# A Search Engine for Discovery of Scientific Challenges and Directions

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Keeping track of scientific challenges, advances and emerging directions is a fundamental part of research. However, researchers face a flood of papers that hinders discovery of important knowledge. In biomedicine, this directly impacts human lives. To address this problem, we present a novel task of extraction and search of scientific challenges and directions, to facilitate rapid knowledge discovery. We construct and release an expert-annotated corpus of texts sampled from full-length papers, labeled with novel semantic categories that generalize across many types of challenges and directions. We focus on a large corpus of interdisciplinary work relating to the COVID-19 pandemic, ranging from biomedicine to areas such as AI and economics. We apply a model trained on our data to identify challenges and directions across the corpus and build a dedicated search engine. In experiments with 19 researchers and clinicians using our system, we outperform a popular scientific search engine in assisting knowledge discovery. Finally, we show that models trained on our resource generalize to the wider biomedical domain and to AI papers, highlighting its broad utility. We make our data, model and search engine publicly available.[1]

## 1 Introduction

Success in scientific efforts hinges on identifying promising and important problems to work on, developing novel and effective solutions, and formulating hypotheses and directions for further exploration. Each new scientific advance helps address gaps in knowledge, including potential extensions and refinements of prior results. New advances often lead to new challenges and directions. With millions of scientific papers published every year, sets of challenges and potential directions for addressing them grow rapidly. A striking recent example is that of literature pertaining to the COVID-19 pandemic [36], which exploded in unprecedented volume with researchers from across diverse fields exploring the many facets of the disease and its societal ramifications. As the pandemic continues worldwide, it is especially urgent to provide scientists with tools for staying aware of advances, problems, and limitations faced by fellow researchers and medical professionals, and of emerging hypotheses or early indications of potential solutions.

Unfortunately, due to the immense scale and siloed nature of the scientific community, it can be difficult for researchers to keep track of their own specialty areas, let alone discover relevant knowledge in areas outside their immediate focus [12, 13, 14, 30]. This can result in poor awareness of failures or limitations reported in recent studies, wasting redundant resources and leading to clinical decision-making uninformed about shortcomings of interventions [5]. Disturbingly, there have been many cases where problems in treatments had been reported but not picked up by sectors of the clinical community [6, 31, 7] leading to higher rates of morbidity and mortality [17, 9, 33].

[1]Redacted for anonymity.

Submitted to NeurIPS 2021 AI for Science Workshop.

Our goal is to bolster the ability of researchers and clinicians to **keep track of difficulties, limitations and emerging hypotheses**. This could help clinical decision making be well-informed, accelerate innovation by surfacing new opportunities to work on, inspire new research directions, and match challenges with potential solutions from other communities [13]. In the face of challenging medical scenarios, such as the rise of a novel virus or situations where standard treatments fail, rapidly finding reports of similar challenges and directions to address them could have dramatic effect [26]. Finally, at the macro level, this ability could assist policymakers and funding agencies (e.g., NIH, NSF) seeking to identify important challenges and promising directions to prioritize research programs; in times of crisis this process needs to be done rapidly[2] but demands substantial human effort.

To address this problem and facilitate discovery of scientific knowledge, we make the following key contributions:

- **Novel Task: Extraction and Search of Scientific Challenges and Directions.** We define semantic categories for 'challenges' and 'directions' that generalize across many types of difficulties, limitations, flaws and hypotheses or potential indications that an issue is worthy of investigation. We focus on COVID-19 literature as the main test bed for our task, as it is known to be highly interdisciplinary [14] with research in many different fields (e.g., AI, climatology, engineering, economics) and relates to a global emergency that urgently demands tools to help researchers and clinicians keep track of challenges and new opportunities.
- **Expert-Annotated Dataset, Publicly Released.** We collect and publicly release a resource of 2.9K expert-annotated texts from full-length COVID-19 papers, labeled by experts for challenges and directions with high inter-annotator agreement. We use the data to train multi-label sentence classification models that achieve high accuracy scores. We analyze model errors, discovering that contextual information can both help and harm results. Based on this finding, we explore a simple technique that integrates multiple ways of encoding context.
- **Novel Scientific Search Engine For Researchers and Clinicians.** We build a novel public search engine that indexes challenges and directions. We apply a model trained on our dataset and apply it to the full corpus of 550K COVID-19 papers to build an index of scientific challenges and potential directions. We create a search engine that allows users to search for combinations of entities (e.g., names of drugs, diseases, etc.) and retrieve challenge/direction sentences that mention them.
- **Evaluating Generality: Zero-Shot Generalization across Biomedicine and AI.** We demonstrate zero-shot generalization, obtaining a high MAP of over 95% when applying the model trained on COVID-19 papers to a broader corpus in the general biomedical domain, and to AI papers in computer science. This indicates the potential value of our resource beyond COVID-19, such as for future pandemics or crises, or for helping AI researchers handle the explosion of research in this area.
- **Evaluating Utility: User Studies with Researchers.** We conduct studies measuring utility. First, we evaluate the system's ability to help researchers with diverse backgrounds discover challenges and directions for a given query (e.g., directions in *drug discovery*). This could also be important for researchers looking into a new area, e.g., AI researchers seeking biomedical problems (Fig. 1). Second, we recruit nine *medical researchers working on COVID-19* in clinical practice and research. These users often require finding information on challenges and directions, during research or treatment planning. In both experiments, totalling 19 researchers and over 70 distinct queries, our prototype outperforms PubMed, the most widely used biomedical search tool, in both quality and utility for discovery of challenges and directions.

## 2 Task Overview & Definitions

We present a novel task of automatically identifying sentences in papers that clearly state *scientific challenges and directions*. We consider the multi-label classification setting, where for a given sentence $\mathcal{X} = \{w_1, w_2, \ldots, w_T\}$ with $T$ tokens, our goal is to output two labels $\mathcal{Y} = \{c, d\}$, where $c$ and $d$ are binary targets indicating if the sentence mentions a challenge/direction, respectively. Additionally, we are also given *context* sentences surrounding $\mathcal{X}$: $(\mathcal{X}_{previous}, \mathcal{X}_{next})$, for the previous and next sentences, respectively, which could be used as further input to models. The multi-label

[2] `https://covidreviews.cochrane.org/`

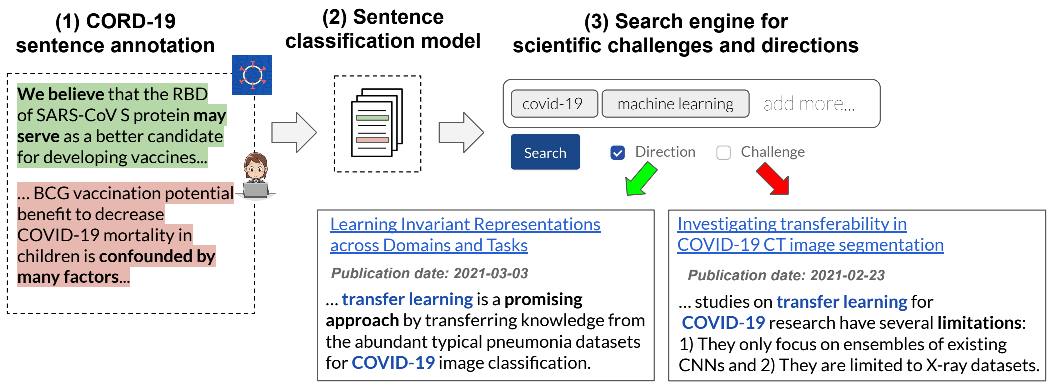


Figure 1: Overview of our system. (1) We collect expert annotations of sentences mentioning challenges and directions from across the CORD-19 corpus. (2) We train a sentence identification model on this data and apply it to the full corpus to extract high-confidence sentences. (3) We build a search engine indexing challenges and directions in COVID-19 literature, allowing users to search for entities and retrieve sentences with their contexts.

setting allows us to capture that in many cases, sentences refer to both challenges and directions at the same time At a high level, our labels are defined as follows.

