises at fine-tuning time for top performance on long premises. Crucially, underperformance is considerable when models only see the long premises at evaluation, evidence that large-scale long-context datasets may indeed be needed for long-context tasks including NLI.

We construct LawngNLI from U.S. legal opinions via the Caselaw Access Project (The President and Fellows of Harvard University, 2018) that have been largely cleaned of in-line citations in order to read more naturally. Its premises are especially long and are multigranular. All examples exist in twin pairs having mutually contradictory hypotheses. LawngNLI's automatic labels derive from the dataset construction using (negation-based) contradiction and (similarity-based) neutralization algorithms. These labels exhibit an accuracy of 88.8% (94.7% for high-confidence human labels) on a subset with human-validated gold labels.

Our work stands within a fast-growing research area on how models can learn to reason over long text. Benchmarks for NLI, or Recognizing Textual Entailment (RTE), stretch back to Dagan et al. (2005). Recently, different "efficient" Transformer architectures have been proposed to address the obstacle of quadratic self-attention complexity in scaling to long sequences (Tay et al., 2020c). Most existing NLI benchmarks, meanwhile, contain largely short premises. Two outliers are two-label DocNLI (Yin et al., 2021) and three-label ConTRoL (Liu et al., 2021)[2]. However, while their premises often exceed the usual 512 maximum sequence length, they still largely are not near the typical maximum sequence lengths of key current long-sequence pre-

[1] Code for obtaining LawngNLI and unfiltered-LawngNLI to be released at [ANONYMIZED]. LawngNLI contains about 140 thousand twinned examples, while unfiltered-LawngNLI (a raw version left for future slicing and not balanced on labels) contains about 4.8 million untwinned candidate examples.

[2] Besides DocNLI (Yin et al., 2021) and ConTRoL (Liu et al., 2021), some previous papers evaluate one or a few efficient Transformer models on longer sequences on different tasks than NLI, specifically for long-context QA: e.g., BigBird (Zaheer et al.), NLQuAD (Soleimani et al., 2021), ETC (Ainslie et al., 2020), and ReadTwice (Zemlyanskiy et al., 2021). To our knowledge, the natural language tasks for existing "fair" benchmarks such as Long Range Arena (Tay et al., 2020b) include only generative or byte-level (albeit longer-byte sequence) tasks (e.g., Huang et al., 2021; Lu et al., 2021; Ma et al., 2021) or classification tasks with larger-than-byte tokenization which fit within 512 maximum sequence length (e.g., Xiong et al., 2021; Wang et al., 2020; Tay et al., 2020a).

trained models (e.g., 90th percentiles of their training examples in Appendix Table 4 are less than one third of 4096). For example, less than 4800 of DocNLI (Yin et al., 2021) training examples exceed 2048 tokens, compared with over 96000 for LawngNLI or over 35000 for its "analysis" subset (Appendix Table 4).

Our experimental evaluation (Section 3) includes both long-sequence and short-sequence models. On our dataset as on the two above, current long-sequence models are outperformed by short-sequence models. On ours, models prepended with a standard retrieval method (BM25 (Robertson and Zaragoza, 2009)) to filter across long premises turn out to perform best on long premises, but all evaluated models fall short when intermediate fine-tuning using only our natural short premises (which derive from human-selected relevant Entail excerpts) or existing NLI datasets as inputs. However, top performance on our dataset requires inputting the full long premises (including with retrieval) rather than only the first 512 tokens (including hypothesis length) or even our short premises.

Overall, our main contributions are: (1) a new NLI benchmark with multigranular premises multiple times longer than in existing NLI benchmarks across percentiles (see Appendix Table 4), (2) a comparison of state-of-the-art NLI models on LawngNLI, doubling as a testbed for AI-based systems for case retrieval/implication scoring which are central to legal research, (3) an evaluation showing how LawngNLI can teach models long-premise NLI, outperforming not only models transferred from existing datasets but also from our own short premises, thus moving from short context to long context directly with the same domain and examples.

2 LawngNLI Dataset

We construct LawngNLI beginning with all citations with parentheticals in official U.S. state and federal case opinions, via the Caselaw Access Project (The President and Fellows of Harvard University, 2018). When judges cite other cases in an opinion, they may highlight content or takeaways from those cases in a parenthetical.[3] Starting with Entail examples, our long premises are the majority opinion cited by the judge, and our short premises are the pages cited by the judge. We extract these parentheticals and the cases and pages they cite (using Eyecite (Cushman et al., 2021)) to build Entail examples, then apply a contradiction algorithm and a neutralization algorithm to convert 1/3 each of the original Entail examples into Contradict and Neutral examples, respectively. Detailed steps are outlined in Appendix Section A.1.

Appendix Tables 2 and 3 show sample examples for each label from our dataset, including distractor premise excerpts (not annotated in the dataset) and other hypotheses paired with the same premise. We compare LawngNLI with existing NLI datasets in Appendix Table 4. LawngNLI's long version of its premises skew much longer than premises in existing datasets: its 10th percentile is near the 90th percentile for the longest existing NLI datasets presented (DocNLI (Yin et al., 2021) and ConTRoL (Liu et al., 2021)). The best-performing models for both use a maximum sequence length of 512, using just initial premise tokens.

2.1 Automatic Labels and Human Assessment

LawngNLI includes only automatic NLI labels. The Entail labels were effectively "annotated" by the judge authoring the (hypothesis) parenthetical citing another case's pages, but our construction algorithms could import some error rate. Thus these labels are assessed for accuracy. Using 300 consensus gold labels from Amazon Mechanical Turk workers (screened on NLI items but not per se for experience with legal text), we find a 88.8% human-validated accuracy (94.7% for high-confidence human labels). Detailed steps are outlined in Appendix Section A.3. Appendix Table 5 shows human-assessed characteristics for LawngNLI's "analysis" subset studied in Section 3.

2.2 Previous NLU Datasets From Legal Text

AutoLAW and CaseHOLD (Mahari, 2021; Zheng et al., 2021) construct datasets for a distinct task of predicting holdings from other cases that support the arguments in the nearby context in the *citing* case. These holdings exhibit an argument support relation with respect to their surrounding context, as opposed to necessarily any NLI relation. Other papers seek to predict legal judgments from cases (Chalkidis et al., 2019).

The legal tasks closest to ours are from the annual COLIEE workshop.[4] In the 2021 formulation,

[3]These explanatory parentheticals are used by, for example, the legal research platform Casetext (Arredondo, 2017).

[4]https://sites.ualberta.ca/~rabelo/

Task 2 requires identifying which paragraph from one Canadian federal case implies a decision in another case. Task 4 requires identifying a yes/no answer to a legal question based on portions of the Japanese civil code. However, these tasks do not fully map to three-label NLI. And the training corpora (in the hundreds of examples) are ballpark 1000 times smaller than usual single-sentence benchmarks, making supervised learning alone insufficient for reliably training models to generalize (Hudzina et al., 2020; Rabelo et al., 2021; Kim et al., 2021; Schilder et al., 2021).

