Evidence Inference 2.0: More Data, Better Models

Jay DeYoung[⋆Ψ], Eric Lehman[⋆Ψ], Ben Nye[Ψ], Iain J. Marshall[Φ], and Byron C. Wallace[Ψ]

[⋆]Equal contribution
[Ψ]Khoury College of Computer Sciences, Northeastern University
[Φ]Kings College London
{deyoung.j,lehman.e,nye.b,b.wallace}@northeastern.edu, mail@ijmarshall.com

Abstract

How do we most effectively treat a disease or condition? Ideally, we could consult a database of evidence gleaned from clinical trials to answer such questions. Unfortunately, no such database exists; clinical trial results are instead disseminated primarily via lengthy natural language articles. Perusing all such articles would be prohibitively time-consuming for healthcare practitioners; they instead tend to depend on manually compiled *systematic reviews* of medical literature to inform care.

NLP may speed this process up, and eventually facilitate immediate consult of published evidence. The *Evidence Inference* dataset (Lehman et al., 2019) was recently released to facilitate research toward this end. This task entails inferring the comparative performance of two treatments, with respect to a given outcome, from a particular article (describing a clinical trial) and identifying supporting evidence. For instance: Does this article report that *chemotherapy* performed better than *surgery* for *five-year survival rates* of operable cancers? In this paper, we collect additional annotations to expand the Evidence Inference dataset by 25%, provide stronger baseline models, systematically inspect the errors that these make, and probe dataset quality. We also release an *abstract only* (as opposed to full-texts) version of the task for rapid model prototyping. The updated corpus, documentation, and code for new baselines and evaluations are available at http://evidence-inference.ebm-nlp.com/.

1 Introduction

As reports of clinical trials continue to amass at rapid pace, staying on top of all current literature to inform evidence-based practice is next to impossible. As of 2010, about seventy clinical trial reports were published daily, on average (Bastian et al., 2010). This has risen to over one hundred thirty trials per day.[1] Motivated by the rapid growth in clinical trial publications, there now exist a plethora of tools to partially automate the systematic review task (Marshall and Wallace, 2019). However, efforts at fully integrating the PICO framework into this process have been limited (Eriksen and Frandsen, 2018). What if we could build a database of **P**articipants,[2] **I**nterventions, **C**omparisons, and **O**utcomes studied in these trials, and the findings reported concerning these? If done accurately, this would provide direct access to which treatments the evidence supports. In the near-term, such technologies may mitigate the tedious work necessary for manual synthesis.

Recent efforts in this direction include the EBM-NLP project (Nye et al., 2018), and Evidence Inference (Lehman et al., 2019), both of which comprise annotations collected on reports of Randomized Control Trials (RCTs) from PubMed.[3] Here we build upon the latter, which tasks systems with inferring findings in full-text reports of RCTs with respect to particular interventions and outcomes, and extracting evidence snippets supporting these.

We expand the Evidence Inference dataset and evaluate transformer-based models (Vaswani et al., 2017; Devlin et al., 2018) on the task. Concretely, **our contributions** are:

- We describe the collection of an additional 2,503 unique 'prompts' (see Section 2) with matched full-text articles; this is a 25% expansion of the original evidence inference dataset that we will release. We additionally have collected an *abstract-only* subset of data intended to facilitate rapid iterative design of models,

[1] See https://ijmarshall.github.io/sote/.

[2] We omit Participants in this work as we focus on the document level task of inferring study result directionality, and the Participants are inherent to the study, i.e., studies do not typically consider multiple patient populations.

[3] https://pubmed.ncbi.nlm.nih.gov/

as working over full-texts can be prohibitively time-consuming.

- We introduce and evaluate new models, achieving SOTA performance for this task.
- We ablate components of these models and characterize the types of errors that they tend to still make, pointing to potential directions for further improving models.

