# Detecting Contradictory COVID-19 Drug Efficacy Claims from Biomedical Literature

**Daniel N. Sosa**[1] **Malavika Suresh**[2] **Christopher Potts**[3] **Russ B. Altman**[4,5]

[1]Department of Biomedical Data Science, Stanford University
[2]School of Computing, Robert Gordon University
[3]Department of Linguistics, Stanford University
[4]Department of Bioengineering, Stanford University
[5]Department of Genetics, Stanford University
{dnsosa,cgpotts,russ.altman}@stanford.edu m.suresh@rgu.ac.uk

## Abstract

The COVID-19 pandemic created a deluge of questionable and contradictory scientific claims about drug efficacy – an "infodemic" with lasting consequences for science and society. In this work, we argue that NLP models can help domain experts distill and understand the literature in this complex, high-stakes area. Our task is to automatically identify contradictory claims about COVID-19 drug efficacy. We frame this as a natural language inference problem and offer a new NLI dataset created by domain experts. The NLI framing allows us to create curricula combining existing datasets and our own. The resulting models are useful investigative tools. We provide a case study of how these models help a domain expert summarize and assess evidence concerning remdisivir and hydroxychloroquine.[1]

## 1 Introduction

The COVID-19 pandemic caused by the novel SARS-CoV-2 virus completely changed modern life. According to the World Health Organization Nov. 16, 2022, situation report, more than 6.5 million people have died as a result of this disease (World Health Organization, 2022). During times of pandemic, treatment options are limited, and developing new drug treatments is infeasible in the short-term (Wouters et al., 2020).

However, if a novel disease shares biological underpinnings with another disease for which a drug treatment already exists, a doctor may be able to repurpose that drug as a treatment for the new disease with positive therapeutic effect (Pushpakom et al., 2019). This strategy has been successful in several contexts (Corsello et al., 2017; Himmelstein et al., 2022; Al-Saleem et al., 2021) and may be the only viable strategy during an emerging pandemic.

Decisions about repurposing drug treatments are predicated on scientific knowledge. Making predictions about how to repurpose an existing drug requires understanding the target disease's mechanism. Because SARS-CoV-2 was a new virus, our knowledge of COVID-19's mechanism rapidly evolved. The biomedical literature about the virus and disease proliferated at an unprecedented rate (Ioannidis et al., 2022a,b). The need for knowledge about the virus and the bottleneck of limited peer reviewers led to many cases of circumventing typical quality control mechanisms for research. To inform their clinical practice, healthcare professionals relied on knowledge sources of lower scientific quality including early clinical reports with small sample sizes and non-peer reviewed manuscripts posted on preprint servers (Nouri et al., 2021). This deluge of rapidly changing information became an "infodemic", and it became infeasible for the average clinician to stay up-to-date with the growing literature (The Lancet Infectious Diseases, 2020).

Automated methods have great potential to help domain experts fight such an infodemic. We illustrate this potential with a case study focused on automatically detecting contradictory research claims in the COVID-19 therapeutics literature. We frame this as a natural language inference (NLI) problem: given pairs of research claims in biomedical literature, we develop models that predict whether they entail, contradict, or are neutral with respect to each other. Our models are trained on a new dataset of these claim pairs extracted from the CORD-19 dataset (Wang et al., 2020a) and annotated by domain experts. Our best models are trained on curricula (Bengio et al., 2009) of existing NLI datasets and our domain-specific one. These models are effective at the NLI task, but the ultimate test of

[1]Our COVID-19 NLI dataset and code are available at https://github.com/dnsosa/covid_lit_contra_claims



*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
*Volume 2: Short Papers*, pages 694–713
July 9-14, 2023 ©2023 Association for Computational Linguistics

their value is whether they can help domain experts. We show how these models could help a domain expert to see early on that hydroxychloroquine was an ineffective COVID-19 treatment and how the story of remdisivir was still emerging.

## 2 COVID-19 NLI Dataset

Our new COVID-19 NLI dataset consists of pairs of research claims describing COVID-19 drug treatment efficacy and safety. These claims came from the subset of the June 17, 2020 (v33) CORD-19 (Wang et al., 2020a) manuscripts containing a COVID-19-related term (e.g., "SARS-CoV-2", "2019-nCov"). Claims were extracted from the articles' full text using the LSTM approach of Achakulvisut et al. (2020). False positive research claims were manually removed.

To begin the annotation process, we inspected pairs of claims on common drugs and topics. This led us to a set of five categories: Strict Entailment, Entailment, Possible Entailment, Strict Contradiction, Contradiction, and Neutral. Our annotation guidelines were developed and refined by clinically trained annotators (nurses and a biomedical researcher) over two preliminary rounds of annotation. In Round 1, four annotators labeled 64 claim pairs (Fleiss' $\kappa = 0.83$). The team discussed this process and refined the guidelines. In Round 2, three annotators (a subset of those from Round 1) annotated 75 claim pairs (Fleiss' $\kappa = 0.84$) using the new guidelines, and then determined that they were ready to scale (Appendix A.1).

For the dataset itself, 1000 pairs of claims were sampled for annotation using three criteria: (1) both claims mention at least one of 7 treatment candidates ({"hydroxychloroquine", "chloroquine", "tocilizumab", "remdesivir", "vitamin D", "lopinavir", "dexamethasone"}), (2) high similarity between the claim's embedding and the embedding for a word in a predefined topic list ({"mortality", "effective treatment", "toxicity"}), using uSIF embeddings (Ethayarajh, 2018), and (3) non-zero polarities of equal or opposite sign using VADER (Hutto and Gilbert, 2014). Appendix A.3 provides further details.

Each annotation requires a large time investment from the annotator and draws heavily on their domain expertise, so each example was annotated by a single annotator, with the inter-annotator agreement rounds and guidelines serving to ensure consistency across the dataset.

| Dataset | # Entail | # Neutral | # Contra |
|---|---|---|---|
| Full | 266 | 610 | 118 |
| D-Train | 129 | 265 | 40 |
| D-Val | 41 | 75 | 41 |
| D-Test | 66 | 100 | 21 |

Table 1: COVID-19 NLI distribution of annotated claim pairs by class for the full dataset and the disjoint (D-*) dataset splits.

Because some claims are present in multiple claim pairs, we selected a subset of pairs such that no claim is present in more than one train, validation, or test split to prevent test-set leakage. From the network of claim pairs (claims are nodes, and co-occurrences in an annotated pair are edges), we selected 3 disjoint subnetworks to comprise the train, validation, and test splits. The resulting dataset contains 778 total claim pairs. Dataset distributions are found in Table 1.

## 3 Model Development

Our goal is to develop a model to help domain experts find and adjudicate contradictory claims in the COVID-19 literature. We explored a wide range of techniques for developing the best model given our available data. The Appendices provide a full overview of all these experiments and comparisons. Here, we provide a high-level summary.

**Pretrained Parameters** All our models begin with pretrained parameters created using the general architecture of BERT (Devlin et al., 2019). Five pre-trained BERT models were evaluated for further fine-tuning: PubMedBERT (Gu et al., 2021), SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2020), BioClinBERT (Alsentzer et al., 2019), and RoBERTa (Liu et al., 2019). We found that PubMedBERT was the best for our task across all fine-tuning regimes (Appendix D).

**Fine-tuning Curricula** For fine-tuning these parameters, we use MultiNLI (Williams et al., 2018), MedNLI (Romanov and Shivade, 2018), ManConCorpus (Alamri and Stevenson, 2016), and our new COVID-19 NLI Dataset (with our six labels collapsed to three as in the other datasets). We found that the best models were achieved with a curriculum that arranged these in the order we gave above. This is intuitively an arrangement from most general to most domain-specific, which aligns with existing results and intuitions for curriculum learn-

| Model | Curriculum | F1 | Contra. Recall |
|---|---|---|---|
| PubMedBERT | Forward | **0.690** | **0.571** |
| | Reverse | 0.428 | 0.381 |
| | Shuffled | 0.523 | 0.416 |
| RoBERTa | Forward | 0.544 | 0.429 |
| | Reverse | 0.411 | 0.476 |
| | Shuffled | 0.239 | 0.119 |
| PubMedBERT Hyp. only | Forward | 0.485 | 0.190 |
| RoBERTa Hyp. only | Forward | 0.433 | 0.095 |

Table 2: Core results. Figure 6 and Table 4 expand these results to include a number of other baselines, most of which perform near chance. Metrics for the shuffled category are averages of the 4 shuffled curricula.

ing (Bengio et al., 2009; Xu et al., 2020; Nagatsuka et al., 2021). For detailed descriptions of these datasets, the range of curricula we explored, and our procedures for hyperparameter tuning, we refer to Appendices B, C, and E, respectively.