3 Experimental Evaluation

Our experiments on LawngNLI test our main research questions. They help illuminate whether large-scale, long-premise NLI datasets are needed at fine-tuning time in order to perform well on long-premise NLI, with implications for other long-context NLP tasks as well.

RQ1: Can models fine-tuned using existing NLI datasets or our short premises perform competitively when evaluated with our long premises, as compared to top performing models fine-tuned using the long premises (including those starting by filtering premises with a standard retrieval method)?

Because of LawngNLI's multigranularity, we can also make a direct comparison for each model.

RQ2: Can models fine-tuned using our short premises perform competitively when evaluated with our long premises, as compared to those same models fine-tuned using our long premises (including with a standard retrieval method)?

3.1 Approach

We choose models that are top performing on existing NLI benchmarks, using their HuggingFace (Wolf et al., 2020) implementation. The full list with rankings is in Appendix Section A.4. We use only LawngNLI's "analysis" subset: with long premises at most 4096 tokens, via a RoBERTa (Liu et al., 2019) tokenizer.

Before moving to LawngNLI, we seek to improve these models' ability on general NLI. We create three versions of each by performing an intermediate fine-tuning on each of the three included existing NLI datasets. We utilize the training sets from three existing NLI benchmarks: three-label ANLI (Nie et al., 2020)[5] which contains MNLI (Williams et al., 2018), three-label ConTRoL (Liu et al., 2021)[6][7], and two-label (Entail, Not Entail) DocNLI (Yin et al., 2021)[8]. Premises in DocNLI and ConTRoL skew longer than most NLI benchmarks, albeit typically not as long as in LawngNLI (see Appendix Table 4).

The three versions are then further fine-tuned on LawngNLI. This fine-tuning is run separately on long premises and short premises. Performance is evaluated both before and after this fine-tuning.

We run fine-tuning on LawngNLI by adapting the code from Xiong et al. (2021)[9]. We used a batch size of 32 and learning rate of 1e-5. See implementation details in Appendix Section A.2.

3.2 Analysis and Results

3.2.1 RQ1: Can Models Compete For Top Overall Performance On Our Long Premises Absent Fine-tuning On Them?

We find a considerable gap in performance with long premises between the top models that have versus have not been fine-tuned on our long premises. Thus at least for our dataset, long premises are needed to perform competitively on our long-premise NLI.

We start with our full evaluation panel: our pretrained models fine-tuned on existing NLI datasets. In Appendix Table 6, we benchmark each by evaluating separately on LawngNLI's long and short premises, both before and after fine-tuning.

Then for further analysis in Table 1, we choose albert-xxlarge-v2_anli, roberta-large_anli, and google_bigbird-roberta-base_anli, as short- and long-sequence models performing at or near the top on both our short and long premises (and for the top setups (4) and (6) for comparison, vanilla roberta-large).

For both fine-tuning and evaluation, we test prepending models with a module using BM25 (Robertson and Zaragoza, 2009) retrieval to filter

COLIEE2021/

[5] https://github.com/facebookresearch/anli

[6] https://github.com/csitfun/ConTRoL-dataset/

[7] Following this paper, we fine-tune on ANLI and then ConTRoL.

[8] https://github.com/salesforce/DocNLI. See Appendix Section A.2 for details about converting LawngNLI to two labels.

[9] https://github.com/mlpen/Nystromformer

Needs long premises for fine-tuning	No				Yes		
Fine-tuning	Short premise			BM25 retrieval on short premise	Long premise	BM25 retrieval on long premise	Hypotheses only
BM25 retrieval on long premise at evaluation	No (1)	Yes (2)	No [512 [tokens] (3)	Yes (4)	No (5)	Yes (6)	(7)
[Entail/Neutral/Contradict. Chance=1/3]							
bigbird_anli	0.613+/-0.015	0.762+/-0.014	0.666+/-0.015	0.767+/-0.013	0.77+/-0.013	0.821+/-0.012	0.55+/-0.016
albert_anli	0.742+/-0.014	0.817+/-0.012	0.742+/-0.014	0.819+/-0.012	0.789+/-0.013	0.868+/-0.011	0.512+/-0.016
roberta_anli	0.716+/-0.014	0.789+/-0.013	0.716+/-0.014	0.81+/-0.013	0.778+/-0.013	0.859+/-0.011	0.538+/-0.016
roberta_vanilla				0.81+/-0.013		0.866+/-0.011	0.555+/-0.016
Maximum of p-values versus (6)	0	0	0	0			
N	3966						

Table 1: Performance of top models (see Appendix Table 6 for versions) and baselines, on *long premises only*: Accuracy on test set within LawngNLI's "analysis" subset (long premise at most 4096 tokens). The error provided is the larger deviation of the Clopper-Pearson (Clopper and Pearson, 1934) exact binomial 95% confidence bounds. The p-values all round to zero from an exact binomial McNemar's (McNemar, 1947) test for a statistically significant difference in accuracies between each model's best version fine-tuning using short premises as inputs (4) and its best version fine-tuning using long premises as inputs (6). For (3), 512 tokens is the overall sequence limit.

the top 5 highest scoring paragraphs across the long premise when querying the hypothesis. While these models outperform those that do not filter, it does not follow that coherent, relevant short premises suffice. There may be less relevant portions of the long premise to filter out. But both with or without retrieval, fine-tuning on a natural candidate (namely our own) for relevant short premises based on human judgment is shown to fall short when evaluated on long premises.

We also evaluate models on hypotheses only, as a test for spurious correlates with the NLI label or artifacts of our contradiction or neutralization algorithms (Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018; Yin et al., 2021). Labels show some modest predictability above random from our hypotheses at 0.555 at the highest, in line with other NLI datasets.[10]

ALBERT-xxlarge-v2$_{ANLI}$ fine-tuned on our long premises with BM25 retrieval performs best on our long premises at 0.868, 0.049 higher than the top model fine-tuned using our short premises.

[10]Similar to ANLI (Nie et al., 2020) A1 at 0.497 and MNLI at 0.55 (Williams et al., 2018; Poliak et al., 2018) and slightly above ANLI later rounds and ConTRoL (Liu et al., 2021) in the 0.40s.

3.2.2 RQ2: Can Models Compete With Their Own Top Performance On Our Long Premises Absent Fine-tuning On Them?

The short-sequence models tend to gain more accuracy from fine-tuning on long (relative to short) premises. Still, all models need long premises as inputs for their best performance on long premises.

4 Conclusion and Future Work

Our results show that state-of-the-art long and short-sequence models need fine-tuning on our long premises to perform competitively on them. Short premises and existing NLI datasets do not suffice. While models fine-tuned on our long premises perform best filtering with a retrieval method, models underperform considerably when fine-tuning on natural short premises (not derived from across the text of our long premises).