2 Annotation

In the *Evidence Inference* task (Lehman et al., 2019), a model is provided with a full-text article describing a randomized controlled trial (RCT) and a 'prompt' that specifies an *Intervention* (e.g., aspirin), a *Comparator* (e.g., placebo), and an *Outcome* (e.g., duration of headache). We refer to these as ICO prompts. The task then is to infer whether a given article reports that the Intervention resulted in a *significant increase*, *significant decrease*, or produced *no significant difference* in the Outcome, as compared to the Comparator.

Our annotation process largely follows that outlined in Lehman et al. (2019); we summarize this briefly here. Data collection comprises three steps: (1) prompt generation; (2) prompt and article annotation; and (3) verification. All steps are performed by Medical Doctors (MDs) hired through Upwork.[4] Annotators were divided into mutually exclusive groups performing these tasks, described below.

Combining this new data with the dataset introduced in Lehman et al. (2019) yields in total 12,616 unique prompts stemming from 3,346 unique articles, increasing the original dataset by 25%.[5] To acquire the new annotations, we hired 11 doctors: 1 for prompt generation, 6 for prompt annotation, and 4 for verification.

2.1 Prompt Generation

In this collection phase, a single doctor is asked to read an article and identify triplets of interventions, comparators, and outcomes; we refer to these as ICO prompts. Each doctor is assigned a unique article, so as to not overlap with one another. Doctors were asked to find a maximum of 5 prompts per article as a practical trade-off between the expense of exhaustive annotation and acquiring annotations over a variety of articles. This resulted in our collecting 3.77 prompts per article, on average. We asked doctors to derive at least 1 prompt from the body (rather than the abstract) of the article. A large difficulty of the task stems from the wide variety of treatments and outcomes used in the trials: 35.8% of interventions, 24.0% of comparators, and 81.6% of outcomes are unique to one another.

In addition to these ICO prompts, doctors were asked to report the relationship between the intervention and comparator with respect to the outcome, and cite what span from the article supports their reasoning. We find that 48.4% of the collected prompts can be answered using only the abstract. However, 63.0% of the evidence spans supporting judgments (provided by both the prompt generator and prompt annotator), are from outside of the abstract. Additionally, 13.6% of evidence spans cover more than one sentence in length.

2.2 Prompt Annotation

Following the guidelines presented in Lehman et al. (2019), each prompt was assigned to a single doctor. They were asked to report the difference between the specified intervention and comparator, with respect to the given outcome. In particular, options for this relationship were: "increase", "decrease", "no difference" or "invalid prompt." Annotators were also asked to mark a span of text supporting their answers: a rationale. However, unlike Lehman et al. (2019), here, annotators were not restricted via the annotation platform to only look at the abstract at first. They were free to search the article as necessary.

Because trials tend to investigate multiple interventions and measure more than one outcome, articles will usually correspond to multiple — potentially many — valid ICO prompts (with correspondingly different findings). In the data we collected, 62.9% of articles comprise at least two ICO prompts with different associated labels (for the same article).

2.3 Verification

Given both the answers and rationales of the prompt generator and prompt annotator, a third doctor — the verifier — was asked to determine the validity of both of the previous stages.[6] We estimate the accuracy of each task with respect to these verification labels. For prompt generation, answers

[4] http://upwork.com.

[5] We use the first release of the data by Lehman et al., which included 10,137 prompts. A subsequent release contained 10,113 prompts, as the authors removed prompts where the answer and rationale were produced by different doctors.

[6] The verifier can also discard low-quality or incorrect prompts.