**Results** To contextualize our results on this hard, novel task, we evaluated a number of baselines using sparse feature representations and simple similarity calculations, as well as hypothesis-only variants of these models and our BERT models. These baselines are described in Appendix F.

Table 2 summarizes our results. We report F1 scores as well as Contradictions Recall, an important category for our case study. The best performance is achieved by the PubMedBERT model trained with the forward curriculum where fine-tuning takes place from general domain to complex, in-domain datasets. This setting outperforms baselines and alternative curricula by a large margin.

## 4 Case Study: Wading Through the Sea of Drug Treatment Literature

The value of our model lies in its potential to help domain experts tackle an infodemic. We used the model to understand the state of knowledge about the efficacy and mechanism of two controversial treatments, hydroxychlorouqine and remdesivir, from the perspective of June 2020.

We first extracted all claims identified from COVID-19 manuscripts concerning a drug treatment, using the same procedure as for our COVID NLI dataset (Section 2), and we filtered that set to pairs of claims that were (1) sufficiently similar (uSIF similarity > 0.5) and (2) both mentioned remdesivir or hydroxychloroquine. We sampled pairs from 50 papers yielding 5,336 total pairs. We then used our best model to make predictions about all these pairs resulting in 322 predicted contradictions. We ranked these by the model's predicted probability of this class, and we inspected the highest probability predictions.

For remdesivir, one claim of limited efficacy from an clinical trial of 233 participants yielded several predicted contradictions:

(1) Remdesivir did not result in significant reductions in SARS-CoV-2 RNA loads or detectability in upper respiratory tract or sputum specimens in this study despite showing strong antiviral effects in preclinical models of infection with coronaviruses (Wang et al., 2020b).

Nineteen unique papers contained a claim that was predicted to contradict this claim – already a striking pattern that might have taken a researcher days to discover by hand by probing full-text articles.

The specific claims that contradict our core claim are illuminating. One reads,

(2) The present study reveals that remdesivir has the highest potential in binding and therefore competitively inhibiting RDRP of SARS-CoV-2, among all known RDRP inhibitors (Choudhury et al., 2021),

indicating strong chemical and pharmacodynamic reasoning supporting a mechanism of action for remdesivir. A second claim describes,

(3) Remdesivir treatment in rhesus macaques infected with SARS-CoV-2 was highly effective in reducing clinical disease and damage to the lungs (Williamson et al., 2020),

surfacing particularly strong pre-clinical evidence. From another ongoing clinical trial including 1,064 patients, authors note:

(4) Preliminary results of this trial suggest that a 10-day course of remdesivir was superior to placebo in the treatment of hospitalized patients with COVID-19. (Beigel et al., 2020)

Overall, we are quickly able to glean how evidence supporting the remdesivir hypothesis was strong from a variety of pre-clinical studies in vastly different settings in 2020. Our original negative claim (1) presents real evidence against the drug. Still, though, the clinical picture was not yet clear, suggesting the need for further clinical investigation or better striation of populations or therapeutic windows for seeing efficacy.

For hydroxychloroquine, one of the earliest drugs considered, a different picture emerges. We focus in on a claim from a medRxiv preprint (5):

(5) In summary, this retrospective study demonstrates that hydroxychloroquine application is associated with a decreased risk of death in critically ill COVID-19 patients without obvious toxicity and its mechanisms of action is probably mediated through its inhibition of inflammatory cytokine storm on top of its ability in inhibiting viral replication. (Yu et al., 2020)

From its predicted contradictions, we immediately identified two clinical studies:

(6) Overall, these data do not support the addition of hydroxychloroquine to the current standard of care in patients with persistent mild to moderate COVID-19 for eliminating the virus. (Tang et al., 2020)

(7) Although a marginal possible benefit from prophylaxis in a more at-risk group cannot be ruled out, the potential risks that are associated with hydroxychloroquine may also be increased in more at-risk populations, and this may essentially negate any benefits that were not shown in this large trial involving younger, healthier participants. (Boulware et al., 2020)

These claims reflect the challenging language typical for the domain including hedging, multiple clauses, important context qualifiers (subpopulations and adverse events), and positive and negative sentiments. From these surfaced contradictions, we find evidence of the drug's inefficacy in mild and moderate cases and are led to discover the early observations of cardiac arrest being associated with hydroxychloroquine treatment. Again, discovering these claims *de novo* is difficult given the size of the corpus of COVID-19 literature. Our NLI model greatly speeds up the process and allows domain experts to home in directly on relevant evidence.

## 5 Stakeholders

There are several biomedical stakeholders who would benefit from models like ours.

**Epidemiologists** Epidemiologists survey public health data to inform policy decisions in collaboration with authoritative bodies like the NIH and WHO. Their recommendations must be conservative, so surfacing results that dispute claims of drug efficacy is critical. Their gold standard resource for aggregating evidence is the meta-analysis, but in the early stages of the pandemic, large randomized controlled trials (RCTs) had not completed, and review articles quickly became outdated.

**FDA Regulators** Regulators too need to make conservative recommendations, as FDA approval signals to clinicians that a treatment is standard-of-care. Surfacing contradictory claims of drug efficacy and safety is essential (Cassidy et al., 2020).

**Researchers** By identifying areas of scientific uncertainty via contradictory evidence at all stages of the pipeline (*in silico*, *in vitro*, *in vivo*, clinical), researchers could have more quickly identified fruitful areas of investigation (Sosa et al., 2021).

**Drug Manufacturers** Manufacturers of repurposing candidates were incentivized to understand in what settings their drug seemed to be effective and by what mechanism. For claims of inefficacy, they were interested in surfacing any mitigating factors qualifying these claims or motivating follow-up analyses.

We note that these models are not intended as the sole source of decision making in clinical or epidemiological settings. To be clinically translatable, further work would need to be conducted on assessing the quality of research claims by relying on contextual information including research setting, demographics, size, and level of evidence. Rather, this work is intended to augment a manual curator's capacity to distill and synthesize a large corpus of literature. This allows trained researchers to use their judgment and conduct last-mile diligence of surfaced research contradictions or corroborations, which will be beneficial to these stakeholders downstream.

## 6 Discussion and Conclusion

In settings where the scale of literature is insurmountable for human readers, as is the case during a pandemic, automated curatorial assistants can be transformative (Lever and Altman, 2021). During COVID-19, meta-analyses and review articles, which are written to synthesize a large body of literature, could not be comprehensive or quickly became outdated. In some cases, it was necessary to create meta-meta-analyses involving hundreds of papers (Chivese et al., 2021).

Our work shows the value of integrating NLP into the domain of meta-science, embracing all the complexities of biomedical research as it naturally exists in literature. We presented an NLI framing for identifying contradictory or corroborating research claims in the challenging domain of COVID-19 drug efficacy. We created a new dataset and designed curricula for optimizing language model fine-tuning for the task. To illustrate the potential of our model, we showed that we were quickly able to distill the state of knowledge about hydroxychlorouqine and remdesivir efficacy as of June 2020, arriving at conclusions that are extremely well-supported in 2022.

Identifying where science is inconsistent is necessary for understanding the current state of human knowledge and reveals frontiers for further research. Significant contradictions can often be found buried in biomedical articles; surfacing these instances nearly as quickly as research is publicly disseminated can generate leads that researchers and curators should pursue. Beyond facilitating search and discovery, our method can help estimate confidence in the consensus of facts in science when creating general knowledge representations (Sosa and Altman, 2022) for downstream applications like predicting novel drug repurposing opportunities *in silico* (Sosa et al., 2020).

## Limitations

We identify three limitations to our approach. First, parsing research claims and automatically classifying a sentence's purpose (its meta-discourse) are not solved problems. It is more prudent to surface novel claims supported by original research than an author's allusion to other research as background context. Second, the domain of biomedical scientific text is complicated by wordy prose, hedging, and long-distance anaphora. These aspects make natural language understanding challenging and present implementational challenges for tokenization, including truncating long sentences and extracting meaning from out-of-vocabulary tokens. Third, commonsense reasoning for detecting contradictions in biomedical text requires expert background knowledge and a working definition of when contexts are sufficiently aligned such that two claims are called contradictory, which may differ depending on the use case. We believe that context sensitivity and interpretability analysis of LLMs for NLI in challenging domains like this using attention mechanisms or frameworks such as maieutic prompting (Jung et al., 2022) are particularly fruitful research directions.