Other aspects of LawngNLI are left for future study. This includes the portion with premises exceeding 4096 tokens. unfiltered-LawngNLI could be re-sliced to vary dataset difficulty. Since LawngNLI consists of legal argumentation, there may be other complexities such as "distractor" counterarguments and hierarchical, multi-factor reasoning across the text.

5 Ethical Considerations

Considerations for general NLI have been explored elsewhere (e.g., for gender bias by Sharma et al. (2021)).

We discuss some considerations for the legal aspect. On the benefit side, NLI is a principal cognitive task in law, so progress here also stands to benefit the legal community: building court cases and advising clients essentially is arguing for and against different natural language inferences from legal texts and facts. Practitioners must move between case text and the entailments and contradictions that they aim to support or counter.

LawngNLI provides a training and test set for developing models for NLI-based case retrieval or implication scoring systems, which could aid in reducing the practitioners' time and industry's annotation costs around legal research. All around the legal system, the pay grade and spare bandwidth of legal counsel is frequently starkly imbalanced between parties with adversarial interests: whether people in the courtroom or settlement conference, consumers or companies in a negotiation boardroom, or in everyday society where behavior is shaped by prospects of legal action. Anything that makes legal research and thus legal counsel cheaper, including lightweight or affordable case retrieval systems, can contribute toward fairer access to legal representation and justice regardless of financial means.

The annual revenue of the legal research industry is in the multiple billions of dollars.[11] And legal research industry size arguably vastly underestimates the full societal cost of suboptimal case retrieval: this cost should also include the time and resources expended by human legal researchers in the loop (paralegals and lawyers) in unnecessary iterating with any suboptimal retrieval in current systems.

Although the leading case retrieval systems that lawyers rely upon (Westlaw, Lexis Advance, etc.) utilize proprietary algorithms, there is some evidence from reverse engineering (Callister, 2020) that they may compare on bag of words or simple embeddings. Even if they use dense retrieval, systems not fine-tuned for NLI are unlikely to retrieve very effectively when querying case text for implications not directly stated in the text or annotations (e.g., those at a different level of specificity or requiring compositional reasoning). Instead, holdings and rules inferable from case text must be extracted through costly human annotation and curation. And even then, lawyers must happen upon keywords for the rules that hold implications for their case. Again, in contrast, an NLI model that performed well on LawngNLI could crosswalk between cases as premises against implications as hypotheses and perform implication-based retrieval automatically.

On the risk side, while prospective human reliance for decision making on erroneous model predictions is an ever-present consideration in NLP, we do not view this as a practical risk for LawngNLI. Everyday people can turn to numerous simple articles online summarizing the law, without digging into complex case retrieval and jurisprudence. And regarding advising others, lawyers bound by professional duties are exclusively authorized to practice law in the U.S. and around the world.[12] Nothing can even be done just knowing the most relevant cases or implications; they must be synthesized by human judgment into an argument sound enough to pass the muster of judges and juries. In other words, legal NLI models are in no way lawyers. Instead, they can work as screening tools for practitioners who then must apply their own judgment to make the results useful. In this way, legal NLI models could help save the resources of lawyers and clients and help improve the quality of legal representation.

[11] As of 2020: e.g., Thomson Reuters, RELX.

[12] https://www.ibanet.org/MediaHandler?id=199b20ec-b7ab-4ef4-99c4-cd45c7b6371b

References

Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. ETC: Encoding Long and Structured Inputs in Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 268–284.

Pablo D. Arredondo. 2017. Harvesting and Utilizing Explanatory Parentheticals. *SCL Rev.*, 69:659. Publisher: HeinOnline.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, and Tri Nguyen. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Yonatan Bilu, Daniel Hershcovich, and Noam Slonim. 2015. Automatic claim negation: Why, how and when. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 84–93.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly, Cambridge.

Paul D. Callister. 2020. Law, artificial intelligence, and natural language processing: A funny thing happened on the way to my search results. *Law Libr. J.*, 112:161. Publisher: HeinOnline.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323.

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.

Charles J. Clopper and Egon S. Pearson. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, pages 404–413. ISBN: 0006-3444 Publisher: JSTOR.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the trec 2019 deep learning track. *arXiv preprint arXiv:2003.07820*.

Jack Cushman, Matthew Dahl, and Michael Lissner. 2021. eyecite: A tool for parsing legal citations. *Journal of Open Source Software*, 6(66):3617. ISBN: 2475-9066.

I. Dagan, O. Glickman, and B. Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of PASCAL first Workshop on Recognising Textual Entailment*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 113–122, Virtual Event Canada. ACM.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient Attentions for Long Document Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436.

John Hudzina, Kanika Madan, Dhivya Chinnappa, Jinane Harmouche, Hiroko Bretz, Andrew Vold, and Frank Schilder. 2020. Information Extraction/Entailment of Common Law and Civil Code. In *JSAI International Symposium on Artificial Intelligence*, pages 254–268. Springer.

Mi-Young Kim, Juliano Rabelo, and Randy Goebel. 2021. BM25 and Transformer-based Legal Information Extraction and Entailment. New York, NY, USA. ACM.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual Constituency Parsing with Self-Attention and Pre-Training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency Parsing with a Self-Attentive Encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pretraining for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2021. Natural Language Inference in Context-Investigating Contextual Reasoning over Long Texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13388–13396. Issue: 15.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. 2021. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*.

Xuezhe Ma, Xiang Kong, Sinong Wang, Chunting Zhou, Jonathan May, Hao Ma, and Luke Zettlemoyer. 2021. Luna: Linear Unified Nested Attention. *arXiv preprint arXiv:2106.01540*.

Robert Zev Mahari. 2021. AutoLAW: Augmented Legal Reasoning through Legal Precedent Prediction. *arXiv preprint arXiv:2106.16034*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Van L, Sofie eghem, Adriane Boyd, Henning Peters, Paul O'Leary McCann, Maxim Samsonov, Jim Geovedi, Jim O'Regan, György Orosz, Duygu Altinok, Søren Lind Kristiansen, Roman, Explosion Bot, Le Fiedler, er, Grégory Howard, Wannaphong Phatthiyaphaibun, Yohei Tamura, Sam Bozek, Murat, Mark Amery, Björn Böing, Pradeep Kumar Tippa, Leif Uwe Vogelsang, Ramanan Balakrishnan, Vadim Mazaev, GregDubbin, Jeannefukumaru, and Walter Henry. 2021. explosion/spaCy: v3.1.4: Python 3.10 wheels and support for AppleOps.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

Juliano Rabelo, R Goebel, y, Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, and Ken Satoh. 2021. Summary of the Competition on Legal Information Extraction/Entailment (COLIEE) 2021. New York, NY, USA. ACM.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*. Citeseer.

Nils Reimers and Iryna Gurevych. 2021. The Curse of Dense Low-Dimensional Information Retrieval for Large Index Sizes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 605–611, Online. Association for Computational Linguistics.