- **Challenge:** A sentence mentioning a problem, difficulty, flaw, limitation, failure, lack of clarity, or knowledge gap.
- **Research direction:** A sentence mentioning suggestions or needs for further research, hypotheses, speculations, indications or hints that an issue is worthy of exploration.

Figure 1 shows examples for each category. Also, we further present the motivation for the task, example annotations, and why the categories are non-trivial for both humans & machines to identify in Technical Appendix §A.1 & §A.2. We note that in addition to biomedical literature discussed in the Introduction, our task is related to a body of research which we cover in Technical Appendix §A.3.

## 3 Data Collection and Models

**Data Collection & Annotation.** We sample 3000 sentences from the full-text papers of CORD-19 ($\sim$180k full-text papers with $\sim$25m sentences).[3] Four expert annotators with biomedical and bioNLP backgrounds annotated the sentences with high agreement.We create a train/dev/test stratified split of 40%/10%/50%, splitting by distinct *papers*.[4] See full details in Appendix §A.4.

**Baseline Models.** We evaluate a range of baseline models for our novel task: simple keyword/sentiment based heuristics, zero-shot inference based on a language model trained for natural language inference (NLI), fine-tuning scientific and non-scientific language models(LMs), and fine-tuning the LMs where they are context-aware (including the surrounding sentences in the training). We also explore two customized approaches: fine-tuning using a Hierarchical Attention Network (HAN) [38], and an approach where we obtain outputs using several variants ("slices") of context-aware and not-context-aware versions of a fine-tuned language model and combine their logits to yield a final pair of logits used for prediction. See Appendix §A.5 & §A.6 for a full description.

**Results.** As seen in Table 1, the best individual classifier by F1 is PubMedBERT with a binary-F1 of 0.770 and 0.766 on the challenge and direction labels, respectively. Our customized approach leads to an improvement of about one F1 point for both labels over the best individual model (standard error of $1.05 \times 10^{-4}$). See full results in Appendix §A.7 & error analysis in Appendix §A.8.

[3]We use a snapshot of CORD-19 from 08-02-2021.

[4]See Table §3 in Appendix §A.4.

| | Challenge | | | Direction | | |
|---|---|---|---|---|---|---|
| Model | P | R | F1 | P | R | F1 |
| Keyword | 0.535 | 0.760 | 0.628 | 0.455 | 0.792 | 0.578 |
| Sentiment | 0.405 | 0.966 | 0.571 | 0.239 | 0.837 | 0.371 |
| NLI | 0.659 | 0.693 | 0.675 | 0.401 | 0.825 | 0.540 |
| RoBERTa-lg | 0.723 (0.042) | 0.824 (0.046) | 0.769 (0.004) | 0.697 (0.065) | 0.825 (0.06) | 0.754 (0.004) |
| SciBERT | 0.729 (0.023) | 0.799 (0.03) | 0.761 (0.007) | 0.719 (0.044) | 0.783 (0.043) | 0.749 (0.01) |
| PubMedBERT | 0.738 (0.018) | 0.804 (0.017) | **0.770** (0.006) | 0.755 (0.017) | 0.778 (0.015) | **0.766** (0.006) |
| HAN | 0.671 (0.02) | 0.863 (0.03) | 0.759 (0.01) | 0.674 (0.04) | 0.804 (0.04) | 0.734 (0.001) |
| Ctx-slices | 0.742 (0.011) | 0.829 (0.012) | **0.783** (0.004) | 0.732 (0.02) | 0.82 (0.03) | **0.773** (0.005) |

Table 1: Model Results. The PubMedBERT model fine-tuned on our multi-label classification task performs best. For the neural models we present the average over 5 training seeds where the number in parentheses is the standard deviation.

**High precision@K** We observe that for 20% recall we obtain well over 90% precision, and for 40% recall about 90% precision, demonstrating the utility of our model for a search engine application (§4) where precision at the top is important.

**Generalization** For directions, we obtain MAP and AUC of around 96% for general biomedicine papers outside CORD-19, and around 95% for AI papers we sample from the computer science domain. For challenges, we reach a MAP and AUC reach around 97-98% for biomedicine and around 96% for AI. See full analysis in Appendix §A.7.

## 4 Search Engine User Studies

We build a search engine that indexes challenges and directions across the ∼550k papers in CORD-19. We perform entity linking to the biomedical KB of MeSH entities [24] which allows us to partially group together all challenges or directions into "topics" referring to a specific fine-grained combination of concepts. See Appendix §A.9.1 for full details.

**Experiment I - diverse researchers** We recruited ten participants with diverse research backgrounds. Each participant was given twenty queries (formulated by a domain expert). We compared the amount of challenges and directions they were able to find in a limited time frame using our system vs. the popular PubMed search system. Our system, on average, yielded ∼2 times as much challenges and ∼3 times as much directions per query. Further details are in Appendix §A.9.2.

**Experiment II - clinical researchers** We recruited nine expert MDs at a large hospital who are involved in clinical research for COVID-19 and for their specialty areas (each have over 1000 citations). Each expert completed randomly ordered search tasks (queries curated by an expert; see Appendix §A.12) using the same systems as in the previous experiment. After all search tasks were completed we use a standardized Post Study System Usability Questionnaire (PSSUQ) [21]. The experts strongly preferred our search engine to PubMed (overall average of 92% versus 59%, with non-normalized scores of 6.42 vs. 4.14). Further details are in Appendix §A.9.3.

## 5 Conclusion

We presented methods for extracting scientific challenges and directions from scholarly papers. We collected 3K expert-labeled sentences and their contexts from COVID-19 papers, and used the dataset to fine-tune scientific language models on our multi-label sentence classification task. We find that the approach can identify challenges and directions with high precision, and that using the model trained on our dataset achieves high zero-shot generalization on general biomedical papers and AI papers in computer science. We harnessed the model to index 950K sentences and build a novel search engine that allows researchers to search for biomedical entities and retrieve sentences mentioning difficulties, limitations, hypotheses and directions. Researchers using our system found that our system provided better support than PubMed in terms of utility and relevance.

# References


[1] Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, et al. Concept annotation in the craft corpus. *BMC bioinformatics*, 2012.

[2] Iz Beltagy, Kyle Lo, and Arman Cohan. 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 3606–3611, 2019.

[3] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 2004.

[4] D. Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *EMNLP*, 2020.

[5] Iain Chalmers, Michael B Bracken, Ben Djulbegovic, Silvio Garattini, Jonathan Grant, A Metin Gülmezoglu, David W Howells, John P A Ioannidis, and Sandy Oliver. How to increase value and reduce waste when research priorities are set. *The Lancet*, 383(9912):156–165, 2014. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(13)62229-1. URL `https://www.sciencedirect.com/science/article/pii/S0140673613622291`.

[6] M. Clarke and S. Hopewell. Many reports of randomised trials still don't begin or end with a systematic review of the relevant evidence. *Journal of the Bahrain medical society*, 24:145–148, 2013.

[7] N. Cooper, David R. Jones, and A. Sutton. The use of systematic reviews when designing studies. *Clinical Trials*, 2:260 – 264, 2005.

[8] Franck Dernoncourt and Ji Young Lee. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, 2017.

[9] R. Gilbert, G. Salanti, M. Harden, and S. See. Infant sleeping position and the sudden infant death syndrome: systematic review of observational studies and historical review of recommendations from 1940 to 2002. *International journal of epidemiology*, 34 4:874–87, 2005.

[10] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*, 2020.

[11] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*, 2020.

[12] Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. Accelerating innovation through analogy mining. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 235–243, 2017.

[13] Tom Hope, Jason Portenoy, Kishore Vasan, Jonathan Borchardt, Eric Horvitz, Daniel S Weld, Marti A Hearst, and Jevin West. Scisight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 135–143, 2020.

