School of Computer Science and Informatics

Master of Science

Intelligent Systems

COVID-LEAP - A specialist cockpit with state-of-the-art search and heuristics for insight into COVID-19 academic research and the vaccine development landscape

Jon-Paul Boyd (P17231743)

June 2021

Supervisor: Professor Aladdin Ayesh

ABSTRACT

As of February 2021, the COVID-19 coronavirus has impacted global society significantly, with over 2.5 million related deaths (Johns Hopkins University COVID-19 Map, 2020) and 255 million job losses (UN News, 2021). The pandemic poses many challenges to the whole spectrum of society, including the scientific community's ability to research an explosion in related academic articles. A second challenge is the accurate and timely tracking of the lightning-fast COVID-19 vaccine development competitive landscape. Insight into current market state and emerging vaccine platform trends support pharma and biotech business leaders in making informed, critical decisions around the financing, formulation, development, evaluation, scale, and manufacturing of their vaccine programmes. A third challenge tasks software engineers with building solutions able to manage complex, often unstructured, and high volume data from disparate sources. When modelled and analysed, it has real potential in delivering efficiency gains and decision support value to researchers and decision-makers.

This project develops a solution with several key features. State-of-the-art BERT (Bidirectional Encoder Representations from Transformers) based neural networks model an extensive collection of COVID-19 academic corpora. Such language modelling supports improved discovery of most meaningful texts according to a biomedical search query. Latent Dirichlet Allocation (LDA) organises the corpus into topics of keywords for summarisation and filtering. Data mining, including heuristics with regular expressions, extracts critical vaccine clinical trials information from several unstructured sources. This supports efficient, visual analysis of the highly competitive vaccine landscape by clinical domain experts and associates academic research articles by referencing trial identifiers. Bibliometrics measure author and article importance.

Empirical evaluation showed dense ANCE transformer model outperforming the lexical BM25 method and "*go-to*" PubMedCentral search site in most tasks. There is a consensus between automated synthetic benchmarking and assessment by our medical expert that ANCE was the strongest model.

The application is Cloud-hosted on the Microsoft Azure platform, facilitating access to a robust, scalable engineering framework suited to deconstructing the problem into pipeline steps that include data acquisition, pre-processing, modelling, evaluation and deployment.

ACKNOWLEDGMENTS

It was a great privilege to work closely with Jonathan Wagg during my 6-month internship at AC Immune. I could not have asked for a better mentor during my foray into exploring the COVID-19 knowledge landscape with NLP. With your kindness, generosity of time, encouragement, and sharp mind, I believe I arrived at a much better place with this project.

I'm very grateful to Anna Pellaz for her patience as I took baby steps into the world of biomedical academic research. In addition, I'm grateful to Ilker Esener and Mark Danton for being so supportive of my efforts and believing in me. Finally, to Marie Kosco-Vilbois, thank you for taking that initial chance on me.

For J and G, love always and forever

For N, thank you for your love and tolerance

For A and D, my shining lights, how blessed I am!

For MC, the last years impossible without you!

Contents

List of Tables

List of Figures

1. Introduction

In 1966 Scottish virologist June Almeida coined the phrase *coronavirus* for a group of viruses that under an electron microscope resembled a solar corona due to their bristly, spikey glycoprotein structure (Chorba, 2020). A zoonotic virus, spreading from animals to human hosts, they were initially considered only a moderate risk to humans (Cyranoski, 2020). That was until 2003, when an outbreak of a new coronavirus, SARS-CoV, first occurred in mainland China before migrating to Hong Kong and then globally, causing 774 deaths (Lanying Du et al., 2009). Then, in 2012, Jordan first recorded cases of yet another novel coronavirus, MERS-CoV (Middle East Respiratory Syndrome), which spread to the rest of the Middle East then beyond to 27 countries with 858 known deaths (Zhu et al., 2020). The reported emergence of the SARS-CoV and MERS-CoV viruses indicated that novel, higher-risk coronaviruses can jump from animals to human hosts to drive outbreaks of deadly infectious diseases.

Fast forward to December 2019. The SARS-CoV-2 virus (Severe Acute Respiratory Syndrome type 2) is reported to originate out of Wuhan, China. Often more widely associated with the disease it causes, COVID-19 (Coronavirus disease 2019) has had a devastating global impact. Approximately one in five contracting COVID-19 become seriously ill, primarily with breathing difficulties (WHO COVID-19, 2020). Those at higher risk include the elderly, individuals undergoing chemotherapy, radiotherapy or have existing severe lung conditions, the obese, and pregnant (NHS.UK, 2020). Tragically, confirmed deaths from the pandemic have surpassed 2.5 million. For further context, in the U.S. alone, by February 2021, the number of deaths passed 500,000, more than the death toll from World War One, World War Two, and the Vietnam war combined.