RELX. Annual Report and Financial Statements 2020. Technical report.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Frank Schilder, Dhivya Chinnappa, Kanika Madan, Jinane Harmouche, Andrew Vold, Hiroko Bretz, and John Hudzina. 2021. A Pentapus Grapples with Legal Reasoning. In *Proceedings of the Eigth International Competition on Legal Information Extraction/Entailment (COLIEE 2021)*, New York, NY, USA. ACM.

Shanya Sharma, Manan Dey, and Koustuv Sinha. 2021. Evaluating Gender Bias in Natural Language Inference. *arXiv preprint arXiv:2105.05541*.

Tom De Smedt and Walter Daelemans. 2012. Pattern for Python. *Journal of Machine Learning Research*, 13(66):2063–2067.

Amir Soleimani, Christof Monz, and Marcel Worring. 2021. NLQuAD: A Non-Factoid Long Question Answering Data Set. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1245–1255.

Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. 2020a. Sparse sinkhorn attention. In *International Conference on Machine Learning*, pages 9438–9447. PMLR.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2020b. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020c. Efficient transformers: A survey. *arXiv preprint arXiv:2009.06732*.

The President and Fellows of Harvard University. 2018. Caselaw Access Project.

Thomson Reuters. Annual Report 2020. Technical report.

Masatoshi Tsuchiya. 2018. Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Chantal van Son, Emiel van Miltenburg, and Roser Morante. 2016. Building a Dictionary of Affixal Negations. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM)*, pages 49–56, Osaka, Japan. The COLING 2016 Organizing Committee.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Thomas Wolf, Julien Chaumond, Lys Debut, re, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, and Sam Shleifer. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. 2021. Nyströmformer: A Nyström-based Algorithm for Approximating Self-Attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14138–14148. Issue: 16.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A Large-scale Dataset for Document-level Natural Language Inference. *arXiv preprint arXiv:2106.09449*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, and Li Yang. Big Bird: Transformers for Longer Sequences.

Yury Zemlyanskiy, Joshua Ainslie, Michiel de Jong, Philip Pham, Ilya Eckstein, and Fei Sha. 2021. ReadTwice: Reading Very Large Documents with Memories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5189–5195.

Lucia Zheng, Neel Guha, Br Anderson, on R., Peter Henderson, and Daniel E. Ho. 2021. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. *arXiv preprint arXiv:2104.08671*.

A Appendix

A.1 Dataset Construction Procedure

A.1.1 Extraction From Caselaw Access Project

LawngNLI is constructed starting with all xml case files from the April 21, 2021 bulk export from the Caselaw Access Project (The President and Fellows of Harvard University, 2018). The word count of the full original corpus before processing at about 12 billion[13] is around three times that of English Wikipedia[14], though for our premises we limit to only the majority opinions.

Entail examples are pairs of citation parentheticals (hypotheses) and excerpts of majority opinions from cited cases with resolvable pincites (premises), extracted from case files using Eyecite (Cushman et al., 2021).[15] In this paper, we only include examples from citations including a resolvable pincite (e.g., does not contain letters).

Examples are dropped or modified by simple "accuracy" filters.[16]

The short version of the premise consists of the resolvable cited pages within the cited case's majority opinion, while the long version of the premise consists of the cited case's full majority opinion.

A.1.2 Identifying (Pivotal) Negation in Hypotheses

Next the Entail examples are automatically labeled by whether their hypotheses contain (pivotal) negation or not, depending on whether the contradiction algorithm described in Appendix Section A.1.4 removes or adds negation, respectively. Pairs with hypotheses rejected for processing by our contradiction algorithm are dropped from the dataset.

[13] https://case.law/docs/site_features/trends

[14] About 4 billion as of December 1, 2021: https://web.archive.org/web/20211201013917/https://en.wikipedia.org/wiki/Special:Statistics

[15] Where Eyecite associates multiple consecutive citations resolving to the same case with the same citation parenthetical, only the first citation and its pincite, if any, is paired with the parenthetical and included as an example.

[16] First, as an overbroad criterion to exclude examples where the (converted or unconverted) original Entail hypothesis was a parenthetical in a case that was later overturned, we drop all examples with hypotheses from cases where a later case shared the same party names in the same or reverse order.

Second, parentheticals with citations including a case history flag (e.g., "acq.","aff'd") are excluded.

Third, we drop examples with hypotheses that contain certain regex keywords ('quoting|en banc|omitted|mphasis|applying|citing|concur|dissent|majority|, in chambers|per curiam|Lexis|opinion| v. |§|¶|[0-9]') associated with parentheticals describing "metadata" about the cited case rather than its content.

Fourth, verbs ending with "ing" followed by "that" at the beginning of remaining hypotheses many times take a supporting stance toward the subsequent subordinate clause, so to adapt such hypotheses to be more similar to a standalone sentence, we remove such initial words and the subsequent "that" in hypotheses.

Finally, sentences are normalized with spaCy 3.1.1 (Montani et al., 2021) to, e.g., process contractions.

Since the absence versus presence of such negation in the hypothesis results in contradictory truth values (and thus also flips the NLI label between Entail and 'Contradict), such negation can be called "pivotal." Negation is defined this way throughout the paper except in Appendix Table 4 when comparing to other datasets, since our contradiction algorithm might exhibit a different error rate on those datasets and confound the comparison. For this reason, greater than 50% of LawngNLI's hypotheses contain negation in Appendix Table 4, even though the dataset is constructed to contain 50% (pivotal) negation hypotheses.

A.1.3 NLI Label Split

Within examples from cases from each state (or federal) and pivotal negation or not, entail examples are randomly assigned to be 1/3 Entail, 1/3 converted to Neutral, and 1/3 converted to Contradict.

A.1.4 Converting Entail Examples to Contradict Examples: Contradiction Algorithm

For examples labeled Contradict in Appendix Section A.1.3, we use our contradiction algorithm to add or remove pivotal negation[17] from the hypothesis, toward aligning the NLI relation with the label.

Our contradiction algorithm builds on the negation algorithm outlined in Section 4.2 of Bilu et al. (2015), which in their paper was annotated by majority vote to have generated an opposing claim with probability 0.79.[18]

The algorithm chooses a random sentence for adding or removing negation and leaves the others unchanged. It finds a non-compound independent clause within the chosen sentence and then makes the first applicable change in the list below. If none of the changes' conditions apply, the hypothesis is rejected for processing by the algorithm.[19]

1. If there are any contradictable indefinite pronouns in the first highest-level noun phrase, the first one is changed to a contradictory pronoun (e.g., "some" to "none" or "neither" to "either").

2. If there are any verb phrases, the first highest-level verb phrase is contradicted using a modified version (e.g., also reversing negation by removing "do"/"does"/"did"+"not") of the negation algorithm from Bilu et al. (2015) mentioned above.

3. If there are any adjective phrases, the first ['no','not','never'] is removed from or else a 'not' is added to the first highest-level adjective phrase or past participle.