[14] Tom Hope, Aida Amini, David Wadden, Madeleine van Zuylen, Sravanthi Parasa, Eric Horvitz, Daniel Weld, Roy Schwartz, and Hannaneh Hajishirzi. Extracting a Knowledge Base of Mechanisms from COVID-19 Papers . 2021.

[15] Ting-Hao Huang, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Yen-Chia Hsu, and C Lee Giles. Coda-19: Using a non-expert crowd to annotate research aspects on 10,000+ abstracts in the covid-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, 2020.

[16] Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. Scirex: A challenge dataset for document-level information extraction. *arXiv preprint arXiv:2005.00512*, 2020.

[17] Katharine Ker, Phil Edwards, Pablo Perel, Haleema Shakur, and Ian Roberts. Effect of tranexamic acid on surgical bleeding: systematic review and cumulative meta-analysis. *BMJ*, 344, 2012. doi: 10.1136/bmj.e3054. URL `https://www.bmj.com/content/344/bmj.e3054`.

[18] Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, 2004.

[19] Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. The genia event extraction shared task, 2013 edition-overview. In *BioNLP Shared Task Workshop*, 2013.

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[21] James R Lewis. Psychometric evaluation of the pssuq using data from five years of usability studies. *International Journal of Human-Computer Interaction*, 2002.

[22] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[23] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

[24] Carolyn E Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

[25] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S. Weld. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of ACL*, 2020. URL `https://arxiv.org/abs/1911.02782`.

[26] Christopher A Longhurst, Robert A Harrington, and Nigam H Shah. A 'green button'for using aggregate patient data at the point of care. *Health affairs*, 33(7):1229–1235, 2014.

[27] Steven Loria. textblob documentation. *Release 0.15*, 2, 2018.

[28] Sunil Mohan and Donghui Li. Medmentions: A large biomedical corpus annotated with umls concepts. In *Automated Knowledge Base Construction (AKBC)*, 2018.

[29] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. Scispacy: Fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*, 2019.

[30] Jason Portenoy, Marissa Radensky, Jevin West, Eric Horvitz, Daniel Weld, and Tom Hope. Bridger: Toward bursting scientific filter bubbles and boosting innovation via novel author discovery. *arXiv preprint arXiv:2108.05669*, 2021.

[31] K. Robinson and S. Goodman. A systematic examination of the citation of prior research in reports of randomized, controlled trials. *Annals of Internal Medicine*, 154:50 – 55, 2011.

[32] Micah Shlain, Hillel Taub-Tabib, Shoval Sadde, and Yoav Goldberg. Syntactic search by example. In *ACL*, 2020.

[33] J. Sinclair. Meta-analysis of randomized controlled trials of antenatal corticosteroid for the prevention of respiratory distress syndrome: discussion. *American journal of obstetrics and gynecology*, 173 1: 335–44, 1995.

[34] Hillel Taub-Tabib, Micah Shlain, Shoval Sadde, Dan Lahav, Matan Eyal, Yaara Cohen, and Yoav Goldberg. Interactive extractive search over biomedical corpora, 2020.

[35] Byron C. Wallace, Joël Kuiper, Aakash Sharma, Mingxi Zhu, and Iain J. Marshall. Extracting pico sentences from clinical trial reports using supervised distant supervision. *J. Mach. Learn. Res.*, 17(1): 4572–4596, January 2016. ISSN 1532-4435.

[36] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. Cord-19: The covid-19 open research dataset. *arXiv preprint arXiv:2004.10706*, 2020.

[37] Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 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, 2020.

[38] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[39] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3905–3914, 2019.

# A Technical Appendix

## A.1 Task Overview Definitions

The CORD-19 corpus [36] curates literature on COVID-19 and related diseases. With many thousands of papers, keeping track is generally hard, and mapping the landscape of scientific challenges and directions to address them is even harder. While "grand" challenges such as designing therapies and handling novel virus variants are broadly known, research focuses on *fine-grained* specific challenges, e.g., difficulties in functional analysis of specific viral proteins, or shortcomings of a specific treatment regime for children. Each challenge, in turn, is associated with potential directions and hypotheses.[5]

As written in the main paper, we present a novel task of automatically identifying sentences in papers that clearly state *scientific challenges and directions*. We consider the multi-label classification setting, where for a given sentence $\mathcal{X} = \{w_1, w_2, \ldots, w_T\}$ with $T$ tokens, our goal is to output two labels $\mathcal{Y} = \{c, d\}$, where $c$ and $d$ are binary targets indicating if the sentence mentions a challenge/direction, respectively. Additionally, we are also given *context* sentences surrounding $\mathcal{X}$: $(\mathcal{X}_{previous}, \mathcal{X}_{next})$, for the previous and next sentences, respectively, which could be used as further input to models. The multi-label setting allows us to capture that in many cases, sentences refer to both challenges and directions at the same time (see Table 2). At a high level, our labels are defined as follows.

- **Challenge:** A sentence mentioning a problem, difficulty, flaw, limitation, failure, lack of clarity, or knowledge gap.
- **Research direction:** A sentence mentioning suggestions or needs for further research, hypotheses, speculations, indications or hints that an issue is worthy of exploration.

These categories allow us to capture important information for scientists that is not captured by existing resources (see §A.3). As part of data annotation we provide annotators with richer explanations and examples of each label (see §A.4) to make these definitions more concrete. Figure 1 shows examples for each category (also see Table 2 in Technical Appendix §A.2 for more discussion).

Many cases of challenges and directions are non-trivial for both humans and machines to identify. We demonstrate two main types of difficulties (see more discussion in Technical Appendix §A.8) — cases of potentially misleading keywords, and cases where deep domain knowledge or context may be required.

- **Misleading keywords**. Consider the following sentence: *"The 15-30 mg/L albumin concentration is a critical value that could indicate kidney problems when it is repeatedly exceeded"*. This text mentions a diagnostic measure that is an indicator of a problem, rather than an actual problem. This is one example out of many other potentially misleading cases, such as cases where a negative outcome occurs to an entity we wish to harm (e.g., "the viral structural integrity is destroyed").
- **Context and domain knowledge**. *"BV-2 cells expressed Mac1 (CD11b) and Mac2 but were negative for the oligodendrocyte marker GalC and the astrocyte marker GFAP."* Deciding whether this sentence contains a challenge is highly non-trivial, since it requires more context and deep domain knowledge to understand whether this outcome is problematic or not.

## A.2 More examples

Table 2 shows example sentences for each category. In the first row (*not challenge, not direction*), the example is purely a factual description of a certain tool. In the second row (*not challenge, direction*), the statement mentions a scientific future direction, but there is no associated challenge that is explicitly mentioned. In the third row (*challenge, not direction*), there is a mention of a disease that is difficult to diagnose, but there is no mention of a suggested hypothesis or direction. Finally, in the last row (*challenge, direction*), a medical concern is presented alongside a scientific speculation on the nature of the signaling in the immune system, therefore reflecting both a challenge and a direction.

[5] While many papers discuss future directions in their concluding section, our task involves capturing all mentions of directions/hypotheses/speculations/early indications appearing throughout *full* paper texts (e.g., in experimental analysis sections).

| Labels | Example |
|---|---|
| Not Challenge, Not Direction | Nowadays, standard structure-based virtual screening has been routinely implemented in drug discovery to quickly prioritize potential compounds for in vitro activity tests. |
| Not Challenge, Direction | Future studies will focus on comparative sequence analysis between the PST isolates reported herein and global isolates of PST to determine the specific geographic origin(s) for this diverse PST population. |
| Challenge, Not Direction | Outbreaks attributed to acute BVDV infections in feedlot calves have been described previously, although definitive diagnosis is often difficult [18]. |
| Challenge, Direction | Thus, both PRRs could be responsible for innate immune signaling during acute DENV infection, perhaps operating in temporally distinct fashion as in WNV infection. |

Table 2: Examples of Challenges and Directions.