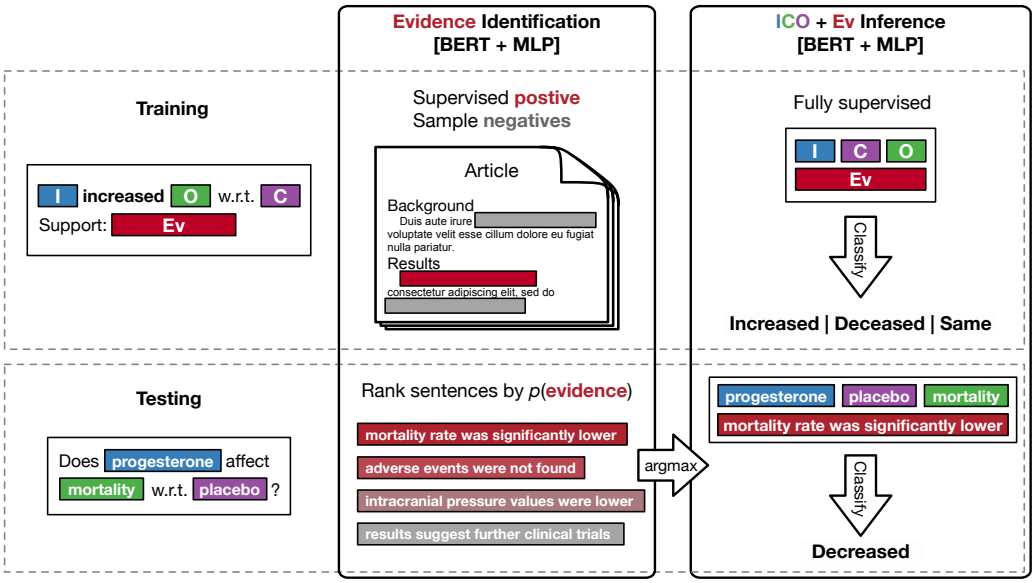

Figure 1: BERT to BERT pipeline. Evidence identification and classification stages are trained separately. The identifier is trained via negative samples against the positive instances, the classifier via only those same positive evidence spans. Decoding assigns a score to every sentence in the document, and the sentence with the highest evidence score is passed to the classifier.

were 94.0% accurate, and rationales were 96.1% accurate. For prompt annotation, the answers were 90.0% accurate, and accuracy of the rationales was 88.8%. The drop in accuracy between prompt generation answers and prompt annotation answers is likely due to confusion with respect to the scope of the intervention, comparator, and outcome.

We additionally calculated agreement statistics amongst the doctors across all stages, yielding a Krippendorf's α of $\alpha = 0.854$. In contrast, the agreement between prompt generator and annotator (excluding verifier) had a $\alpha = 0.784$.

2.4 Abstract Only Subset

We subset the articles and their content, yielding 9,680 of 24,686 annotations, or approximately 40%. This leaves 6375 prompts, 50.5% of the total.

3 Models

We consider a simple BERT-based (Devlin et al., 2018) pipeline comprising two independent models, as depicted in Figure 1. The first *identifies* evidence bearing sentences within an article for a given ICO. The second model then *classifies* the reported findings for an ICO prompt using the evidence extracted by this first model. These models place a dense layer on top of representations yielded from (Gururangan et al., 2020), [7] a variant of RoBERTa (Liu et al., 2019) pre-trained over scientific corpora,[8] followed by a Softmax.

Specifically, we first perform sentence segmentation over full-text articles using `ScispaCy` (Neumann et al., 2019). We use this segmentation to recover evidence bearing sentences. We train an evidence *identifier* by learning to discriminate between evidence bearing sentences and randomly sampled non-evidence sentences.[9] We then train an evidence *classifier* over the evidence bearing sentences to characterize the trial's finding as reporting that the Intervention *significantly decreased, did not significantly change,* or *significantly increased* the Outcome compared to the Comparator in an ICO. When making a prediction for an (ICO, document) pair we use the highest scoring evidence sentence from the identifier, feeding this to the evidence classifier for a final result. Note that the evidence classifier is conditioned on the ICO frame; we prepend the ICO embedding (from Biomed RoBERTa) to the embedding of the identified evidence snippet. Reassuringly, removing this signal degrades performance (Table 1).

For all models we fine-tuned the underlying BERT parameters. We trained all models using the Adam optimizer (Kingma and Ba, 2014) with a BERT learning rate 2e-5. We train these models for 10 epochs, keeping the best performing version on a nested held-out set with respect to

[7] An earlier version of this work used SciBERT (Beltagy et al., 2019); we preserve these results in Appendix C.

[8] We use the `[CLS]` representations.