## Ethics Statement

COVID research has been misinterpreted or selectively promoted leading to disinformation muddling public understanding of COVID-19 science. Any research in this space is at risk of being misapplied, and models like ours in principle could be used to distort rather than clarify the current state of research, especially by cherry picking results that fit a particular world view.

Creating a method for surfacing contradictory claims in science may also create unwanted incentives for researchers. For instance, if writing simpler and more polar claims causes our NLI model to include these claims in contradictory pairs, researchers may choose to write in such a way as to make their results more sensational, discoverable, and desirable for publishing (Ioannidis and Trikalinos, 2005). Unwanted bias may be incurred from cultural norms around how much to hedge research claims. A second important caveat is that claims surfaced with this model should be given proper due diligence. This model makes no assumptions about the quality of the underlying research and may give visibility to low-quality manuscripts. Diligence should always be maintained concerning the context, scope, relevancy, and timeliness of the research being surfaced, and our model should only serve as an initial exploratory aid.

## References


Titipat Achakulvisut, Chandra Bhagavatula, Daniel Acuna, and Konrad Kording. 2020. Claim extraction in biomedical publications using deep discourse model and transfer learning. *arXiv:1907.00962*.

Jacob Al-Saleem, Roger Granet, Srinivasan Ramakrishnan, Natalie A. Ciancetta, Catherine Saveson, Chris Gessner, and Qiongqiong Zhou. 2021. Knowledge graph-based approaches to drug repurposing for COVID-19. *Journal of Chemical Information and Modeling*, 61(8):4058–4067.

Abdulaziz Alamri and Mark Stevenson. 2016. A corpus of potentially contradictory research claims from cardiovascular research abstracts. *Journal of Biomedical Semantics*, 7.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Antonio Gonçalves, Julie Bertrand, Ruian Ke, Emmanuelle Comets, Xavier de Lamballerie, Denis Malvy, Andrés Pizzorno, Olivier Terrier, Manuel Rosa Calatrava, France Mentré, Patrick Smith, Alan S Perelson, and Jérémie Guedj. 2020. Timing of antiviral treatment initiation is critical to reduce SARS-Cov-2 viral load. *medRxiv*, page 2020.04.04.20047886.

John H. Beigel, Kay M. Tomashek, Lori E. Dodd, Aneesh K. Mehta, Barry S. Zingman, Andre C. Kalil, Elizabeth Hohmann, Helen Y. Chu, Annie Luetkemeyer, Susan Kline, Diego Lopez de Castilla, Robert W. Finberg, Kerry Dierberg, Victor Tapson, Lanny Hsieh, Thomas F. Patterson, Roger Paredes, Daniel A. Sweeney, William R. Short, Giota Touloumi, David Chien Lye, Norio Ohmagari, Myoung-don Oh, Guillermo M. Ruiz-Palacios, Thomas Benfield, Gerd Fätkenheuer, Mark G. Kortepeter, Robert L. Atmar, C. Buddy Creech, Jens Lundgren, Abdel G. Babiker, Sarah Pett, James D. Neaton, Timothy H. Burgess, Tyler Bonnett, Michelle Green, Mat Makowski, Anu Osinusi, Seema Nayak, and H. Clifford Lane. 2020. Remdesivir for the treatment of COVID-19: a final report. *New England Journal of Medicine*, 383(19):1813–1826.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 41–48, New York, NY, USA. Association for Computing Machinery.

Oliver Borchers. 2019. Fast sentence embeddings. `https://github.com/oborchers/Fast_Sentence_Embeddings`.

David R. Boulware, Matthew F. Pullen, Ananta S. Bangdiwala, Katelyn A. Pastick, Sarah M. Lofgren, Elizabeth C. Okafor, Caleb P. Skipper, Alanna A. Nascene, Melanie R. Nicol, Mahsa Abassi, Nicole W. Engen, Matthew P. Cheng, Derek LaBar, Sylvain A. Lother, Lauren J. MacKenzie, Glen Drobot, Nicole Marten, Ryan Zarychanski, Lauren E. Kelly, Ilan S. Schwartz, Emily G. McDonald, Radha Rajasingham, Todd C. Lee, and Kathy H. Hullsiek. 2020. A randomized trial of hydroxychloroquine as postexposure prophylaxis for COVID-19. *New England Journal of Medicine*, 383(6):517–525.

Christine Cassidy, Danielle Dever, Laura Stanbery, Gerald Edelman, Lance Dworkin, and John Nemunaitis. 2020. FDA efficiency for approval process of COVID-19 therapeutics. *Infectious Agents and Cancer*, 15(1):73.

Tawanda Chivese, Omran A.H. Musa, George Hindy, Noor Al-Wattary, Saif Badran, Nada Soliman, Ahmed T.M. Aboughalia, Joshua T. Matizanadzo, Mohamed M. Emara, Lukman Thalib, and Suhail A.R. Doi. 2021. Efficacy of chloroquine and hydroxychloroquine in treating COVID-19 infection: A meta-review of systematic reviews and an updated meta-analysis. *Travel Medicine and Infectious Disease*, 43:102135.

Shuvasish Choudhury, Debojyoti Moulick, Purbajyoti Saikia, and Muhammed Khairujjaman Mazumder. 2021. Evaluating the potential of different inhibitors on RNA-dependent RNA polymerase of severe acute respiratory syndrome coronavirus 2: A molecular modeling approach. *Medical Journal Armed Forces India*, 77:S373–S378.

Ka-Tim Choy, Alvina Yin-Lam Wong, Prathanporn Kaewpreedee, Sin Fun Sia, Dongdong Chen, Kenrie Pui Yan Hui, Daniel Ka Wing Chu, Michael Chi Wai Chan, Peter Pak-Hang Cheung, Xuhui Huang, Malik Peiris, and Hui-Ling Yen. 2020. Remdesivir, lopinavir, emetine, and homoharringtonine inhibit SARS-CoV-2 replication in vitro. *Antiviral Research*, 178:104786.

Steven M. Corsello, Joshua A. Bittker, Zihan Liu, Joshua Gould, Patrick McCarren, Jodi E. Hirschman, Stephen E. Johnston, Anita Vrcic, Bang Wong, Mariya Khan, Jacob Asiedu, Rajiv Narayan, Christopher C. Mader, Aravind Subramanian, and Todd R. Golub. 2017. The drug repurposing hub: A next-generation drug library and information resource. *Nature Medicine*, 23(4):405–408.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh. 2018. Unsupervised random walk sentence embeddings: A strong but simple baseline. In *Proceedings of the Third Workshop on Representation Learning for NLP*, pages 91–100, Melbourne, Australia. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):2:1–2:23.

Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. 2022. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726.

Oliver James Hulme, Eric-Jan Wagenmakers, Per Damkier, Christopher Fugl Madelung, Hartwig Roman Siebner, Jannik Helweg-Larsen, Quentin F. Gronau, Thomas Lars Benfield, and Kristoffer Hougaard Madsen. 2021. A Bayesian reanalysis of the effects of hydroxychloroquine and azithromycin on viral carriage in patients with COVID-19. *PLOS ONE*, 16(2):e0245048. Publisher: Public Library of Science.

CJ Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

John P. A. Ioannidis, Eran Bendavid, Maia Salholz-Hillel, Kevin W. Boyack, and Jeroen Baas. 2022a. Massive covidization of research citations and the citation elite. *Proceedings of the National Academy of Sciences*, 119(28):e2204074119.

John P. A. Ioannidis, Maia Salholz-Hillel, Kevin W. Boyack, and Jeroen Baas. 2022b. The rapid, massive growth of COVID-19 authors in the scientific literature. *Royal Society Open Science*, 8(9):210389.

John P. A. Ioannidis and Thomas A. Trikalinos. 2005. Early extreme contradictory estimates may appear in published research: The proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, 58(6):543–549.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1279, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jake Lever and Russ B. Altman. 2021. Analyzing the vast coronavirus literature with CoronaCentral. *Proceedings of the National Academy of Sciences*, 118(23):e2100766118.

Yue-hua Li, Cheng-hui Zhou, Han-jun Pei, Xian-liang Zhou, Li-huan Li, Yong-jian Wu, and Ru-tai Hui. 2013. Fish consumption and incidence of heart failure: A meta-analysis of prospective cohort studies. *Chinese Medical Journal*, 126(5):942–948.

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. *arXiv:1907.11692*.