A.1.5 Filtering

Now we apply simple "difficulty" filters: examples with hypotheses containing quotation marks or fewer than four words or with at least 50% bigram overlap with their premise are dropped.

A.1.6 Converting Entail Examples to Neutral Examples: Neutralization Algorithm

For examples labeled Neutral in Appendix Section A.1.3, we use our neutralization algorithm to match the hypothesis with a different premise, toward aligning the NLI relation with the label. To balance attrition, the neutralization algorithm is applied to all examples regardless of NLI label, but only the hypotheses from Neutral examples are actually re-paired with the assigned premise.

The candidates for matching with each hypothesis are the premises from all examples that are from cases in the same state as the original premise (or from a federal case if the original premise is from a federal case). Excluded from candidacy are premises from cases citing or cited by the case containing the original hypothesis.

A hypothesis is paired with a candidate premise as follows. The short version of the premise is used for this step.

First, the top 30 (dot-product) nearest neighbors of the hypothesis among the candidates are retrieved using FAISS (?)[20] on msmarco-distilbert-

[17] "Pivotal" negation is negation the absence versus presence of which results in at least some contradictory truth values for the hypothesis, flipping its NLI label from Entail to Contradict.

[18] Hypotheses are parsed with the Berkeley Neural Parser 0.2.0 'benepar_en3' with spaCy 3.1.1 'en_core_web_lg' (Kitaev et al., 2019, Kitaev and Klein, 2018, Montani et al., 2021). Verb tense is modified using NLTK 3.6.2 WordNet Lemmatizer and Pattern 3.6 conjugate function (?; Bird et al., 2009; Smedt and Daelemans, 2012). We explored attempting to negate adjectives and verbs using the lexical negation dictionary compiled by van Son et al. (2016) but ultimately limited to just using direct negation.

[19] This includes rejecting hypotheses consisting of verb phrases not nested within independent clauses; since these are rarely found in negated form in the original dataset, including them would leave an artifact of this contradiction algorithm. So for these hypotheses, we prioritize balance across labels over coverage of candidate examples.

[20] https://github.com/facebookresearch/faiss

base-tas-b embeddings (Hofstätter et al., 2021)[21] via Sentence-Transformers (https://github.com/UKPLab/sentence-transformers, Reimers and Gurevych, 2021).

Second, candidate premises with which a hypothesis has at least 50% bigram overlap are dropped. This step preserves the filter applied earlier to all examples through the re-pairing for the Neutral examples.

Finally, Neutral hypotheses only are paired with their remaining candidate premise with respect to which it has the highest BM25 (Robertson and Zaragoza, 2009) score via Gensim 3.8.3 (Rehurek and Sojka, 2010). For hypotheses of all labels, if no candidate premises remain, their example is dropped.

A.1.7 Balancing

We split the dataset into "analysis"/non-"analysis" subsets by the inclusion criterion for this paper's experimental evaluation (Section 3): whether the sequence length of an example's long premise is at most 4096 tokens, via a RoBERTa (Liu et al., 2019) tokenizer.

Within each of the "analysis"/non-"analysis" subsets, the dataset is then downsampled by randomly sampling each of the three label-plus-negation groups closed under the contradiction operation (Entail+negation plus Contradict+non-negation; Contradict+negation plus Entail+non-negation; Neutral+negation plus Neutral+non-negation) down to the minimum of their example counts. A 90/5/5 train/val/test split is stratified by "analysis"/non-"analysis" subset and these groups.

Each example is then complemented with its contradictory twin: the same premise paired with the hypothesis modified by adding or removing pivotal negation (so applying the contradiction algorithm). Neutral labels are unchanged from the original example, while Entail and Contradict labels are flipped. This twinning balances the dataset within the "analysis"/non-"analysis" subsets on NLI label by pivotal negation versus not.

[21] https://huggingface.co/sebastian-hofstaetter/distilbert-dot-tas_b-b256-msmarco. The Sentence-Transformers www.sbert.net documentation shows retrieval using dot-product similarity on this model's embeddings to perform best among several models on TREC-DL 2019 (Craswell et al., 2020) and the MS Marco Passage Retrieval dataset (Bajaj et al., 2016).

A.1.8 Citation Removal Algorithm and Prepending

Our algorithm here attempts to remove as many in-line citations from premises as it can so that the premises are more customary English-language texts. The processed premises are studied in this paper. But the dataset obtainable from code to be released will include the preprocessed premises as well for future study. Finally, we copy and prepend at the beginning of premises the minimum number of paragraphs from the end that contain 512 tokens, to limit models from relying on cues for the NLI label near the start.

A.2 Implementation Details

For our intermediate fine-tuning, we adapt the code and largely follow the respective model hyperparameters and fine-tuning settings of the three existing NLI benchmarks. The settings that we modify rather than follow are: attention gradient checkpointing, GPU setup while not changing accumulated batch size, and maximum sequence length (with our sequence lengths longer for certain models, we also train for 3 epochs instead of 5 on DocNLI (Yin et al., 2021)). Maximum sequence lengths for intermediate fine-tuning are the lesser of the model maximum and 2048 (except for a maximum sequence length of 156 for pretrained short-sequence models fine-tuned on ANLI, consistent with Nie et al. (2020)[22]).

After intermediate fine-tuning, the long-sequence models' maximum sequence lengths are increased to 4096 for further fine-tuning on LawngNLI. We adapt the code from Xiong et al. (2021)[23]. We adapted this code in order to allow compatibility with their suite of efficient Transformers, but ultimately we did not pretrain them and did not further explore including them after several (initialized with copied RoBERTa-base (Liu et al., 2019) embeddings) did not rise far above random accuracy for LawngNLI fine-tuning under some initial hyperparameters explored. This does reflects little on these models since we did not pretrain them.

For fine-tuning on LawngNLI, we use a batch size of 32 and learning rate of 1e-5. We explored hyperparameters among those explored by RoBERTa (Liu et al., 2019) for GLUE (Wang et al.,

[22] https://github.com/facebookresearch/anli

[23] https://github.com/mlpen/Nystromformer

2018), along with batch size 128 so that all of our models in Appendix Section A.4 would start to converge during fine-tuning starting from their initial losses and accuracies. Beyond this, we did not conduct a full hyperparameter search based on model performance.

NVIDIA 12GB TITAN Xp, 11GB GeForce GTX 1080 Ti, 11GB GeForce RTX 2080 Ti, 24GB TITAN RTX GPUs, and NVIDIA 48GB RTX A6000 GPUs were used for intermediate fine-tuning and fine-tuning on LawngNLI.