[9] We train this via negative sampling because the vast majority of sentences are not evidence-bearing.

macro-averaged f-scores. When training the evidence identifier, we experiment with different numbers of random samples per positive instance. We used `Scikit-Learn` (Pedregosa et al., 2011) for evaluation and diagnostics, and implemented all models in `PyTorch` (Paszke et al., 2019). We additionally reproduce the end-to-end system from Lehman et al. (2019): a gated recurrent unit (Cho et al., 2014) to encode the document, attention (Bahdanau et al., 2015) conditioned on the ICO, with the resultant vector (plus the ICO) fed into an MLP for a final significance decision.

4 Experiments and Results

Our main results are reported in Table 1. We make a few key observations. First, the gains over the prior state-of-the-art model — which was not BERT based — are substantial: 20+ absolute points in F-score, even beyond what one might expect to see shifting to large pre-trained models.[10] Second, conditioning on the ICO prompt is key; failing to do so results in substantial performance drops. Finally, we seem to have reached a plateau in terms of the performance of the BERT pipeline model; adding the newly collected training data does not budge performance (evaluated on the augmented test set). This suggests that to realize stronger performance here, we perhaps need a less naive architecture that better models the domain. We next probe specific aspects of our design and training decisions.

Impact of Negative Sampling As negative sampling is a crucial part of the pipeline, we vary the number of samples and evaluate performance. We provide detailed results in Appendix A, but to summarize briefly: we find that two to four negative samples (per positive) performs the best for the end-to-end task, with little change in both AUROC and accuracy of the best fit evidence sentence. This is likely because the model needs only to maximize discriminative capability, rather than calibration.

Distribution Shift In addition to comparable Krippendorf-α values computed above, we measure the impact of the new data on pipeline performance. We compare performance of the pipeline with all data "Biomed RoBERTa (BR) Pipeline" vs. just the old data "Biomed RoBERTA (BR) BERT Pipeline 1.0" in Table 1. As performance stays relatively constant, we believe the new data to be well-aligned with the existing release. This also suggests that the performance of the current simple pipeline model may have plateaued; better performance perhaps requires inductive biases via domain knowledge or improved strategies for evidence identification.

Model	Cond?	P	R	F
BR Pipeline	✓	.784	.777	.780
BR Pipeline	✗	.513	.510	.510
BR Pipeline abs.	✓	.776	.777	.776
Baseline	✓	.526	.516	.514
Diagnostics:				
BR Pipeline 1.0	✓	.762	.764	.763
Baseline 1.0	✓	.531	.519	.520
BR ICO Only		.522	.515	.511
BR Oracle Spans	✓	.851	.853	.851
BR Oracle Sentence	✓	.845	.843	.843
BR Oracle Spans	✗	.806	.812	.808
BR Oracle Sentence	✗	.802	.795	.797
BR Oracle Spans abs.	✓	.830	.823	.824
Baseline Oracle 1.0	✓	.740	.739	.739
Baseline Oracle	✓	.760	.761	.759

Table 1: **Classification Scores**. BR Pipeline: Biomed RoBERTa BERT Pipeline. *abs*: Abstracts only. *Baseline*: model from Lehman et al. (2019). **Diagnostic models**: *Baseline* scores Lehman et al. (2019), BR Pipeline when trained using the Evidence Inference 1.0 data, BR classifier when presented with only the ICO element, an entire human selected evidence span, or a human selected evidence sentence. Full document BR models are trained with four negative samples; abstracts are trained with sixteen; Baseline oracle span results from Lehman et al. (2019). In all cases: *'Cond?'* indicates whether or not the model had access to the ICO elements; P/R/F scores are macro-averaged.

Oracle Evidence We report two types of Oracle evidence experiments - one using ground truth evidence spans "Oracle *spans*", the other using *sentences* for classification. In the former experiment, we choose an arbitrary evidence span[11] for each prompt for decoding. For the latter, we arbitrarily choose a sentence contained within a span. Both experiments are trained to use a matching classifier. We find that using a span versus a sentence causes a marginal change in score. Both diagnostics provide an upper bound on this model type, improve over the original Oracle baseline by approximately 10 points. Using Oracle evidence as opposed to a trained evidence identifier leaves an end-to-end performance gap of approximately 0.08 F1 score.