Koichi Nagatsuka, Clifford Broni-Bediako, and Masayasu Atsumi. 2021. Pre-training a BERT with curriculum learning by increasing block-size of input text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 989–996.

Shayan N. Nouri, Yosef A. Cohen, Mahesh V. Madhavan, Piotr J. Slomka, Ami E. Iskandrian, and Andrew J. Einstein. 2021. Preprint manuscripts and servers in the era of coronavirus disease 2019. *Journal of Evaluation in Clinical Practice*, 27(1):16–21.

Sudeep Pushpakom, Francesco Iorio, Patrick A. Eyers, K. Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, Tim Guilliams, Joanna Latimer, Christine McNamee, Alan Norris, Philippe Sanseau, David Cavalla, and Munir Pirmohamed. 2019. Drug repurposing: Progress, challenges and recommendations. *Nature Reviews Drug Discovery*, 18(1):41–58.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Connie Schardt, Martha B. Adams, Thomas Owens, Sheri Keitz, and Paul Fontelo. 2007. Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Medical Informatics and Decision Making*, 7(1):16.

Daniel N Sosa and Russ B Altman. 2022. Contexts and contradictions: A roadmap for computational drug repurposing with knowledge inference. *Briefings in Bioinformatics*, 23(4):bbac268.

Daniel N. Sosa, Binbin Chen, Amit Kaushal, Adam Lavertu, Jake Lever, Stefano Rensi, and Russ Altman. 2021. Repurposing biomedical informaticians for COVID-19. *Journal of Biomedical Informatics*, 115:103673.

Daniel N. Sosa, Alexander Derry, Margaret Guo, Eric Wei, Connor Brinton, and Russ B. Altman. 2020. A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases. *Pacific Symposium on Biocomputing*, 25:463–474.

Wei Tang, Zhujun Cao, Mingfeng Han, Zhengyan Wang, Junwen Chen, Wenjin Sun, Yaojie Wu, Wei Xiao, Shengyong Liu, Erzhen Chen, Wei Chen, Xiongbiao Wang, Jiuyong Yang, Jun Lin, Qingxia Zhao, Youqin Yan, Zhibin Xie, Dan Li, Yaofeng Yang, Leshan Liu, Jieming Qu, Guang Ning, Guochao Shi, and Qing Xie. 2020. Hydroxychloroquine in patients with mainly mild to moderate coronavirus disease 2019: Open label, randomised controlled trial. *BMJ*, 369:m1849.

The Lancet Infectious Diseases. 2020. The COVID-19 infodemic. *The Lancet Infectious Diseases*, 20(8):875.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020a. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics.

Yeming Wang, Dingyu Zhang, Guanhua Du, Ronghui Du, Jianping Zhao, Yang Jin, Shouzhi Fu, Ling Gao, Zhenshun Cheng, Qiaofa Lu, Yi Hu, Guangwei Luo, Ke Wang, Yang Lu, Huadong Li, Shuzhen Wang, Shunan Ruan, Chengqing Yang, Chunlin Mei, Yi Wang, Dan Ding, Feng Wu, Xin Tang, Xianzhi Ye, Yingchun Ye, Bing Liu, Jie Yang, Wen Yin, Aili Wang, Guohui Fan, Fei Zhou, Zhibo Liu, Xiaoying Gu, Jiuyang Xu, Lianhan Shang, Yi Zhang, Lianjun Cao, Tingting Guo, Yan Wan, Hong Qin, Yushen Jiang, Thomas Jaki, Frederick G. Hayden, Peter W. Horby, Bin Cao, and Chen Wang. 2020b. Remdesivir in adults with severe COVID-19: A randomised, double-blind, placebo-controlled, multicentre trial. *Lancet (London, England)*, 395(10236):1569–1578.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Brandi N. Williamson, Friederike Feldmann, Benjamin Schwarz, Kimberly Meade-White, Danielle P. Porter, Jonathan Schulz, Neeltje van Doremalen, Ian Leighton, Claude Kwe Yinda, Lizzette Pérez-Pérez, Atsushi Okumura, Jamie Lovaglio, Patrick W. Hanley, Greg Saturday, Catharine M. Bosio, Sarah Anzick, Kent Barbian, Tomas Cihlar, Craig Martens, Dana P. Scott, Vincent J. Munster, and Emmie de Wit. 2020. Clinical benefit of remdesivir in rhesus macaques infected with SARS-CoV-2. *Nature*, 585(7824):273–276.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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. Association for Computational Linguistics.

World Health Organization. 2022. Weekly epidemiological update on COVID-19 - 16 November 2022. WHO report.

Olivier J. Wouters, Martin McKee, and Jeroen Luyten. 2020. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *JAMA*, 323(9):844–853.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104. Association for Computational Linguistics.

Bo Yu, Chenze Li, Peng Chen, Ning Zhou, Luyun Wang, Jia Li, Hualiang Jiang, and Dao Wen Wang. 2020. Hydroxychloroquine application is associated with a decreased mortality in critically ill patients with COVID-19. *medRxiv*.

## Supplementary Materials

## A Further Details about the COVID-19 NLI Dataset

In this appendix we provide additional details about the creation of the COVID-19 NLI dataset. Our annotators are experts in the domain having trained as healthcare providers (nursing) and annotation. The research annotator is a specialist in the biomedical domain with background in molecular biology and computer science. Annotators have also provided span annotations in several cases of drug mention, polarity, context, and expressions of uncertainty to aid in the annotation task. We plan to release the dataset under a Creative Commons Attribution 4.0 International license.[2]

### A.1 Inter-Annotator Analysis

Two rounds of inter-annotator analysis were conducted to converge on a set of annotation guidelines for scaling and to measure consistency between multiple annotators. In the first round four annotators (three clinical annotators, one researcher) were presented with 64 pairs of extracted research claims and an initial set of annotation guidelines. Classification was conducted across five classes including a Strict Entailment and Strict Contradiction class indicating two claims were entailing or contradicting in a strict logical sense as opposed to a common-reasoning sense. Global Fleiss' $\kappa$ for this round was 0.83. For the second round, three annotators (two clinical annotators, one researcher) annotated 75 claim pairs with updated guidelines and achieved similar consistency at $\kappa = 0.84$. Further minor modifications were made to the annotation guidelines resulting in the final guidelines used for the scaling round (Table 3).

| Criteria | Annotation |
|---|---|
| All drugs, context, and sentiment match | STRICT ENTAILMENT |
| At least one drug matches, the sentiment is the same but the context is at least similar | ENTAILMENT |
| All drugs and context match but the sentiment is opposing | STRICT CONTRADICTION |
| At least one drug matches, the sentiment is opposing but the context is at least similar | CONTRADICTION |
| The context or sentiment statement cannot be compared | NEUTRAL |
| There is no mention of a drug OR none of the drugs match | NEUTRAL |
| One claim contains both a POSITIVE and a NEGATIVE statement and the other claim contains a POSITIVE or NEGATIVE statement | CONTRADICTION |
| One claim is POSITIVE or NEGATIVE statement and the other is EXPRESSION_OF_UNCERTAINTY | NEUTRAL |
| Both claims are EXPRESSION_OF_UNCERTAINTY | ENTAILMENT |

Table 3: Annotation guidelines for the COVID-19 NLI dataset.

[2] https://creativecommons.org/licenses/by/4.0/

### A.2 Qualitative Error Analysis

Two challenges facing annotators during inter-annotator analysis rounds were making judgments about uncertainty and context. Research claims may have important meta-discourse cues describing speculation, hedging, or prior knowledge. For example, the statement:

(8) Randomized controlled trials are currently underway and will be critical in resolving this uncertainty as to whether [hydroxychloroquine] and [azithromycin] are effective as a treatment for COVID-19 (Hulme et al., 2021),

created discrepancy among annotators where one annotator indicated that "uncertainty as to whether [hydroxychloroquine] and [azithromycin] are effective as a treatment" is a negative statement and another indicated that "are effective as a treatment for COVID-19" was a positive statement. This led to different conclusions about efficacy of hydroxychloroquine, as the authors are describing the uncertainty in the field as background knowledge without staking a claim themselves. This motivated the creation of a span annotation, EXPRESSION_OF_UNCERTAINTY, and the criterion that when one of the claims contains this type of span, the pair is called Neutral.