### A.3 Related Work

In addition to biomedical literature discussed in the Introduction, our work on extracting challenges and directions from scientific papers is related to a large body of research.

**Scientific information extraction and text classification.** The goal in this line of work is to extract structured information from literature, such as sentence-level classification into categories including objectives/methods/findings [8] or extracting entities and relations [23, 35, 19]. Unlike previous work, our labelling schema encapsulates underexplored facets such as difficulties, flaws, uncertainties (challenges) and suggestions, hypotheses, indications that an issue is worthy of additional exploration (directions). Our coarse-grained schema covers diverse variants of challenges/directions and can help generalize across the interdisciplinary COVID-19 literature [14].

**COVID-19 IE and search tools.** Recent work includes visualizing COVID-19 concepts and relations [13], a syntactic search engine [32], and a search engine for causal and functional relations [14]. Ours system is focused on challenges and directions, not captured by existing tools. Recent work [15] has used crowd workers to annotate *abstracts* (not full-texts as in this paper) for Background, Purpose, Method, Finding/Contribution. As discussed in §A.4, we find that crowd workers fail on our task, even though recruited with high quality assurance standards.

### A.4 Data Collection Procedure

We recruited four expert annotators with biomedical and bioNLP backgrounds to annotate sentences sampled across CORD-19. Annotators were given detailed annotation guidelines[6] and had a one-hour training session for reviewing the guidelines and discussing more examples. The guidelines included simple explanations of challenges and directions along with introductory examples. We sampled sentences from full-text papers, aiming to capture diverse, fine-grained challenges/directions that often do not appear in abstracts. The subset of full-text papers in CORD-19 numbers roughly 180K papers with around 25 millions sentences.[7] We also provide surrounding sentences around the target sentence as context.

[6]Annotation guidelines are available in REDACTED FOR ANONYMITY.

[7]We use a snapshot of CORD-19 from 08-02-2021.

| Labels | Train | Dev | Test | All |
|---|---|---|---|---|
| Not Challenge, Not Direction | 602 | 146 | 745 | 1493 |
| Not Challenge, Direction | 106 | 25 | 122 | 253 |
| Challenge, Not Direction | 288 | 73 | 382 | 743 |
| Challenge, Direction | 155 | 40 | 210 | 405 |

Table 3: Distribution of labels across data splits. Splits are stratified with no overlap in papers.

Randomly sampling sentences for annotation is highly unlikely to lead to enough challenge/direction cases. To increase this likelihood, two annotators curate 280 keywords or phrases with affinity to one of the two categories.[8] Sentences mentioning at least one keyword (lemmatized) are upsampled. For example, words such as *unknown, limit, however* provide weak signal indicating a potential mention of a challenge; words like *suggest, future work, explore* are weak indicators of a direction. To expand the list further, annotators made use of SPIKE [34] which also has a vocabulary explorer that allows browsing keywords similar to an input term. Overall, the 280 keywords covered around a third of sentences in CORD-19, demonstrating their breadth. We note that for most keywords context can completely change their meaning; for instance, "limit" can appear in the context of "we limit the discussion" which has no relation to challenges. Our set of terms with weak correlation to the label (e.g., the word *may* that very weakly relates to directions) favors high recall rather than precision.

Finally, to further increase coverage, we sampled at random roughly a quarter of sentences from the remaining sentences that did *not* contain any of the keywords, obtaining in total 3000 sentences. We filter sentences that are not in English, mostly numeric/mathematical, or that are very short/long (often due to PDF parsing issues), resulting in 2894 sentences and their surrounding contexts, from 1786 papers.

**Annotator agreement:** 60% of the sentences were labeled by all annotators[9], with high average pairwise agreement. Following common practice we measure micro-F1 and macro-F1, treating labels from one annotator as ground-truth and the other as predicted, obtaining 85% for challenges and 88% for directions for micro-F1, and 84% and 82% for macro-F1. Positive label proportions are 39.66% and 22.74% for challenges/directions, respectively. We create a train/dev/test stratified split of 40%/10%/50%(Table §3), splitting by distinct *papers*. We opt for a large, diverse test set for model evaluation [4]. The sampled sentences originate from papers published in 1108 journals.

#### A.4.1 A note on crowdsourcing.

We also attempted crowdsourcing to scale the collection process.[10] However, despite multiple trials and strict quality assurance, the nuanced nature of the task was found to be difficult for crowd workers, especially due to false negatives.

### A.5 Baseline models

The classification task at hand is a multi-label sentence classification problem, with the goal of predicting whether a sentence mentions a challenge, a research direction, both, or neither. The definitions of the challenge and direction categories are as described in §2. We evaluate a range of baseline models we examine for our novel task.

- **Keyword-based**: A simple heuristic based on the lexicon we curated for data collection (§A.4) — sentences with a challenge keyword are labeled as challenge, and similarly for direction.
- **Sentiment**: Challenge statements potentially have a negative tone, and directions are potentially more positive. We score the sentiment of each sentence using an existing tool [27] and classify negative sentiment sentences as challenges and positive ones as directions.
- **Zero-shot inference**: In zero-shot classification, models predict labels they were not trained on [39]. This could be particularly relevant in emerging domains such as COVID-19, where collecting

[8]Our list of keywords is available in REDACTED FOR ANONYMITY.

[9]Final labels selected by majority vote, with ties (fewer than 100 cases) adjudicated by a member of the research team.

[10]Using the Appen platform `https://appen.com/`.

large amounts of labeled data could be prohibitive. We use a language model trained for natural language inference (NLI), letting the model infer whether the input text *entails* the label name. See Appendix §A.5.1 for full details.

- **Scientific language models**: We also experiment with fine-tuning language models that were pre-trained on scientific papers. We report results for PubMedBERT-abstract-fulltext [10] which was pre-trained on PubMed paper abstracts & full texts, and for SciBERT [2], trained on a corpus of biomedical and computer science papers. In addition, we also experiment with a non-scientific language model, RoBERTa-large, which has been shown to obtain excellent results when fine-tuned on scientific texts [11]. We also experimented with other language models, with very similar results.

  For all language models we fine-tune we use the Hugging Face library [37][11]. We use hyperparameter tuning with the objective of maximizing the F1-score on the development set using grid search over batch size ([8,16,32]), learning rate ([1e-05, 2e-05, 3e-05, 5e-05]) and epochs (maximal value of 25 epochs). We use the Adam optimizer[20] with a dropout rate of 0.3 for all neural models, using a binary cross-entropy (BCE) loss over our two labels. For the sentiment analysis model, we tune its threshold on the development set.

See customized models at Technical Appendix §A.6.

### A.5.1 Zero-shot baseline and variations specification

We use BART-MNLI-large [22], a pre-trained NLI model. We find that simply feeding in "challenge" and "direction" as label names, or similar variants, performs poorly, likely due to the nuanced complexity of these labels. Instead of using one name, we find that enumerating multiple variants of challenges (e.g., difficulty, limitation, failure) provides better results.

We define the following sub-labels enumerating challenges and directions. We take the different variants of challenges that we use in our definition of this label — [challenge, problem, difficulty, flaw, limitation, failure, lack of clarity, gap of knowledge] — and similarly for directions ( [direction, suggestion, hypothesis, need for further research, open question, future work]). Denote the former list by $\mathcal{L}_c$, and the latter $\mathcal{L}_d$. For each category, we compute the probability of each $l \in \mathcal{L}_c$ ($\mathcal{L}_d$, respectively) and take the maximal value for each set of sub-labels, denoted by $m_c$ and $m_d$. If $m_c \geq 0.9$ we label the sentence as a challenge, and similarly for $m_d$ and directions, using the same threshold. Otherwise, the input is classified as negative.

We briefly examine a few variations on the zero-shot classification baseline, in terms of the class/label names given as input, to study their effect. We use the same binary threshold of 0.9 for all variants.