[10] To verify the impact of architecture changes, we experiment with randomly initialized and fine-tuned BERTs. We find that these perform worse than the original models in all instances and elide more detailed results.

[11] Evidence classification operates on a single sentence, but an annotator's selection is *span* based. Furthermore, the prompt annotation stage may produce different evidence spans than prompt generation.

Ev. Cls	ID Acc.	Predicted Class Sig $\ominus$	Sig $\sim$	Sig $\oplus$
Sig $\ominus$	.667	.684	.153	.163
Sig $\sim$	.674	.060	.840	.099
Sig $\oplus$	.652	.085	.107	.808

Table 2: Breakdown of the conditioned Biomed RoBERTa pipeline model mistakes and performance by evidence class. ID Acc. is the "identification accuracy", or percentage of . To the right is a confusion matrix for end-to-end predictions. 'Sig $\ominus$' indicates significantly decreased, 'Sig $\sim$' indicates no significant difference, 'Sig $\oplus$' indicates significantly increased.

Conditioning As the pipeline can optionally condition on the ICO, we ablate over both the ICO and the actual document text. We find that using the ICO alone performs about as effectively as an unconditioned end-to-end pipeline, 0.51 F1 score (Table 1). However, when fed Oracle sentences, the unconditioned pipeline performance jumps to 0.80 F1. As shown in Table 3 (Appendix A), this large decrease in score can be attributed to the model losing the ability to identify the correct evidence sentence.

Mistake Breakdown We further perform an analysis of model mistakes in Table 2. We find that the BERT-to-BERT model is somewhat better at identifying *significantly decreased* spans than it is at identifying spans for the *significantly increased* or *no significant difference evidence* classes. Spans for the *no significant difference* tend to be classified correctly, and spans for the *significantly increased* category tend to be confused in a similar pattern to the *significantly decreased class*. End-to-end mistakes are relatively balanced between all possible confusion classes.

Abstract Only Results We report a full suite of experiments over the abstracts-only subset in Appendix B. We find that the pipeline models perform similarly on the abstract-only subset; differing in score by less than .01F1. Somewhat surprisingly, we find that the abstracts oracle model falls behind the full document oracle model, perhaps due to a difference in language reporting general results vs. more detailed conclusions.

5 Conclusions and Future Work

We have introduced an expanded version of the Evidence Inference dataset. We have proposed and evaluated BERT-based models for the evidence inference task (which entails identifying snippets of evidence for particular ICO prompts in long documents and then classifying the reported finding on the basis of these), achieving state of the art results on this task.

With this expanded dataset, we hope to support further development of NLP for assisting Evidence Based Medicine. Our results demonstrate promise for the task of automatically inferring results from Randomized Control Trials, but still leave room for improvement. In our future work, we intend to jointly automate the identification of ICO triplets and inference concerning these. We are also keen to investigate whether pre-training on related scientific 'fact verification' tasks might improve performance (Wadden et al., 2020).

Acknowledgments

We thank the anonymous BioNLP reviewers.

This work was supported by the National Science Foundation, CAREER award 1750978.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Hilda Bastian, Paul Glasziou, and Iain Chalmers. 2010. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med*, 7(9):e1000326.

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mette Brandt Eriksen and Tove Faber Frandsen. 2018. The impact of patient, intervention, comparison, outcome (PICO) as a search strategy tool on literature search quality: a systematic review. *Journal of the Medical Library Association*, 106(4).

Suchin Gururangan, Ana Marasovi, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *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)*, pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Iain J. Marshall and Byron C. Wallace. 2019. Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8(1):163.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

David Wadden, Kyle Lo, Lucy Lu Wang, Shanchuan Lin, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claim. In *Association for Computational Linguistics (ACL)*.

Appendix

A Negative Sampling Results

We report negative sampling results for Biomed RoBERTa pipelines in Table 3 and Figure 2.