For two claims to be considered comparable, they need to have sufficient contextual overlap. As an example, in the pair

(9) Remdesivir, lopinavir, emetine, and homoharringtonine inhibit SARS-CoV-2 replication *in vitro* (Choy et al., 2020)

and

(10) Overall our results emphasize that the PK/PD properties of lopinavir/ritonavir, IFN-$\beta$-1a and hydroxychloroquine make them unlikely to have a dramatic impact on viral load kinetics in the nasopharynx if they are administered after symptom onset (Antonio Gonçalves et al., 2020),

the key contexts are "*in vitro*"and "viral load kinetics in the nasopharynx". The first indicates experimental results in a controlled lab setting whereas the second indicates data collected from the noses of live patients. The decision about whether or not these contexts are sufficiently similar to decide that these claims can be compared requires the judgment of annotators, paralleling how research builds from different levels of evidence to create a grander picture about drug mechanism and efficacy. Because science is not predicated on hard and fast rules as such, annotator judgment was not always consistent.

### A.3 Preparing Claims for Annotation

For this work, resources were available for our team of highly skilled annotators to label 1000 pairs of claims. Sampling pairs of research claims at random from all extracted claims would yield pairs that are predominantly Neutral to one another. Thus, we biased the sampling procedure using heuristics for improving the balance of pairs across the three classes for annotators. The intuition behind the heuristic procedure is that two claims describing at least one drug in common and concerning a common topic may be an entailing pair if they have the same overall polarity or a contradictory pair if they have opposing polarity. Many annotated pairs were still expected to be neutral despite the biasing procedure. This was borne out by the annotated data distribution.

We considered three topics, $t \in T = \{$ "mortality", "effective treatment", "toxicity"$\}$ and seven drugs, $d \in D = \{$"hydroxychloroquine", "chloroquine", "tocilizumab", "remdesivir", "vitamin D", "lopinavir", "dexamethasone"$\}$. For each pair, $(t, d)$, the following procedure (Algorithm 1) was used to generate candidate claim pairs from the set of true research claims, $C$. Additionally given *pol*(.), a function for calculating the polarity of a claim; $k$, the number of claims to sample that are relevant to a drug and topic and have a given polarity (positive or negative); and $N$, the total number of pairs to subsample, we define our heuristic algorithm for generating candidate non-trivial pairs in Algorithm 1.

**Algorithm 1** Heuristic sampler for generating candidate non-trivial pairs

**Input:** Topic set $T$, drug set $D$, claim set $C$, polarity function $pol(.) : c \rightarrow [-1, 1]$, drug topic claim sample size $k$, total subsample size $N$

**Output:** Set of $N$ claim pairs $P_N$ concerning a common drug and topic and non-neutral predicted polarity

1: $P \leftarrow \emptyset$
2: **for** $(d, t) \in D \times T$ **do**
3: Retrieve claims $C_d := \{c \in C : d$ is a substring of $c\}$
4: Define $C_{d,t,k,pos}$ := top $k$ claims $c$ relevant to $t$ from $C_d$ s.t. $pol(c) > 0$
5: Define $C_{d,t,k,neg}$ := top $k$ claims $c$ relevant to $t$ from $C_d$ s.t. $pol(c) < 0$
6: Enumerate all combinations of claim pairs, $P_{d,t,2k}$, from claims in set $C_{d,t,k,pos} \cup C_{d,t,k,neg}$
7: Remove copy claim pairs, $P_{d,t,2k} \leftarrow P_{d,t,2k} \setminus \{(c_1, c_2) \in P_{d,t,2k} : c_1 = c_2\}$
8: $P \leftarrow P \cup P_{d,t,2k}$
9: **end for**
10: Sample $N$ pairs uniformly from $P$, $P_N$

We set $k = 7$ and $N = 1000$. To evaluate claim relevancy (lines 4 and 5), we calculate the cosine similarity between an embedding of the topic and sentence embeddings of claims using uSIF (Ethayarajh, 2018). Polarity, $pol(.)$, is calculated using Vader scores (Hutto and Gilbert, 2014).

# B Curriculum Datasets

We included four datasets for fine-tuning our language models, which comprise general language and multiple biomedically-focused domains. All our datasets use the labels Entailment, Contradiction, and Neural. For our COVID-19 NLI dataset, we collapse Strict Entailment with Entailment and Strict Contradiction with Contradiction.

## B.1 MultiNLI

MultiNLI is an NLI dataset consisting of 433k premise-hypothesis pairs taken from 5 general domains (Williams et al., 2018). To create the dataset, annotators were shown a premise and were asked to provide hypothesis statements that were entailed by, contradicted by, or were neutral to the prompt premise. In this work, we used the *matched* validation set for evaluation, which we split into two equal sized validation and test sets. The licensing situation for MultiNLI is somewhat complex (see Williams et al. 2018, section 2.2), but the dataset is widely used in the research community.

## B.2 MedNLI

MedNLI is an NLI dataset consisting of 14k premise-hypothesis pairs where premises are extracted from doctor's notes in electronic medical records (Romanov and Shivade, 2018). The annotation task for generating premise-hypothesis pairs was analogous to that for MultiNLI. As far as we know, MedNLI does not have an associated license, but it is widely used in the research community.[3]

## B.3 ManConCorpus

ManConCorpus is a dataset of research claims taken from biomedical systematic reviews (Alamri and Stevenson, 2016). These reviews compile together studies that investigate a common research questions and consider their findings in aggregate. The research question, which conforms to the standardized PICO criteria (Schardt et al., 2007), yields a binary answer, so findings from the associated review will take explicit "yes" or "no" stances. One such PICO question is "In elderly populations, does omega 3 acid from fatty fish intake, compared with no consumption, reduce the risk of developing heart failure?" (Li et al., 2013).

Pairs of claims manually annotated from these works can be paired together for NLI classification by matching claims that take the same stance on a common question as entailing pairs, those that take

[3] https://archive.physionet.org/physiotools/mimic-code/mednli/

opposite stances on a common question as contradicting pairs, and those taken from two different reviews about different questions as neutral pairs. The dataset's 16 PICO questions are split into 12, 4, and 4 questions for the train, validation, and test splits, respectively, and the neutral class is downsampled to be the same size as the next largest class in all splits. The resulting dataset has 2.8k claim pairs in total. The ManConCorpus is covered under a CC-BY-NC-SA license.[4]

## C Curriculum Design

To create an effective curriculum for the ultimate task of detecting contradictions in the COVID-19 treatment domain, we conducted a set of experiments analyzing the effect of multiple design decisions for incorporating domain-adjacent corpora in training.

### C.1 Experiments

#### C.1.1 Shuffled and Combined Curricula

To understand the importance of sequencing the curriculum, we evaluated BERT models trained using various sequences of domain-adjacent corpora in equal proportion. We consider three types of curricula: forward, reverse, and shuffled. The forward curriculum proceeds with fine-tuning a pre-trained BERT model in sequence from the most general domain (MultiNLI) to MedNLI to ManConCorpus to the most relevant domain (COVID-19 NLI). The reverse curriculum begins with the most relevant domain and proceeds in the opposite direction. The shuffled curricula were sampled from the 22 possible random orderings of the four domains excluding the forward and reverse sequences. We sampled three shuffled domains to assess the background from non-intentional curriculum design. Finally, we considered a "combined" curriculum where data from the four corpora are concatenated together and shuffled, thus ablating the notion of intentional sequencing in the curriculum. To ensure no dataset dominated training, each dataset, $D_{train}$, is subsampled such that $N_{D_{train}} = \min(d, |D_{train}|)$ samples are present in the curriculum.

#### C.1.2 Ordered Curriculum Subsequence Fine-Tuning

To assess the contribution to performance from specific domains during sequencing as well as the effect of curriculum size, we evaluated forward curriculum subsequences. Ten subsequences were evaluated: the full forward curriculum, two three-dataset subsequences, three two-dataset subsequences, and the four single corpora. As in C.1.1, $N_{D_{train}}$ samples are present in the curriculum from dataset $D_{train}$.

#### C.1.3 Perturbing Dataset Proportion in Sequential Curricula

To assess whether changing the ratio of training data used from the various corpora yielded better performance or to dilutive biases from larger corpora, we modulated the data ratio parameter. We define data ratio, $r$, as the multiplicative factor larger a dataset is from the next dataset in the curriculum sequence. Specifically, given $r$, we calculate the sample size of the dataset, $D_{train}$, to be used in the $i$th step (1-index) of a size-$k$ fine-tuning curriculum as $N_{D_{train}} = \min(r^{k-i}d, |D_{train}|)$. We considered three curricula: the full forward curriculum and the two sequential three-dataset curricula.