- Class-name: Using only the class names, i.e., "challenge" and "direction", rather than more fine-grained label names.
- Template: Using "challenge" and "future direction" as part of a template sentence following the approach in Yin et al. [39]. Specifically, "This sentence is about a challenge", "This sentence is about a future direction".
- Concatenated: Instead of [challenge, problem, difficulty, flaw, limitation, failure, lack of clarity, gap of knowledge] as standalone inputs, we concatenate them into one string — "challenge, problem, difficulty, flaw, limitation, failure, lack of clarity, gap of knowledge"; the same was done for directions.

Table 4 in Technichal Appendix §A.7.2 shows the results of the zero-shot variant models.

## A.6 Context Modelling Variants

We also experiment with models motivated by examination of baseline errors (see Technical Appendix §A.8). Specifically, we find that adding context helps in certain cases: For example, in the sentence *"... the patient had an extreme elevation of procalcitonin without signs of bacterial infection."* which was misclassified as a non-challenge, adding context helped identify the unexplained elevation as problematic. However, context can also introduce noise (see Table 1). We explore different ways in which the context can affect predictions — during training, and during inference. In addition to simply fine-tuning PubMedBERT with full context, we explore two main customized approaches.

[11] See our code attached in the REDACTED FOR ANONYMITY.

**Hierarchical Attention Network (HAN) [38]** Recall Section §2, where candidate sentences are denoted by $\mathcal{X}$ and their surrounding context by $\mathcal{X}_{previous}, \mathcal{X}_{next}$. Denote by $\mathcal{X}_{context}$ the concatenation: `[CLS]` $\mathcal{X}_{previous}$ `[SEP]` $\mathcal{X}$ `[SEP]` $\mathcal{X}_{next}$ `[SEP]`. We compute a weighted average of `[CLS]` and the first two `[SEP]` tokens using attention weights, and use this average embedding for final classification. The weights are learned as part of end-to-end training. [12] While this model can potentially learn to re-weight the context, it encodes the full $\mathcal{X}_{context}$ jointly before this weighting takes place, which can lead to noise propagating early on. We thus test a different approach to help mitigate this issue.

**Context Slice + Combine** Let $f_{\mathcal{X}}(\mathbf{x})$ denote the label logits emitted from the final layer of the PubMedBERT model which was fine-tuned on $\mathcal{X}$ only, for some input text $\mathbf{x}$. Likewise, denote by $f_{\mathcal{X}_{context}}(\mathbf{x})$ the logits from PubMedBERT fine-tuned using the *full context*. At inference time, we obtain outputs using the following variants ("slices") of $f$ and $\mathbf{x}$: (1) $l_1 = f_{\mathcal{X}}(\mathcal{X})$, (2) $l_2 = f_{\mathcal{X}context}(\mathcal{X}_{context})$, (3) $l_3 = f_{\mathcal{X}}(\mathcal{X}_{context})$, and (4) $l_4 = f_{\mathcal{X}context}(\mathcal{X})$. We then average ("combine") all four, yielding a final pair of logits used for prediction. (1) and (2) are just the models reported in Table 1 – feeding $\mathcal{X}$ as input to PubMedBERT fine-tuned on $\mathcal{X}$, and similarly for $\mathcal{X}_{context}$. (3) and (4) switch between training and inference inputs: in (3) $f_{\mathcal{X}}$ takes $\mathcal{X}_{context}$ as input during inference, and in (4) $\mathcal{X}$ is fed as input into $f_{\mathcal{X}_{context}}$. The reason we include these is to tease apart different ways in which the context may introduce noise or signal, during training using context (3) and during inference (4). We empirically find all four are in agreement in roughly 70%  83% of the cases for challenges  directions; 3 out of 4 agree in 20%  11%, and the rest are tied. This suggests each variant can potentially capture complementary information.

### A.7 Results

**Classification Results.** As seen in Table 1, fine-tuned scientific language models outperform the Zero-Shot model, which still does well considering it had no supervision and was pre-trained on non-scientific texts. The sentiment analysis and keyword-based classifiers, both based on large lists of "positive/negative" keywords, have good recall but poor precision. The best individual classifier by F1 is PubMedBERT with a binary-F1 of 0.770 and 0.766 on the challenge and direction labels, respectively. The HAN approach was able to increase recall substantially for problems, but at the cost of reduction in precision, leading to overall inferior F1 on par with PubMedBERT+context.

The Slice-Combine approach leads to an improvement of about one F1 point for both labels over the best individual model (standard error of $1.05 \times 10^{-4}$). In an ablation experiment we compute the averaged logits of $l_1, l_2$ and $l_3, l_4$ *separately*, and also simply ensemble four model runs of fine-tuned PubMedBERT, both leading to inferior results (see these additional results in Technical Appendix §A.7.1). Finally, an oracle that selects the best logit $l_1$-$l_4$ for each input based on ground truth labels has F1 of 0.907 and 0.896 for challenges/directions, suggesting much room for future work on adaptive use of context during training and inference. See in-depth analysis of additional model errors in Technical Appendix §A.8.

**Precision@Recall** Our primary focus is a novel search engine application (§4). For such applications, it is often more important to have high precision for top retrieved results. We examine precision for a range of values of recall, shown in Figure 2. We observe that for 20% recall we obtain well over 90% precision, and for 40% recall about 90% precision.

**Evaluating predictions across CORD-19** To further ensure quality, we run the PubMedBERT model across all sentences in CORD-19. Out of all sentences indexed in our search engine as either a challenge or a direction, we sample roughly 350 sentences, with higher sampling weight given to high-confidence predictions. About 190 sentences have confidence greater than $0.9$, 130 have confidence lower than $0.5$, with 90 sentences with confidence within the range of $(0.25, 0.75)$. These sentences are labeled by an expert annotator following the same criteria used to annotate our dataset (§A.4). As shown in Figure 3, we obtain very high mean average precision (MAP) of 98% and area under the precision-recall curve (AUC) of over 97% for directions, and 97% / 96% for challenges. We conclude that for high-confidence challenge and direction sentences indexed in our search engine, accuracy is expected to be overall considerably high. Our test set consists of considerably harder examples, explaining the gap in performance (see discussion in §A.8).

[12] See Yang et al. [38] for details about the general framework.

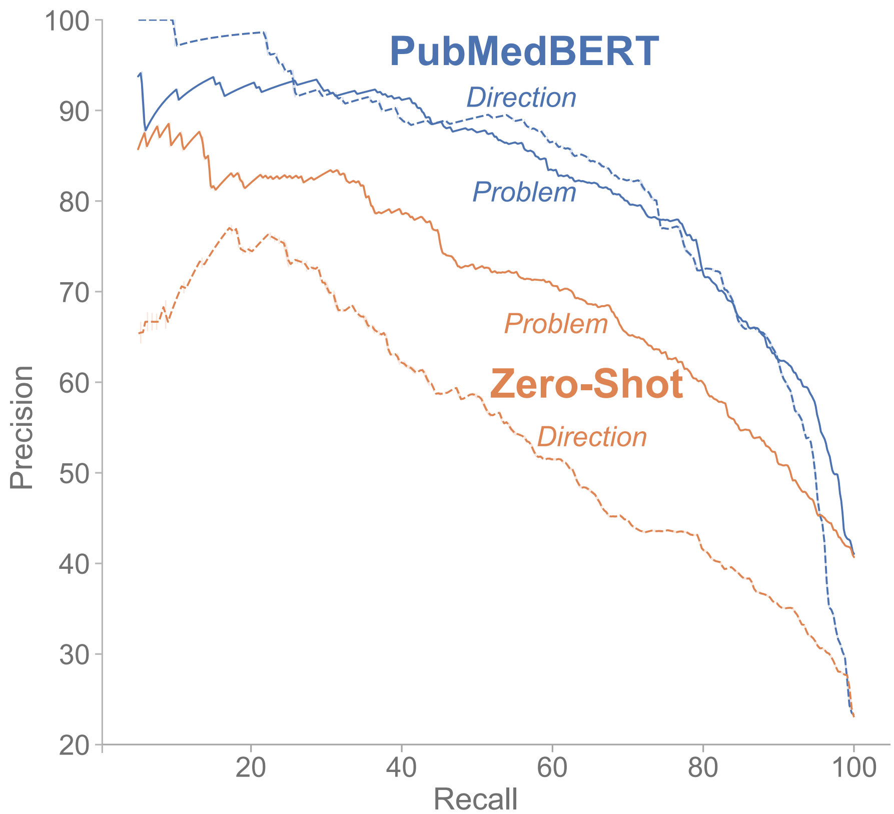


Figure 2: Precision/Recall results for the PubMedBERT model, and the zero-shot model. Precision for PubMedBERT is high for reasonably large values of recall.