External code is from GitHub repositories, with repository forking permitted under contemporaneous GitHub's Terms of Service. External models are from HuggingFace Transformers (Wolf et al., 2020; contemporaneously governed by an Apache License 2.0 permitting modification, distribution, etc.) or from GitHub repositories. Cases from the Caselaw Access Project (The President and Fellows of Harvard University, 2018) are used to construct our datasets. Any dataset sharing will comply with Caselaw Access Project (The President and Fellows of Harvard University, 2018) terms of access or else any separate agreement with the licensor. In particular, if necessary to ensure this compliance, we will share code for constructing our datasets rather than the datasets themselves.

A.2.1 Existing NLI Datasets

For models in Appendix Section A.4 with fine-tuned checkpoints provided at `https://github.com/facebookresearch/anli` (ALBERT-xxlarge-v2 (Lan et al., 2019), BART-large (Lewis et al., 2020), and RoBERTa-large (Liu et al., 2019)), we used these model checkpoints. Otherwise we fine-tuned the models, aiming to replicate the original hyperparameters.

To transfer learning from two-label DocNLI, the models intermediate-fine-tuned on DocNLI are further fine-tuned and evaluated on a two-label version of LawngNLI (where the Entail examples are duplicated and then (Entail, Neutral and Contradict) labels are mapped to (Entail, Not Entail)). This construction balances the two-label version between (Entail, Not Entail). For further fine-tuning these models on LawngNLI, the number of epochs is then halved. This is equivalent to splitting the Neutral and Contradict examples (now labeled Not Entail) in the original three-label dataset in half across pairs of consecutive original epochs (1 and 2, 3 and 4, and so on) so that the fine-tuning example count is 2/3 of the original dataset times the original number of epochs. Except that example shuffling also pools examples between these consecutive original epochs.

A.3 Procedure for Human Assessment

Human assessment was limited to Amazon Mechanical Turk Master Workers based in the U.S.

Assessed accuracy of examples with long premises is lower than for with short premises, even though the former arguably should have a higher accuracy against the ground truth: they are a superset of the information in the short premise, thereby providing additional context while being written to be internally consistent. It may be then that the human-assessed error rates for the automatic labels are themselves imperfect against the ground truth, especially for more difficult examples.

Human assessment proceeded as follows:

- Examples were each reviewed by two workers in batches of 28 examples, which were drawn from a first and then second set of 504 examples with sequence length at most 4096. Each set consists of a stratified random sample of test examples. The stratification is as follows: First, balance over the Cartesian product of the automatic label and pivotal negation versus not. Then half using the short premise and half using the long premise.

- Workers provided NLI labels for batches effectively without a time limit (batches due 1 week after assignment). Batches were issued until there were 300 non-screening examples with their two worker labels in agreement. The accuracy of these examples' automatic labels was then evaluated against those agreed labels (as gold).

- Workers were advised that they were providing NLI labels to be used in an academic analysis evaluating a new dataset.

- Workers were paid above the U.S. federal minimum wage on "reasonable" (as opposed to actual) time spent: 2 hours per batch, but workers may have spent more or less time on any batch up to 1 week. In addition, a performance bonus was provided for each label deemed correct on a screening example.

- Worker screening was as follows:

- First, workers needed to qualify by answering at least 4 examples correct (credit was sometimes given for an incorrect label with defensible reasoning) on an initial pre-screen of six screening examples within a half hour. Several batches not meeting the minimum performance described in the instructions (which was itself below the qualification threshold) were rejected.
- Because NLI is multiple choice, there is a risk that the initial screening may be insufficient or that workers may not consider examples thoroughly in selecting options (or even guess somewhat randomly). Though we saw evidence directly in the gold dataset, we included screening examples in the ongoing batches. We excluded two workers' examples for falling below a threshold.
  * Each batch contains 3 screening examples and 25 non-screening examples.
  * Labels on screening examples were selected by a co-author. Screening examples were not included in the 300 examples in the gold dataset.
  * Workers could continue completing the batches of 28 unless at a time of audit their cumulative accuracy on screening examples fell below 50% (after at least 5 screening examples). If their cumulative accuracy fell below this threshold, they were still paid for all completed batches but the examples they labeled were not included in the gold dataset.
- Workers provided labels via a six-option scale: 'definitely entail', 'probably entail', 'definitely neutral', 'probably neutral', 'definitely contradict', 'probably contradict'.
- For examples that workers labeled as entail or contradict, they also copied and pasted a portion of the premise relevant to determining the label they chose.
- We temporarily experimented with having a different version of the dataset assessed, but no workers labeled the same examples in that experiment and the current assessment set.

A.3.1 Instructions for the main NLI task

This is an italicized text version of the instructions, with some bolding omitted here.

" *How is text1 related to text2: Entail, Neutral, or Contradict?*

***To qualify for this task, you must first perform sufficiently well on the [SCREENING TASK], listed under the same requester**.*

This task is natural language inference/RTE (same as under our [SCREENING TASK]).

*Note: In completing this task, you are providing NLI labels to be used in an academic analysis evaluating a new dataset. **These items were constructed from legal texts and probabilistically selected from a larger dataset without screening. They may include sensitive or derogatory language.***

Items are in batches of 28. The time limit is one week, so that you can spend more time on certain difficult items if you wish.

I. REWARD

The reward per batch was calculated based on two hours of work. However, feel free to work at your own pace as the time limit is much longer.

As a percent of the base reward, there is an overall bonus of 20

II. EVALUATION

Assignments that are not accompanied by evidence of diligence in reading passages (snippets from text1 that are relevant) may be rejected.

If your running average accuracy on "validation items" falls too low, you may no longer be able to access further batches.

"Validation items" are a random subset of the items that are separately labeled: correct labels are determined independent of your response, though these labels will not be visible to you.

III. INSTRUCTIONS

Feel free to use the Find function to search the text. Some text1 portions standing alone could suggest an incorrect label and yet should be consistent with a correct label in the context of text1 overall. For example, if a text mentions a claim by a lower court or different case and then proceeds to reject that claim, then the text overall rejects that claim.

text2 may well refer to text1, and any of the three labels may apply. Note that in some items though, text2 may not be referring to text1 at all (e.g., it may be discussing a scenario from a text not included here, with different entities or conclusions neither entailed nor contradicted by text1). In the latter *cases, the correct label would be Neutral.*

IV. RESPONSES

(1 - Required.) SELECT EXACTLY ONE LABEL AMONG THE THREE that best describes how text1 is related to text2. As part of the label, select a level of confidence: PROBABLY or DEFINITELY (this confidence level will in no way affect the evaluation or reward, so you can just be honest).

(2 - Required if Entail or Contradict label is provided. PASTE A SNIPPET FROM TEXT1 SUPPORTING THE LABEL) This text can be short or incomplete. It is simply to help demonstrate reasonable diligence, only for if that is not already clear from the labels. A tiny fraction of the time should be spent on this step. ”

A.3.2 Instructions for the pre-screen task

This is an italicized text version of the instructions, with some bolding omitted here. An included illustrative example is also omitted here. Note that some earlier workers saw earlier versions.

“ *How is text1 related to text2: Entail, Neutral, or Contradict?*

[If this task is visible to you, the batch of items should be new to you and you may complete the task.]