### C.2 Evaluation

#### C.2.1 Setup

We evaluated multiple BERT models pre-trained using general and biomedical corpora for curriculum-based fine-tuning (Devlin et al., 2019). Each fine-tuning step involves training for 4 epochs with learning rate $l = 10^{-5}$ and batch size $b = 8$. For all experiments, $d = 500$, and for data ratio experiments, $r \in \{1, 2\}$. Pre-trained models were loaded from the Hugging Face Transformers library (Wolf et al., 2020). All fine-tuning was conducted on 2 Tesla-V100-SXM2 and 2 Tesla-A100-PCIe GPUs. Experiments in curriculum design were evaluated with the pre-trained PubMedBERT model (Gu et al., 2021). Other pre-trained BERT models were evaluated on forward curriculum subsequences (Appendix D).

[4] http://staffwww.dcs.shef.ac.uk/people/M.Stevenson/resources/bio_contradictions/

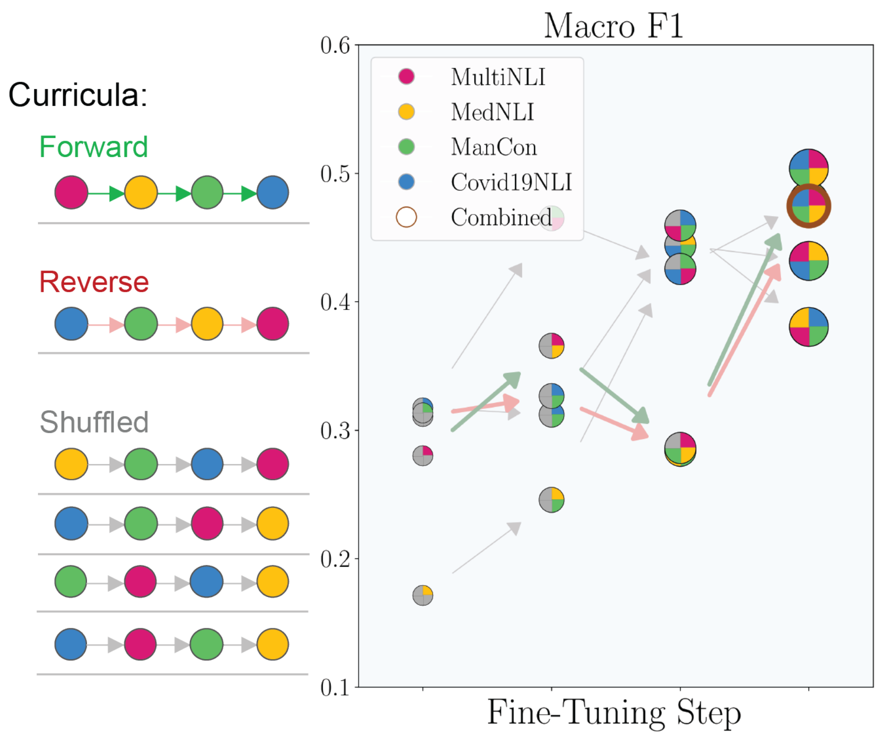


Figure 1: Macro F1 evaluation on the COVID-19 NLI validation set for different orderings of the fine-tuning curriculum and the combined curriculum (brown). Pie point position indicate F1 value, colors indicate which corpora have been already introduced to the model at a given fine-tuning step, and arrows indicate curriculum sequences. Pie size is proportional to amount of training data seen by the model.

#### C.2.2 Evaluation Metrics

The primary NLI evaluation metric for fine-tuned BERT models was macro F1 on the COVID-19 NLI validation set. We also investigated recall of the contradictions class as an important metric in evaluating the ability to detect contradictory research claims.

#### C.2.3 Shuffled and Collapsed Curricula

Of the six tested four-dataset curricula, the forward curriculum performed highest with an F1 of 0.503. The reverse curriculum, starting with the most relevant and challenging curriculum first, achieved an F1 of 0.474. The shuffled curricula yielded F1 scores of 0.380, 0.432, and 0.478. The collapsed curriculum, in which the four corpora are concatenated and shuffled, achieved competitive performance as well, yielding an F1 score of 0.475 (Figure 1).

#### C.2.4 Ordered Subsequences

From the 10 curriculum subsequences, the model trained with the full forward curriculum yielded highest performance with an F1 of 0.503. Among the two three-domain sequences, the one including the in-domain COVID-19 NLI dataset achieved greater performance than that without, yielding F1 scores of 0.440 and 0.296 respectively. Similarly, with the two-domain subsequences, the sequence with ManConCorpus and COVID-19 performed best with F1 of 0.434, and the subsequence containing MedNLI and ManConCorpus performed worst with F1 of 0.275. Among the single domain curricula, the in-domain training on our dataset was best with F1 of 0.311 (Figure 2).

#### C.2.5 Variable Dataset Proportions

In all three curricula, the condition with data ratio $r = 2$ outperformed the $r = 1$ equal data proportion condition. The highest performing curriculum was the $r = 2$ forward curriculum achieving an F1 of 0.638. In the in-domain three-dataset sequence, F1 increased from 0.416 with $r = 1$ to 0.461 with $r = 2$.

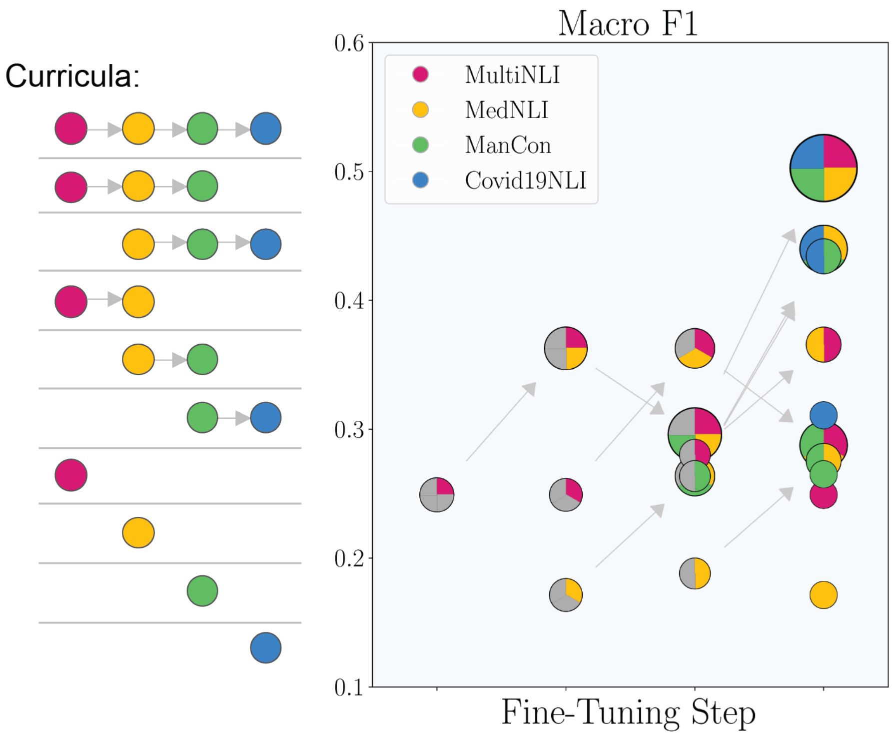


Figure 2: Macro F1 evaluation for various subsequences of the forward curriculum.

The out-of-domain three-sequence saw a similar increase in performance favoring $r = 2$ over $r = 1$ (Figure 3).

## D BERT Pretraining

Five pre-trained BERT models were evaluated for further fine-tuning: PubMedBERT (Gu et al., 2021), SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2020), BioClinBERT (Alsentzer et al., 2019), and RoBERTA (Liu et al., 2019). We conducted fine-tuning experiments under the same 10 subsequences and parameter settings as in Section C.1.2 and evaluated performance on the validation split of the COVID-19 NLI dataset. For PubMedBERT, SciBERT, and RoBERTa, the full forward curriculum yielded the greatest macro F1 scores at 0.503, 0.448, and 0.590, respectively. The greatest performance was achieved by the MedNLI-ManCon-COVID-19 NLI subsequence for BioBERT and BioClinBERT models yielding F1 scores of 0.433 and 0.354 (Figure 4). The models were used according to the licensing information provided at the Hugging Face pages for the models.[5]

## E BERT Hyperparameter Tuning

We evaluated macro F1 and contradictions recall on the COVID-19 NLI validation set over a parameter sweep of learning rates, $lr \in \{5\text{e}{-6}, 1\text{e}{-5}, 3\text{e}{-5}, 5\text{e}{-5}, 1\text{e}{-4}, 3\text{e}{-4}\}$ and batch sizes, $b \in \{4, 8, 16, 32\}$ for PubMedBERT and RoBERTa models. For both models the highest macro F1 setting was $lr = 3\text{e}{-5}$ and $b = 4$ yielding $F1 = 0.61$ and $F1 = 0.64$ for PubMedBERT and RoBERTa, respectively. These settings yielded the greatest contradictions recall of 0.51 for PubMedBERT, and settings of $lr = 5\text{e}{-6}, b = 4$ yielded the highest contradictions recall value of 0.39 for RoBERTa (Figure 5).