**Zero-shot generalization to biomedicine and AI domains**. We perform a preliminary experiment examining whether a model trained on our dataset can, with no additional training, generalize to identify challenges and directions in *general* biomedical papers, which we sample from S2ORC, a larger corpus with millions of papers [25], and also AI papers sampled from a corpus of full-text computer science papers [16]. In total, we sample about 1000 sentences across the two datasets, following the same procedure as described above for CORD-19 sentences: we randomly sample 100 papers that did not appear in the CORD-19 corpus to ensure no leakage of information from our training set (we filter with a paper identifier shared by both resources). From these papers, we sample sentences in the same way as above for annotation. The annotator labels 630 sentences: 430 sentences with confidence scores greater than $0.9$ and 200 sentences with scores lower than $0.9$, 150 of those with scores lower than $0.5$. For AI papers, we follow the same procedure, with 300 sampled sentences. CORD-19 papers are highly interdisciplinary in terms of the areas they cover [14, 13], raising the possibility of using our dataset to train models that can be applied to other domains without additional costly data collection. As seen in Figure 3, identification accuracy is high in this sample too. For directions, we obtain MAP and AUC of around 96% for biomedicine, and around 95% for AI. For challenges, MAP and AUC reach around 97-98% for biomedicine and around 96% for AI. These preliminary results focus mostly on high-confidence predictions relevant for search applications, and generalization could be explored more extensively in future work.

### A.7.1 Additional Slice-Combine Context Models Results

In addition to the experiments reported, we tested multiple ways to combine information from four variants: (i) apply average or a median on the logits, (ii) majority voting, (iii) log-odds extremization, (iv) training a router model based on the logit differences, (v) running logistic regression with the embedding (final layer) of each of the four input encoders and their logits as features. Aside from the simple averaging, logistic regression was a close runner-up. We explore the average weights the

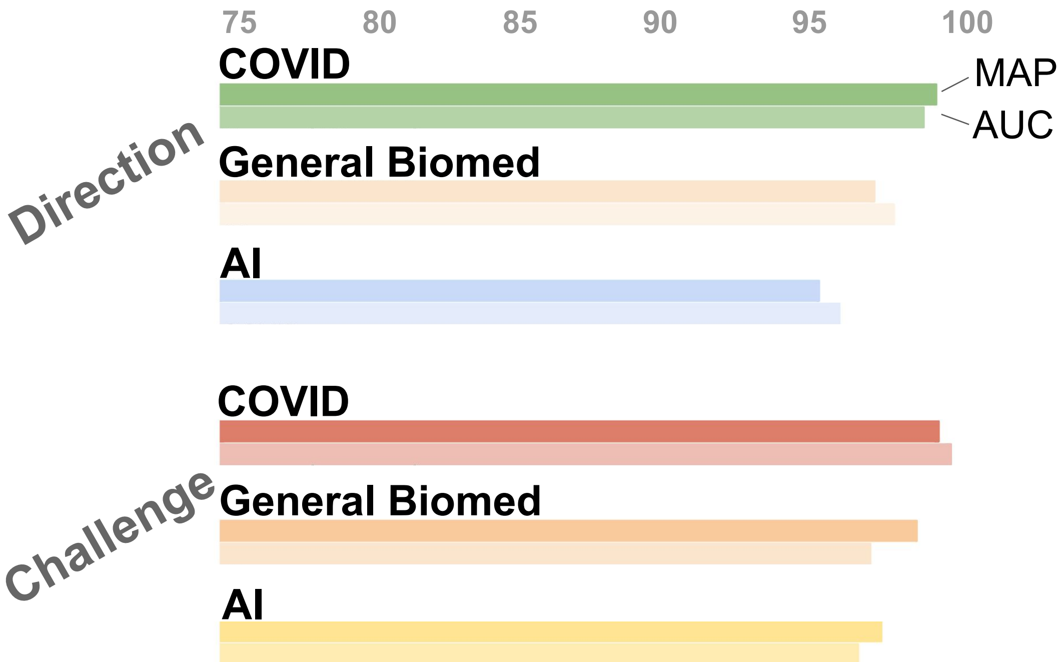


Figure 3: Evaluating predictions beyond our test set. We use a model trained on our data to identify challenges and directions across CORD-19 (denoted by *COVID*), S2ORC (general biomedical papers, denoted *general biomed*) and SciRex (full-text AI papers, denoted *AI*). Accuracy is considerably high. Zero-shot generalization over non-COVID papers, even non-biomedical papers, is encouragingly high, indicating the utility of our resource beyond COVID-19.

| | **Challenge** | | | **Direction** | | |
|---|---|---|---|---|---|---|
| Variation | P | R | F1 | P | R | F1 |
| NLI-Zeroshot | 0.659 | 0.693 | **0.675** | 0.401 | 0.825 | 0.54 |
| Class-name | 0.789 | 0.440 | 0.565 | 0.618 | 0.065 | 0.119 |
| Template | 0.439 | 0.941 | 0.599 | 0.589 | 0.401 | 0.478 |
| Concatenated | 0.849 | 0.107 | 0.190 | 0.491 | 0.724 | **0.585** |

Table 4: Zero-Shot Baseline Ablation Results. We provide the baseline different variants of class descriptors for challenges and directions, respectively.

logistic regression assigned to the four context variants. For challenges 0.14, 0.21, 0.21, and 0.35 for (1)-(4) respectively; and for directions 0.32, 0.2, 0.07 and 0.45. Interestingly this suggests that training the model end-to-end with context could be useful even when the context is not available at inference.

**Sanity testing the Slice-Combine Context Models** As a sanity check we simply ensemble 4 runs of PubMedBERT, resulting in inferior F1 of 0.772 and 0.764 for challenges/directions, which further indicates the complimentary value of our four context variants beyond simpler ensembling.

### A.7.2 Zero-Shot Baseline Ablation Results

Results for the ablation experiments are in Table 4. As can be observed, for challenges, the variant reported in Table 1 achieves the best results by a large margin. For directions, the variant that concatenates class descriptors into one string does marginally better.

### A.8 Error analysis

We study the cases where the best fine tuned model failed to classify sentences correctly. In order to do so we randomly sampled and analyzed roughly 20% of the false positive and false negative errors across both labels.

**Challenges** The most common error that accounts for a third of wrong predictions (both false positive and negative) is that in some cases deciding whether an outcome is positive or negative requires a more profound understanding of the biomedical entities involved and of the context. For example, the sentence consider the sentence *"The surprising conclusion of the study was that relative to primary rat Schwann cells undergoing myelination, only 2 cell lines expressed high levels of mRNA coding for myelin proteins and none of the cell lines expressed all of the myelin proteins typically expressed in myelinating Schwann cells."* The model classifies the sentence as a challenge. However, an expert who read the text concluded that the outcome is non-problematic since with further downstream analysis the mentioned conclusion may represent a more accurate model for future analysis of myelin gene promoters. Conversely, the sentence *"It is remarkable that the patient had an extreme elevation of procalcitonin without signs of bacterial infection."*, was not classified as a challenge, but an expert annotator did identify the issue of having a strong biological indicator for a serious condition without a clear explanation for its elevation as problematic. We note that in multiple cases presenting the model with the context aids with these issues. For instance, in the above example, when providing the context which includes a reference to the the risk factor, the prediction flips to the right call. However, as discussed in §A.5, context can also add noise in some cases. This observation led to our Context Slice+Combine approach described in §A.6.