I. OVERVIEW

6 items limited to 30 minutes. The reward is $3.75, with a bonus for 5 or 6 answers deemed correct for a total reward of $4 or $4.50, respectively.

For EACH item you answer, SELECT ONE LABEL AMONG THE THREE (there is exactly one reference answer for EACH item) and PASTE A SNIPPET FROM TEXT1 SUPPORTING THE LABEL. Any item with multiple or zero labels selected or not accompanied by relevant snippet from text1 (see II) will be marked incorrect.

Feel free to use the Find function to search the text. Some text1 portions standing alone could suggest an incorrect label and yet should be consistent with a correct label in the context of text1 overall. For example, if a text mentions a claim by a lower court or different case and then proceeds to reject that claim, then the text overall rejects that claim.

text2 may well refer to text1, and any of the three labels may apply. Note that in some items though, text2 may not be referring to text1 at all (e.g., it may be discussing a scenario from a text not included here, with different entities or conclusions neither entailed nor contradicted by text1). In the latter cases, the correct label would be Neutral.

Assignments with too few answers deemed correct (e.g., two or fewer) that are not accompanied by evidence of diligence in reading passages (relevant snippet from text1) may be rejected.

II. RESPONSES

(1 - Required.) Select the label that best describes how text1 is related to text2.

(2 - Required if Entail or Contradict label is provided. paste relevant snippet from text1) This text can be short or incomplete. It is simply to help demonstrate reasonable diligence, only for if that is not already clear from the labels. A tiny fraction of the time should be spent on this step.

(3 - Optional - include explanation for your label) Answers matching reference label will be given full credit regardless of if an explanation is provided. For other labels, it may depend on explanation.

III. LABELS (Entail, Contradict, and Neutral)

text1 in this task may be substantially longer than below. See "Instructions" for examples. text1 and text2 are sometimes known as premise and hypothesis, respectively.

Quoting from https://aclweb.org/aclwiki/index.php?title=Textual_Entailment_Portal and http://u.cs.biu.ac.il/~dagan/publications/RTEChallenge.pdf:

""An example of a positive TE (text entails hypothesis) is:

text: If you help the needy, God will reward you. hypothesis: Giving money to a poor man has good consequences.

An example of a negative TE (text contradicts hypothesis) is:

text: If you help the needy, God will reward you. hypothesis: Giving money to a poor man has no consequences.

An example of a non-TE (text does not entail nor contradict) is:

text: If you help the needy, God will reward you. hypothesis: Giving money to a poor man will make you better person.

The entailment need not be pure logical - it has a more relaxed definition: "t entails h (t $\Rightarrow$ h) if, typically, a human reading t would infer that h is most likely true."[1]"" ”

A.4 Evaluation Panel: List of State-Of-The-Art Pretrained Models

- Longformer-base (Beltagy et al., 2020)
- BigBird-RoBERTa-base (Zaheer et al.)

- ALBERT-xxlarge-v2 (Lan et al., 2019). It is ranked highest besides T5 models and third overall on MNLI (Williams et al., 2018)[24]. It also ranked highest on ANLI test A2 and A3[25].
- BART-large (Lewis et al., 2020). It ranked first on ConTRoL (Liu et al., 2021), after fine-tuning on ANLI (Nie et al., 2020).
- Custom Legal-BERT (Zheng et al., 2021). Pre-trained on the Caselaw Access Project (The President and Fellows of Harvard University, 2018) corpus.
- LEGAL-BERT-base-uncased, also known as LEGAL-BERT-SC (Chalkidis et al., 2020). It is pretrained on legal text from fields such as legislation, cases, and contracts.
- RoBERTa-large (Liu et al., 2019). It performed the better out of two models (over Longformer (Beltagy et al., 2020)) on DocNLI (Yin et al., 2021) and ranked second on ANLI test A1[26].

A.5 Appendix Tables

[24] https://paperswithcode.com/sota/natural-language-inference-on-multinli

[25] https://paperswithcode.com/sota/natural-language-inference-on-anli-test

[26] https://paperswithcode.com/sota/natural-language-inference-on-anli-test

Sample twin Entail/Contradict examples from LawngNLI	
Hypotheses from "analysis" subset	• *Contradict:* city acted affirmatively to create or increase risk of harm on city street by ignoring residents' requests to reduce speed limit or by taking down residents' signs indicating drivers should adhere to a lower speed limit • *Entail:* city did not act affirmatively to create or increase risk of harm on city street by ignoring residents' requests to reduce speed limit or by taking down residents' signs indicating drivers should adhere to a lower speed limit
Additional hypotheses	• *Entail:* failing to enforce or lower the speed limit on a residential street "did not create a 'special danger' to a discrete class of individuals..[ed.: excerpted]..as opposed to a general traffic risk to pedestrians and other automobiles" • *Contradict:* traffic laws and enforcement practices did not pose "a general traffic risk to pedestrians and other automobiles"
Relevant premise excerpts	• [ed.: Plaintiffs] ...submit that the City of Fort Thomas..violated their son's substantive due process rights by failing to act upon their request (and the requests of others) to lower the speed limit on the street..The police also removed signs posted by residents indicating that drivers should adhere to a 15 mile-per-hour speed limit.. • [ed.: Plaintiffs] ...alleged that the City's failure to maintain safe conditions on Garrison Avenue violated their son's substantive due process rights..established a "state-created danger" under DeShaney.. • ...DeShaney's holding..precludes [ed.: Plaintiffs'] argument that the Due Process Clause constitutionalizes a locality's choices about what speed limit to adopt for a given street or how to enforce that speed limit.. • There are two exceptions to the DeSha-ney rule..Under the second exception..a plaintiff may bring a substantive due process claim by establishing (1) an affirmative act by the State that either created or increased the risk that the plaintiff would be exposed to private acts of violence.. • [ed.: Plaintiffs] fail to satisfy any of the three requirements for establishing our circuit's "state-created danger" exception to DeShaney. First, the creation of a street and the management of traffic conditions on that street are too attenuated and indirect to count as an "affirmative act"..
Distractor premise excerpts	• ...After all, the City was told about the risks of not lowering the speed limit to 15 miles per hour (more accidents); it intentionally chose not to heed this warning (taking on the risk of more accidents); and the alleged risk came to pass when..was killed (an accident).. • ...For in one sense, it could be said that all governing bodies act with deliberate indifference when they consider and reject a traffic-safety proposal of this sort that comes with known risks..

Table 2: Sample twin Entail/Contradict examples from LawngNLI, also in the "analysis" subset analyzed in our experiments (Section 3): sequence length of long premise at most 4096. Each hypothesis pairs with the excerpted premise in a separate example. For those specific "Additional hypotheses" above, the examples containing them are in unfiltered-LawngNLI (see GitHub link in first footnote) but not LawngNLI, the core dataset studied in this paper.