## F Test Set Evaluation and Baselines

We evaluated test set statistics for the COVID-19 NLI using PubMedBERT and RoBERTa (Liu et al., 2019) models fine-tuned with the forward curriculum of MultiNLI $\rightarrow$ MedNLI $\rightarrow$ ManCon $\rightarrow$ COVID-19

[5] https://huggingface.co/models

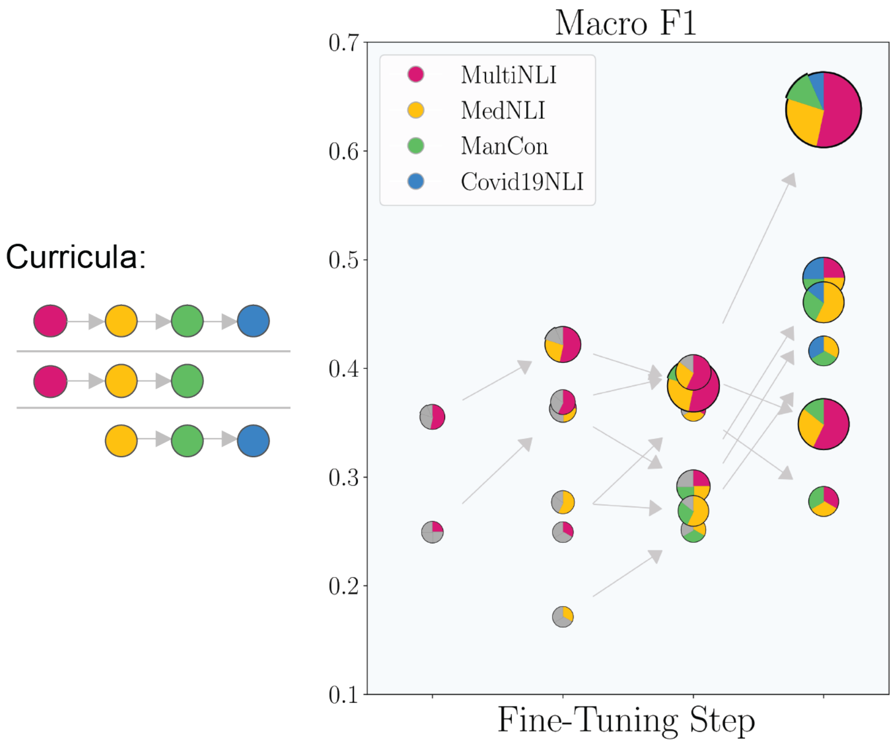


Figure 3: Evaluation of three subsequences of the forward curriculum with two different data ratio proportions, $r \in \{1, 2\}$.

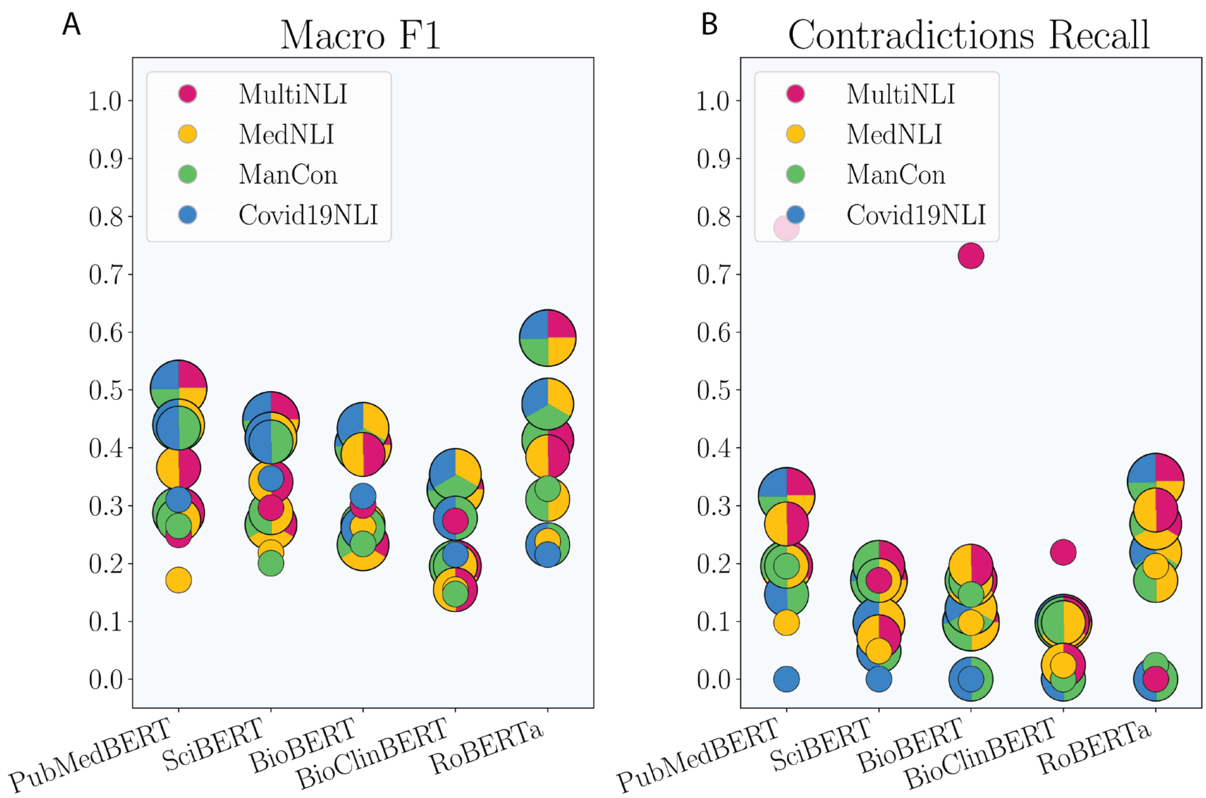


Figure 4: Macro F1 (A) and contradictions recall (B) for five pre-trained BERT models: PubMedBERT, SciBERT, BioBERT, BioClinBERT, and RoBERTA fine-tuned with subsequences of the forward curriculum.