The second biggest cause for false positives is sentences that provide a general description of a condition rather than a challenge. An example is *"The location of the headache might vary depending on which sinuses are affected[...]"*. Such texts are tricky since they are essentially facts about a condition rather than a description of an explicit challenge (e.g., the headache may be trivial to treat). To make the distinction clearer, consider the text *"Colitis is a chronic digestive disease"*. It presents a definition of a disease, and not an instance of an explicit problem one needs to address.

The second biggest cause we observe for false negatives is sentences that mention *partial* solutions that can mitigate a problem. For instance, *"[...] the cellular apoptotic process is immediately triggered as an innate defense mechanism in response to infection, but is abruptly suppressed during the middle stage of infection."*. In this example a defense mechanism is mentioned that can mitigate the problem, but the challenge nonetheless remains. Combined, the above error causes account for roughly 2/3 of the false positives and negatives. Labelers were provided input on how to deal with these issues in annotation, but since they are nuanced, models may require more data or sophistication to classify them correctly.

**Directions** The most common cause for error in Direction is the identification of future action items which are general vague suggestions as directions. For example, *"In agreement with the authors of that study, we believe that communication between laboratory specialists and clinicians should be intensified and improved."*. The example suggests a policy or action that does not constitute a research direction or hypothesis. This error accounts for 60% of the false positives.

In terms of recall errors we find that sentences that suggest a hypothesis or other more implicit research directions account for roughly 50% of the errors. For example *"Persistent viral shedding may indicate different levels of virulence, host immune response and infectiousness"*. The guidelines stipulate that these should be positive since researchers need to verify these directions, and indeed in most cases they are correctly identified, but some instances cause false negatives.

The rest of the errors in challenges or directions were anecdotal rather than systemic.

**A note on expected errors in the downstream tasks.** Our data contains an over-sampled proportion of tricky keywords (e.g., the word "hard" appears in both "hard material" and "hard task"), and thus we expect fewer errors in the search application task (see also Figure 3). In addition, in our downstream task of search we rank results by prediction confidence, and precision for high values of confidence is high even on our harder test set (Figure 2). Indeed, Figure 3 shows that predictions appear to have high overall accuracy in a sampled set.

### A.9 User Studies

We now explore user studies designed to evaluate our framework's utility. First, we explore whether our system can be helpful for quick discovery of challenges/directions. Second, we conduct a study with nine medical researchers working on COVID-19 treatment and research. In total our studies include 19 researchers and over 70 distinct search queries.

#### A.9.1 Search Engine

**Challenge and direction indexing** We build a search engine that indexes challenges and directions across the entire CORD-19 corpus up to and including August 2021. To build the search engine (see Figure 5 in the Technical Appendix for a screen capture), we first apply PubMedBERT to 550K papers, totalling 29M sentences. 180K of the papers are with full text, the rest are abstracts. We then clean poorly tokenized sentences, non-English sentences, very short sentences or texts with latex code. We classify the remaining sentences leaving 2.2M sentences — about 950K sentences with high-confidence predictions for at least one of challenge/direction and their surrounding context sentences. We select high-confidence sentences by using a threshold of 0.99 for both challenges and directions, using a thresholds leading to well over 90% precision at top-10% on our test set.

**Entity-based indexing** For each sentence in our set of 2.2M, we add another layer of indexing, by extracting entities and linking them to knowledge base entries. This allows us to partially group together all challenges or directions into "topics" referring to a specific fine-grained combination of concepts (e.g., *AI* + *diagnosis* + *pneumonia*), and facilitate entity-centric faceted search which is known to be useful in scientific exploratory search [13, 14].[13] We extract a range of biomedical entities and link them to a biomedical KB of MeSH (Medical Subject Headings) entities [24]. See Technical Appendix §A.10 for full technical details.

In the experiments that follow, we compare our system with a strong real-world system — PubMed biomedical search engine[14], a leading search site that clinicians and researchers regularly peruse as their go-to tool. While PubMed was not designed to find challenges and directions, no existing tool is; PubMed allows users to search for entities such as MeSH terms, is supported by a KB of biomedical entities used for automatic query expansion, and has many other functions — and as such is a strong real-world baseline.

#### A.9.2 Challenge/Direction Exploration

We recruited ten participants with education and experience in medicine, microbiology, public health, molecular, cellular, and developmental biology, biochemistry, chemical & biological engineering, environmental science, and mathematics. Participants are paid $50 per hour of work, comparing query results from our system and PubMed. Participants were given guidelines, which include definitions for research challenges and directions with simple examples.[15]

Each participant was given twenty queries, split into two sections for challenges and directions, respectively. For each query, participants were asked to find as many research challenges as possible in no more than 3 minutes. The total number of unique queries among the participants is 65. Some examples of queries used for the challenges section include "antibodies" and "inflammation, lung", with the paired entities being searched jointly; example queries for the directions section include "telemedicine" and "vaccines, technology". All queries were curated by a domain expert.

As seen in Figure 4, our system yielded a greater number of challenges and directions, on average, than the PubMed tool. Users found roughly 4.46 challenges and 6.43 directions per query using our system compared to the 2.24 challenges and 2.03 directions per query found using PubMed (p-value of .00192 for challenges and .000529 for directions using a paired t-test). For each participant we included 5 challenges and 5 directions that were overlapping across all participants, in order to control and compare between results for the same queries. We find that on average across users, 70.0% of the query results using our system led to a strictly larger number of challenges discovered than

[13] Other forms of challenge grouping, such as with embedding-based sentence clustering, are interesting to explore in future work.

[14] `https://pubmed.ncbi.nlm.nih.gov/`

[15] Full annotation guidelines are included in REDACTED FOR ANONYMITY.

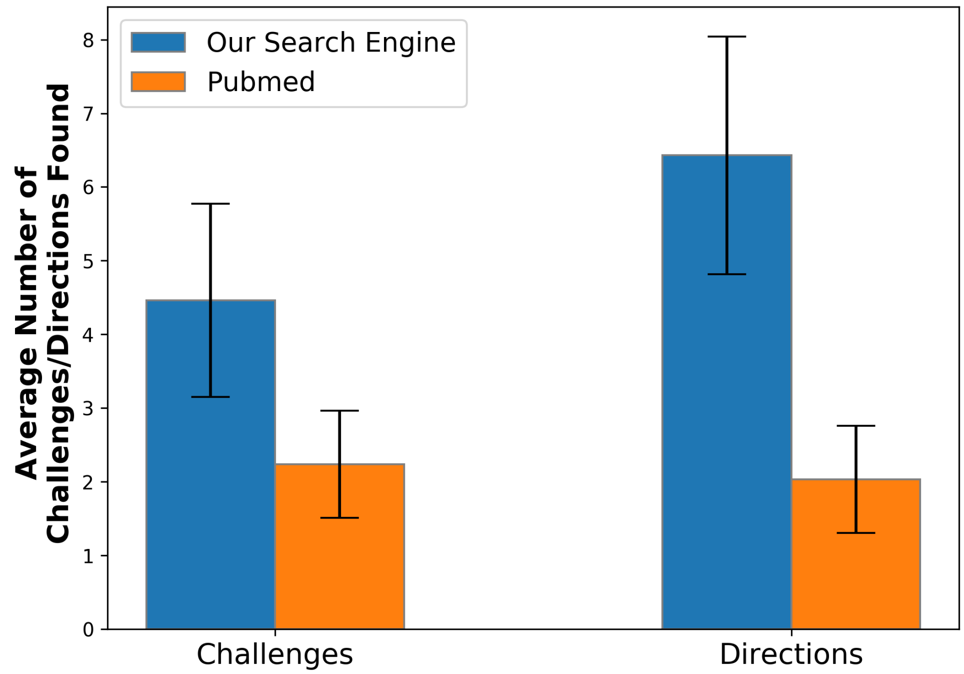


Figure 4: Study with researchers with diverse backgrounds. Participants using our search engine were able to find substantially more cases of challenges and research directions they considered useful than with PubMed. Error bars represent 90% confidence intervals.

| Metric | Chal./Dir. Search | PubMed |
|---|---|---|
| Search | **90%** | 48% |
| Utility | **94%** | 57% |
| Interface | **91%** | 68% |
| Overall | **92%** | 59% |

Table 5: Nine medical researchers expressed much higher satisfaction with our system (Chal./Dir.) than PubMed.

the respective query results using PubMed, and 22% were ties. For directions, we find a larger gap between the two systems, with 96.0% of the query results using our system yielding strictly more directions than PubMed, and 2% yielding ties.