Sample twin Neutral examples from LawngNLI	
Hypotheses from "analysis" subset	• *Neutral:* a parade permit requirement did not violate the First Amendment • *Neutral:* a parade permit requirement violated the First Amendment
Distractor premise excerpts	• ...Section 13k prohibits two distinct activities: it is unlawful either "to parade, stand, or move in processions or assemblages in the Supreme Court Building or grounds,".. • ...we shall address only whether the proscriptions of 13k are constitutional as applied to the public sidewalks..

Table 3: Sample twin Neutral examples from LawngNLI, but not in the "analysis" subset analyzed in our experiments (Section 3): sequence length of long premise at most 4096. Each hypothesis pairs with the excerpted premise in a separate example.

LawngNLI	Long premises		Short premises	
	"Analysis" subset	Full	"Analysis" subset	Full
Premise length	[970, 1527, 2339, 3154, 3693]	[1285, 2179, 3692, 6044, 9238]	[301, 462, 711, 925, 1397]	[331, 498, 746, 966, 1581]
Hypothesis length	21.758	21.464	21.758	21.464
Hypothesis negation	[0.579, 0.583, 0.586]	[0.574, 0.578, 0.583]	[0.579, 0.583, 0.586]	[0.574, 0.578, 0.583]
Training examples	71442	128520	71442	128520
Existing datasets	**MNLI**	**anli**	**DocNLI**	**ConTRoL-dataset**
Premise length	[10, 15, 23, 34, 46]	[14, 28, 63, 80, 95]	[57, 73, 115, 557, 1050]	[55.6, 138, 333, 996, 1147]
Hypothesis length	14.271	13.608	56.797	16.323
Hypothesis negation	[0.13, 0.141, 0.358]	[0.074, 0.069, 0.197]	[0.187, 0.202]	[0.094, 0.078, 0.107]
Training examples	392702	3233665	942314	6719

Table 4: Descriptive statistics of NLI datasets. Negation words ['no','not','never','none','nobody','nothing', 'neither','nor','cannot'] or contains "n't". Proportions are by label: Entail/Neutral/Contradict or Entail/Not entail. *About 50% of LawngNLI's hypotheses contain pivotal negation*, even though over 50% contain negation under the keyword definition (used here for comparability across datasets). See Appendix Section A.1 on dataset construction. Token lengths are [10, 25, 50, 75, 90] percentiles or an average via a RoBERTa (Liu et al., 2019) tokenizer.

LawngNLI automatic labels	Short premise		Long premise	
	All	Negation	All	Negation
Agreed-upon (gold) labels				
Accuracy	0.92	0.901	0.888	0.87
N	160	76	140	66
High-confidence agreed-upon (gold) labels				
Accuracy	0.972	0.976	0.947	0.905
N	81	39	68	31
Full assessment set				
Worker agreement	0.758	0.71	0.761	0.75
High confidence, if agreement	0.506	0.513	0.486	0.47
N	211	107	184	88

Table 5: Human assessment by two workers per example of a stratified random sample of LawngNLI's *"analysis" subset* (sequence length of long premise at most 4096). The split refers to *pivotal* negation. Provided accuracies are equally weighted averages of the accuracies by label. High-confidence labels are when both workers chose "definitely" rather than "probably" their label.

Evaluation	Long premise		Short premise	
Fine-tuning	No	Yes	No	Yes
[Entail/Neutral/Contradict. Chance=1/3]				
google_bigbird-roberta-base_anli	0.342+/-0.015	0.77+/-0.013	0.403+/-0.015	0.84+/-0.012
albert-xxlarge-v2_anli	0.501+/-0.016	0.789+/-0.013	0.551+/-0.016	0.882+/-0.01
roberta-large_anli	0.353+/-0.015	0.778+/-0.013	0.374+/-0.015	0.884+/-0.01
allenai_longformer-base-4096_anli	0.367+/-0.015	0.691+/-0.015	0.402+/-0.015	0.802+/-0.013
zlucia_custom-legalbert_anli	0.499+/-0.016	0.776+/-0.013	0.536+/-0.016	0.843+/-0.012
nlpaueb_legal-bert-base-uncased_anli	0.478+/-0.016	0.767+/-0.013	0.514+/-0.016	0.849+/-0.012
facebook_bart-large_anli	0.345+/-0.015	0.76+/-0.014	0.532+/-0.016	0.879+/-0.011
allenai_longformer-base-4096_ConTRoL-dataset	0.355+/-0.015	0.693+/-0.015	0.375+/-0.015	0.791+/-0.013
google_bigbird-roberta-base_ConTRoL-dataset	0.354+/-0.015	0.757+/-0.014	0.383+/-0.015	0.845+/-0.012
zlucia_custom-legalbert_ConTRoL-dataset	0.445+/-0.016	0.782+/-0.013	0.462+/-0.016	0.839+/-0.012
nlpaueb_legal-bert-base-uncased_ConTRoL-dataset	0.423+/-0.016	0.761+/-0.014	0.471+/-0.016	0.839+/-0.012
facebook_bart-large_ConTRoL-dataset	0.407+/-0.015	0.758+/-0.014	0.468+/-0.016	0.876+/-0.011
albert-xxlarge-v2_ConTRoL-dataset	0.434+/-0.016	0.781+/-0.013	0.478+/-0.016	0.878+/-0.011
roberta-large_ConTRoL-dataset	0.429+/-0.016	0.761+/-0.014	0.478+/-0.016	0.872+/-0.011
[Entail/Not entail. Chance=1/2]				
allenai_longformer-base-4096_DocNLI	0.5+/-0.014	0.777+/-0.011	0.496+/-0.014	0.834+/-0.01
google_bigbird-roberta-base_DocNLI	0.513+/-0.014	0.817+/-0.011	0.508+/-0.014	0.863+/-0.01
zlucia_custom-legalbert_DocNLI	0.412+/-0.013	0.822+/-0.011	0.577+/-0.013	0.857+/-0.01
nlpaueb_legal-bert-base-uncased_DocNLI	0.512+/-0.014	0.833+/-0.01	0.38+/-0.013	0.873+/-0.009
facebook_bart-large_DocNLI	0.497+/-0.014	0.641+/-0.013	0.531+/-0.014	0.874+/-0.009
albert-xxlarge-v2_DocNLI	0.637+/-0.013	0.847+/-0.01	0.421+/-0.013	0.907+/-0.008
roberta-large_DocNLI	0.448+/-0.014	0.843+/-0.01	0.412+/-0.013	0.91+/-0.008
N	3966			

Table 6: Performance of full intermediate-fine-tuned model panel: Accuracy on test set within LawngNLI's "analysis" subset (long premise at most 4096 tokens). Fine-tuning on the LawngNLI subset is on premises with the same granularity as evaluation. The error provided is the larger of the two deviations of the Clopper-Pearson (Clopper and Pearson, 1934) exact binomial 95% confidence bounds from the point estimate.