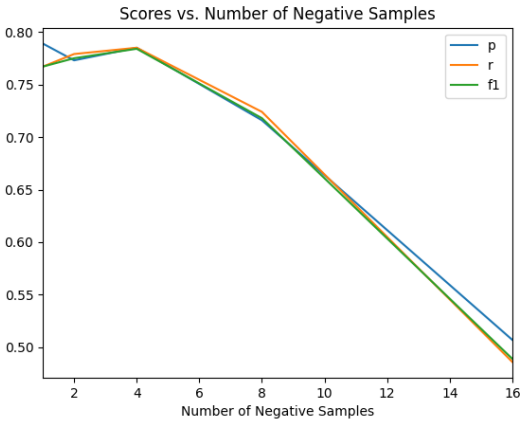

Figure 2: End to end pipeline scores for different negative sampling strategies with Biomed RoBERTa.

Neg, samples	Cond?	AUROC	Top1 Acc
1	✓	0.973	0.682
2	✓	0.972	0.700
4	✓	0.972	0.671
8	✓	0.961	0.492
16	✓	0.590	0.027
1	✗	0.915	0.236
2	✗	0.921	0.226
4	✗	0.925	0.251
8	✗	0.899	0.165
16	✗	0.508	0.015

Table 3: Evidence Inference v2.0 evidence identification validation scores varying across negative sampling strategies using Biomed RoBERTa in the pipeline.

B Abstract Only Results

We repeat the experiments described in Section 4. Our primary findings are that the abstract-only task is easier and sixteen negative samples perform better than four. Otherwise results follow a similar trend to the full-document task. We document these in Table 4, 5, 6 and Figure 3.

C SciBERT Results

We report original SciBERT results in Tables 7, 8, 9 and Figures 4, 5. Table 7 contains the Biomed RoBERTa numbers for comparison. Note that original SciBERT experiments use the evidence inference v1.0 dataset as v2.0 collection was incomplete at the time experiment configurations were determined. Biomed RoBERTa experiments use the v2.0 set for calibration. We find that Biomed RoBERTa generally performs better, with a notable exception in performance on abstracts-only Oracle span classification.

Model	Cond?	P	R	F
BR Pipeline	✓	.776	.777	.776
BR Pipeline	✗	.513	.510	.510
Diagnostics:				
ICO Only		.545	.543	.537
Oracle Spans	✓	.830	.823	.824
Oracle Sentence	✓	.845	.843	.843
Oracle Spans	✗	.814	.809	.809
Oracle Sentence	✗	.802	.795	.797

Table 4: **Classification Scores**. Biomed RoBERTa Abstract only version of Table 1. All evidence identification models trained with sixteen negative samples.

Neg. Samples	Cond?	AUROC	Top1 Acc
1	✓	0.983	0.647
2	✓	0.982	0.664
4	✓	0.981	0.680
8	✓	0.978	0.656
16	✓	0.980	0.673
1	✗	0.944	0.351
2	✗	0.953	0.373
4	✗	0.947	0.334
8	✗	0.938	0.273
16	✗	0.947	0.308

Table 5: Abstract only (v2.0) evidence identification validation scores varying across negative sampling strategies using Biomed RoBERTa.

C.1 Negative Sampling Results

We report SciBERT negative sampling results in Table 9 and Figure 4.

C.2 Abstract Only Results

We repeat the experiments described in Section 4 and report results in Tables 10, 11, 12 and Figure 5. Our primary findings are that the abstract-only task is easier and eight negative samples perform better than four. Otherwise results follow a similar trend to the full-document task.