### A.9.3 Evaluation with Medical Researchers

We now report on an evaluation of our search engine performed with nine medical researchers at a large hospital.[16]

**Study.** We recruited nine expert MDs with a wide range of specialization including cardiology, pulmonary and critical care medicine, gastroenterology and general medicine who are actively involved in clinical research both for COVID-19 and specialty areas, and each have over 1000 citations. Each expert completed randomly ordered search tasks (challenge/direction queries curated by an expert medical researcher; see Appendix §A.12) using both PubMed and our system. Experts using our UI viewed sentences and their contexts (previous/next sentences). In addition we also displayed metadata such as paper title, date, url. After all search tasks were completed for both systems, experts were given seven-point Likert-scale questions to judge system utility, interface, and

[16]In addition to the motivation discussed in the Introduction, see §Appendix A.13 for a more detailed example scenario where medical researchers need to search for challenges and directions.

search quality. Following [14], we use a standardized Post Study System Usability Questionnaire (PSSUQ) [21], widely used in system quality research, and added questions designed to evaluate search and exploration utility: *overall search accuracy*, *results that are not only relevant but interesting or new*, *finding papers interesting to read*, and *ability to understand and judge each individual result quickly without additional context*. Each question is asked twice, once for PubMed and once for our system, leading to $15 \times 2 \times 6$ = 180 responses.

**Results.** Table 5 shows the average Likert scores (normalized to [0%,100%]) across all questions and users for our system and PubMed. We group questions by three types for brevity. The results show that the medical experts strongly prefer our search engine to PubMed (overall average of 92% vs. 59%, with non-normalized scores of 6.42 vs. 4.14). On average across all questions, the majority of the nine MDs assigned our system a higher score than PubMed, at an average rate of 85% per question. When considering ties, the average rate is 92%. We found that our system significantly outperformed PubMed across all questions (Wilcoxon signed rank test p-value is significant at $5.409 \times 10^{-6}$). These results further resonate in light of the experts' strong familiarity with PubMed and the bare-bones nature of our UI.

### A.10 Entity-based Indexing

We employ the SciSpacy library [29] to extract entities using five different NER models: one trained on MedMentions [28] (a dataset with general mentions of UMLS [3] entities covering a wide range of concepts), and four trained on more specialized sources (CRAFT [1], JNLPBA [18], BC5CDR [23], BIONLP13CG [19]). Each entity is then automatically linked to a biomedical KB of MeSH (Medical Subject Headings) entities [24] using SciSpacy's entity linking functionality that performs character-trigram matching on MeSH entity names and aliases. We filter for high-confidence linked entities,[17] and for entities that appeared in at least 10 sentences, then selecting the top 30K unique entities to be indexed by our search engine. At search time, we match user queries to MeSH aliases with an autocomplete dropdown for users to select from as they type. After one entity is selected, the user can search for more from a narrower list of entities that co-occur with it.

### A.11 Search UI

Figure 5 shows a screen capture of our search user interface.

### A.12 Expert queries

Below are examples for our expert queries:

- **Find problems/limitations/flaws** related to COVID-19 and each of (1) *hospital infections*, (2) *diagnosis*, (3) *vaccines for children*, (4) *probiotics and the gastrointestinal tract*.
- **Find directions/hypotheses/potential indications** related to COVID-19 and each of (1) *mechanical ventilators*, (2) *liver*, (3) *artificial intelligence*, (4) *drug repositioning*.

### A.13 Medical Search Scenario

In addition to the motivation discussed in the Introduction, we briefly discuss in more detail how a search engine for challenges and directions could help medical researchers when conducting literature reviews. Many research ideas come to MDs with a challenge they perceive during clinical care. If they are unable to find a solution to the problem based on prior experience, they then search for the available scientific literature for possible guidance. When no sufficient guidance is found, research projects are often commenced, starting with literature search which often involves understanding and mapping out associated challenges and directions to help with formulating research questions. Physicians and trainees still spend a significant amount of time doing this form of literature search for fine-tuning their research question. If such a process can be simplified with automation, it could potentially cut down the time and effort needed to formulate and narrow down research questions.

[17] Using a threshold of 0.9.

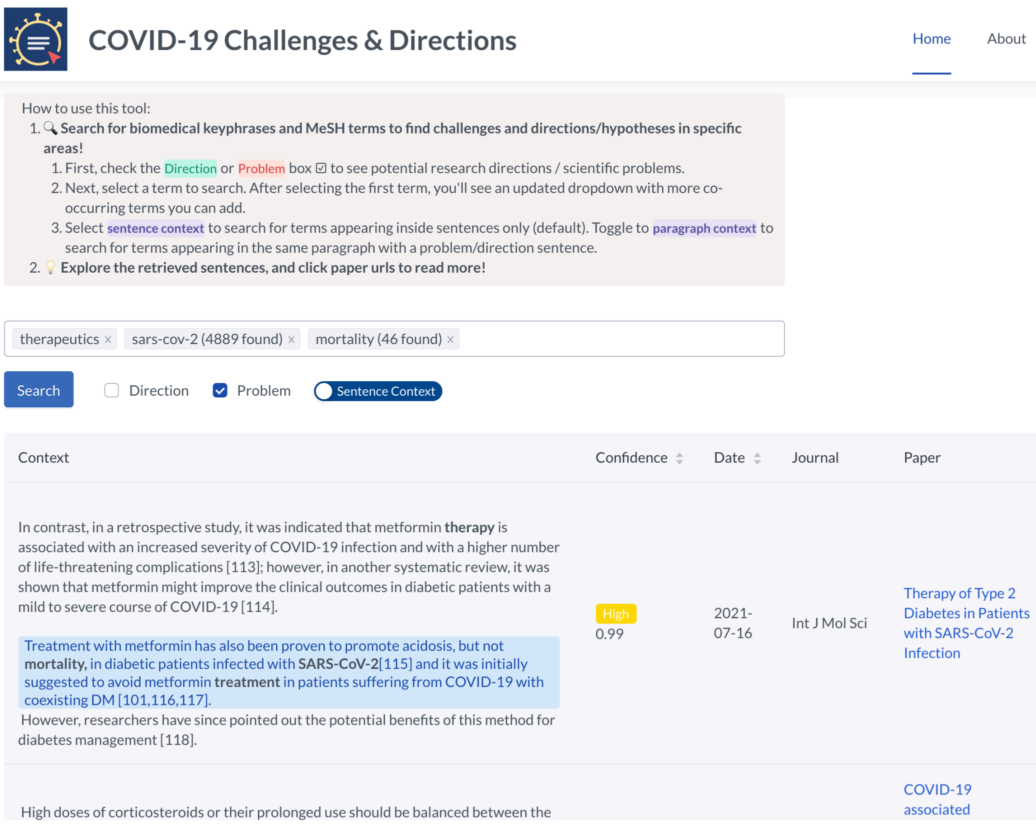


Figure 5: Screen capture of our search user interface.