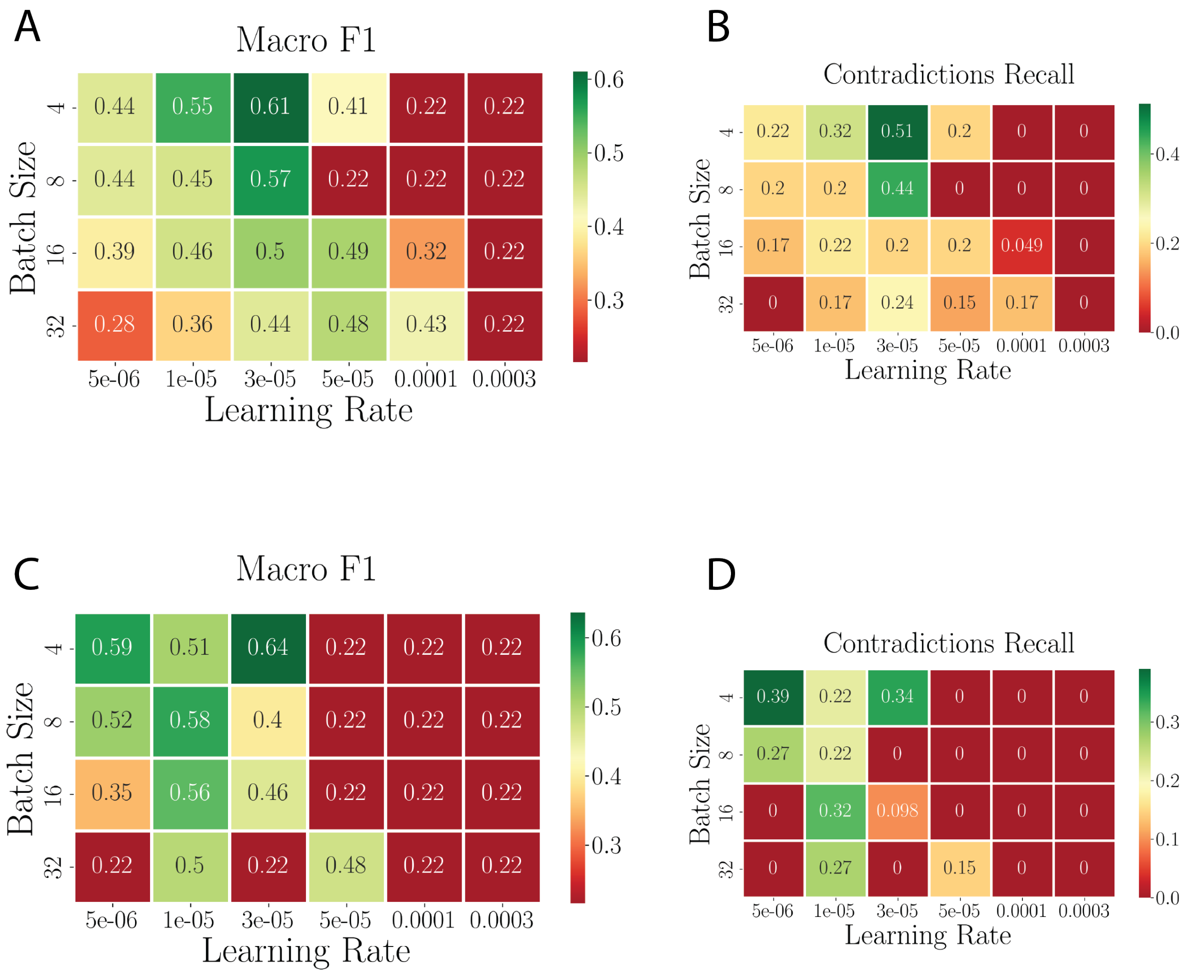


Figure 5: Batch size and learning rate hyperparameter optimization evaluated with the COVID-19 NLI validation metrics for PubMedBERT (A, B) and RoBERTa (C, D) models.

NLI. We set data ratio as being equal between the four corpora ($r = 1$) (see Appendix C.1.3), and after hyperparameter tuning of learning rate and batch size (Appendix E) set parameters $l_{HP} = 3 * 10^{-5}$ and $b_{HP} = 4$.

We compared performance of our trained BERT models to several NLI baselines.

- **Hypothesis-Only Unigrams** Softmax classification using unigram counts in the hypothesis (single claim).
- **Word Overlap** Softmax classification over counts of overlapping unigrams from the two claims.
- **Word Cross-Product** Softmax classification over counts of pairs of words in the cross-product between the two claims.
- **Similarity + Polarity** Softmax classification using similarity of the two claims as calculated using uSIF sentence embeddings (Ethayarajh, 2018; Borchers, 2019) and polarity of each claim using Vader polarity scores (Hutto and Gilbert, 2014).
- **Hypothesis-Only BERT** BERT classification where one of the two claims has been ablated.

Figure 6 offers a comparison of these baselines with our proposed models, focusing on the forward curriculum condition. We also evaluated the optimized PubMedBERT and RoBERTa models with the reverse curriculum and four shuffled curricula 4. We note the consistent result that the forward curriculum performs best overall.

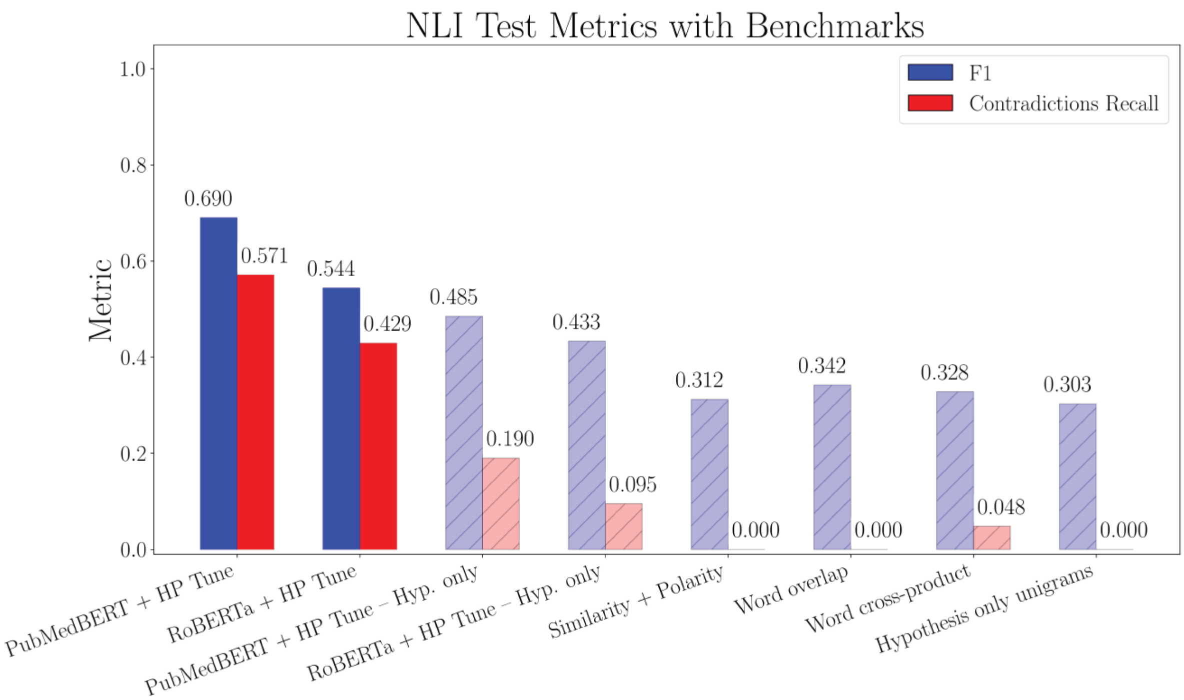


Figure 6: Macro F1 and contradictions recall for PubMedBERT and RoBERTa models fine-tuned with the forward curriculum compared to NLI benchmarks. 'HP Tune' indicates hyperparameter tuning (Appendix E).

| Model | Curriculum | F1 | Contra. Recall |
| --- | --- | --- | --- |
| PubMedBERT | Multi → Med → ManCon → Covid | **0.690** | **0.571** |
| | Covid → ManCon → Med → Multi | 0.428 | 0.381 |
| | Covid → Multi → ManCon → Med | 0.486 | 0.381 |
| | Covid → ManCon → Multi → Med | 0.581 | 0.571 |
| | Med → ManCon → Covid → Multi | 0.446 | 0.381 |
| | ManCon → Multi → Covid → Med | 0.579 | 0.333 |
| RoBERTa | Multi → Med → ManCon → Covid | 0.544 | 0.429 |
| | Covid → ManCon → Med → Multi | 0.411 | 0.476 |
| | Covid → Multi → ManCon → Med | 0.319 | 0.476 |
| | Covid → ManCon → Multi → Med | 0.232 | 0 |
| | Med → ManCon → Covid → Multi | 0.174 | 0 |
| | ManCon → Multi → Covid → Med | 0.232 | 0 |

Table 4: Test set performance on optimized PubMedBERT and RoBERTa models trained with various fine-tuning curricula.

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*Final required Limitations section*

☑ A2. Did you discuss any potential risks of your work?
*Final recommended Ethics section*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*Abstract and Section 1*

☒ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B ☑ Did you use or create scientific artifacts?

*Section 2*

☑ B1. Did you cite the creators of artifacts you used?
*Section 3, Appendix B, Appendix D*

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*Appendix A, Appendix B, Appendix D*

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*Appendix A, Appendix B, Appendix D*

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*All the data are sampled from publicly available research papers.*

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Section 2, Appendix A*

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*Section 2, Appendix A*

### C ☑ Did you run computational experiments?

*Section 3, Appendix C-F*

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*Appendix C.2.1*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Section 3, Appendix C-F*

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*Section 3, Appendix C-F*

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*Section 2, Appendix A*

**D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Section 2, Appendix A*

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*Section 2, Appendix A*

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*Not applicable. Annotators are professional clinical annotators.*

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*Not applicable. All the data are sampled from publicly available research papers.*

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Not applicable. Our research does not classify as human subjects research.*

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*Not applicable. The group of professional annotators is too small for such reporting and would violate privacy.*