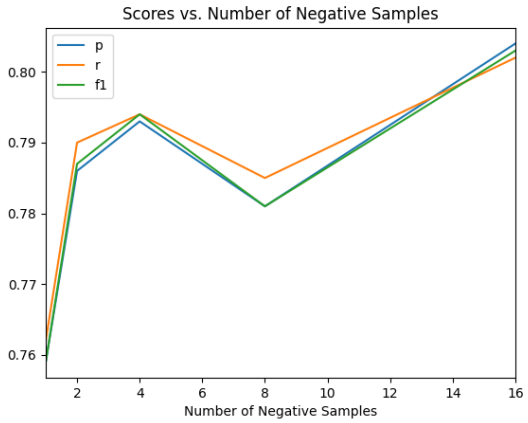

Figure 3: End to end pipeline scores on the abstract-only subset for different negative sampling strategies with Biomed RoBERTa.

		Conf. Cls		
Ev. Cls	ID Acc.	Sig $\ominus$	Sig $\sim$	Sig $\oplus$
Sig $\ominus$	.728	.761	.067	.172
Sig $\sim$	.691	.130	.802	.068
Sig $\oplus$	.573	.123	.109	.768

Table 6: Breakdown of the abstract-only conditioned Biomed RoBERTa pipeline model mistakes and performance by evidence class. ID Acc. is breakdown by final evidence truth. To the right is a confusion matrix for end-to-end predictions.

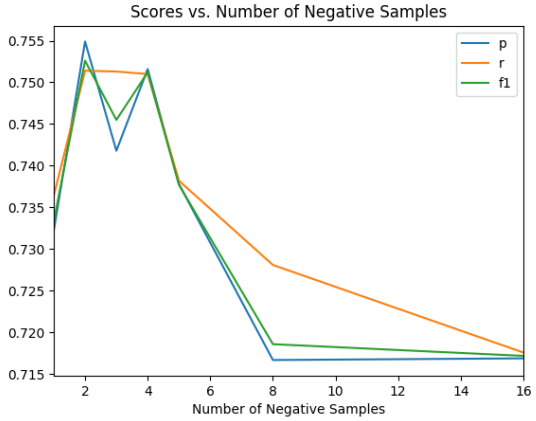

Figure 4: End to end pipeline scores for different negative sampling strategies for SciBERT.

Model	Cond?	P	R	F
BR Pipeline	✓	.784	.777	.780
SB Pipeline	✓	.750	.750	.749
BR Pipeline	✗	.513	.510	.510
SB Pipeline	✗	.489	.486	.486
BR Pipeline abs.	✓	.776	.777	.776
SB Pipeline abs.	✓	.803	.798	.799
Baseline	✓	.526	.516	.514
Diagnostics:				
BR Pipeline 1.0	✓	.762	.764	.763
SB Pipeline 1.0	✓	.749	.761	.753
Baseline 1.0	✓	.531	.519	.520
BR ICO Only		.522	.515	.511
SB ICO Only		.494	.501	.494
BR Oracle Spans	✓	.851	.853	.851
SB Oracle Spans	✓	.840	.840	.838
BR Oracle Sentence	✓	.845	.843	.843
SB Oracle Sentence	✓	.829	.830	.829
BR Oracle Spans	✗	.806	.812	.808
SB Oracle Spans	✗	.786	.789	.787
BR Oracle Sentence	✗	.802	.795	.797
SB Oracle Sentence	✗	.780	.770	.773
BR Oracle Spans abs.	✓	.830	.823	.824
SB Oracle Spans abs.	✓	.866	.862	.863
Baseline Oracle 1.0	✓	.740	.739	.739
Baseline Oracle	✓	.760	.761	.759

Table 7: Replica of Table 1 with both SciBERT and Biomed RoBERTa results. **Classification Scores**. BR Pipeline: Biomed RoBERTa BERT Pipeline, SB Pipeline: SciBERT Pipeline. *abs*: Abstracts only. *Baseline*: model from Lehman et al. (2019). **Diagnostic models**: *Baseline* scores Lehman et al. (2019), BR Pipeline when trained using the Evidence Inference 1.0 data, BR classifier when presented with only the ICO element, an entire human selected evidence span, or a human selected evidence sentence. Full document BR models are trained with four negative samples; abstracts are trained with sixteen; Baseline oracle span results from Lehman et al. (2019). In all cases: *'Cond?'* indicates whether or not the model had access to the ICO elements; P/R/F scores are macro-averaged over classes.

		Predicted Class		
Ev. Cls	ID Acc.	Sig $\ominus$	Sig $\sim$	Sig $\oplus$
Sig $\ominus$	.711	.697	.143	.160
Sig $\sim$	.643	.076	.838	.086
Sig $\oplus$	.635	.146	.141	.713

Table 8: Replica of Table 2 for SciBERT. Breakdown of the conditioned BERT pipeline model mistakes and performance by evidence class. ID Acc. is the "identification accuracy", or percentage of . To the right is a confusion matrix for end-to-end predictions. 'Sig $\ominus$' indicates significantly decreased, 'Sig $\sim$' indicates no significant difference, 'Sig $\oplus$' indicates significantly increased.

Neg. Samples	Cond?	AUROC	Top1 Acc
1	✓	.969	.663
2	✓	.959	.673
4	✓	.968	.659
8	✓	.961	.627
16	✓	.967	.593
1	✗	.894	.094
2	✗	.890	.181
4	✗	.843	.083
8	✗	.862	.170
16	✗	.403	.014

Table 9: Evidence Inference v1.0 evidence identification validation scores varying across negative sampling strategies for SciBERT.

Model	Cond?	P	R	F
BERT Pipeline	✓	.803	.798	.799
BERT Pipeline	✗	.528	.513	.510
Diagnostics:				
ICO Only		.480	.480	.479
Oracle Spans	✓	.866	.862	.863
Oracle Sentence	✓	.848	.842	.844
Oracle Spans	✗	.804	.802	.801
Oracle Sentence	✗	.817	.776	.783

Table 10: **Classification Scores**. SciBERT/Abstract only version of Table 1. All evidence identification models trained with eight negative samples.

Neg. Samples	Cond?	AUROC	Top1 Acc
1	✓	0.980	0.573
2	✓	0.978	0.596
4	✓	0.977	0.623
8	✓	0.950	0.609
16	✓	0.975	0.615
1	✗	0.946	0.340
2	✗	0.939	0.342
4	✗	0.912	0.286
8	✗	0.938	0.313
16	✗	0.940	0.282

Table 11: Abstract only (v1.0) evidence identification validation scores varying across negative sampling strategies for SciBERT.

		Conf. Cls		
Ev. Cls	ID Acc.	Sig ⊖	Sig ∼	Sig ⊕
Sig ⊖	.767	.750	.044	.206
Sig ∼	.686	.092	.816	.092
Sig ⊕	.591	.109	.064	.827

Table 12: Breakdown of the abstract-only conditioned SciBERT pipeline model mistakes and performance by evidence class. ID Acc. is breakdown by final evidence truth. To the right is a confusion matrix for end-to-end predictions.

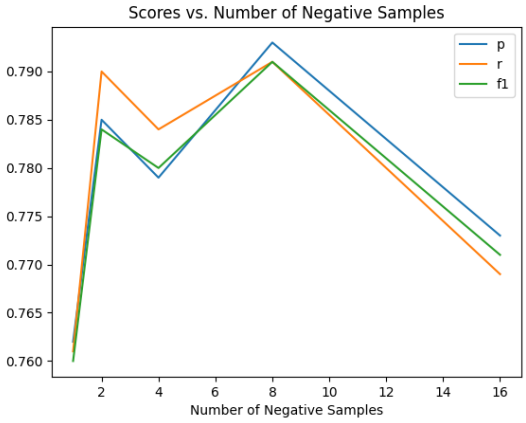

Figure 5: End to end pipeline scores on the abstract-only subset for different negative sampling strategies for SciBERT.

	Train	Dev	Test	Total
Number of prompts	10150	1238	1228	12616
Number of articles	2672	340	334	3346
Label counts (-1 / 0 / 1)	2465 / 4563 / 3122	299 / 544 / 395	295 / 516 / 417	3059 / 5623 / 3934

Table 13: Corpus statistics. Labels -1, 0, 1 indicate *significantly decreased*, *no significant difference* and *significantly increased*, respectively.