The pandemic has disrupted every aspect of life, testing the resilience of healthcare systems while severely affecting food supply chains and transportation systems. Many millions have been made unemployed. Over a billion are at severe risk of livelihood loss (WHO News, 2020). Mental health continues to be significantly impacted. Loss of or concern for loved ones, social isolation, over-consumption of news media reporting the crisis, and lack of exercise are just some of the causes of higher stress levels, anxiety, depression, and substance abuse (CDC, 2020).

Alongside frontline healthcare workers are scientists battling COVID-19, working tirelessly on developing vaccines and antiviral drugs to fight infection as a way out of this crisis. These are unprecedented times, with pharma and biotech progressing vaccine developments and overlapping trials at lightning pace. There is close collaboration with

regulatory authorities and ethics committees to validate vaccines for emergency use and globally distribute to priority populations (WHO Covid Timeline, 2020).

However, scientists researching the pandemic face a big issue, namely the explosion of literature in the knowledge landscape. A complex interaction of factors, including the biological (low child deaths, autoimmune disorders), clinical (vaccine trials), psychological (post-traumatic stress disorder), social (schooling, inequality), political (institutional trust, polarisation), and economic (job losses, evictions) are contributing to a noisy information crisis. Knowledge coming into the public domain from many sources is ever-increasing. One figure states over 3000 related, peer-reviewed new papers published daily (Lee et al., 2019), while another indicates the doubling of papers every 20 days, among the biggest explosions of scientific literature (Brainard, 2020).

Given that such knowledge reflects the evolving pandemic, the changing research topics of interest, and the complexity of scientific discovery, no researcher has the time to evaluate all available information manually. Critical knowledge will be missed. Researchers are making pragmatic decisions faster than ever with limited analysis in the face of overwhelming uncertainty.

This project presents COVID-LEAP (Literature Exploration Analysis Platform) to support the exploration of COVID-19 knowledge and alleviate clinical researcher cognitive overload. It addresses the leveraging and connecting of unstructured academic articles and clinical trials knowledge, applying natural language processing (NLP) and deep learning (DL) methods at scale to support clinical researchers and business leaders when they are making quick decisions under extreme time pressure. Although commercial knowledge providers offer pandemic landscape dashboards and curated article collections, they are only available at cost to subscribers. With COVID-LEAP, our first focus is on state-of-the-art information retrieval (IR) methods explicitly calibrated for searching the CORD-19 (COVID-19 Open Research Dataset, CORD-19, 2020) archive of coronavirus literature. It intelligently surfaces and organises scholarly article content most relevant to an information need, providing direct answer snippets as evidence. This will help researchers more efficiently identify sources of knowledge directly relevant to their research. Our second focus is on providing the latest summarisation of critical vaccine development landscape knowledge.

Automation removes the potential for human errors in compiling and aggregating clinical trial knowledge. An additional benefit is the release of key resources from a very labour-intensive task. Furthermore, the lag between publication of trial updates and availability to the consumer is reduced. Various methods are applied to knowledge capture, mining, and visualisation, resulting in a tool allowing researcher interaction for discovery, insight, and pattern identification. It will be architected as a cloud-hosted application on Microsoft Azure.

1.1 Research Objectives

Having recognised the problem of extracting relevant information to clinical questions from a large, dynamic corpus of COVID-19 knowledge, this project deconstructs the issue into the following research problems. Note the organisation of sub-objectives as hypothesis (H*n*) grouped by research question (RQ*n*).

1. **RQ1 Feasibility**: Can a search strategy including semantic models, and vaccine tracking related to the pandemic, be delivered through to execution?

   H1 *Discover the available resources providing the foundational knowledge*
   H2 *Identify computing resources necessary for search over large volume corpora*

2. **RQ2 Functionality**: Can NLP support the medical community with the intelligent search of the pandemic literature corpus and insight into the associated vaccine landscape, enabling better next-step research decisions?

   H3 *Consider rapid prototyping of main features*
   H4 *Verify NLP models answer questions with the most relevant literature, evaluated with extrinsic human expert criteria, including questions of differing formats*
   H5 *Can data mining be used to develop insights, findings, patterns, emerging trends, and tables of facets from the corpus?*
   H6 *Ascertain if the application reduces the overall time spent on COVID-19 research*
   H7 *Verify the tool is user-friendly to medical researchers*

3. **RQ3 Technical Performance**:
   H8 *Verify query response over full corpus completes with acceptable latency*
   H9 *Determine intrinsic performance, assessing models with benchmark suites*

1.2 Personal Development Objectives

The author has a passion for A.I. in health sciences, having developed a fuzzy logic system classifying breast cancer tumours that outperformed neural networks and other machine learning (ML) approaches. As a father living in a changing world with COVID-19, the described problem domain allows for doing something good, worthy, and insightful. The project has provided the opportunity for blending technical know-how with clinical expertise through close collaboration with biotech specialists. While supporting the scientific community with the development of research tools, an understanding of their research objectives, methods, and terminology has been developed.

The project also presents the researcher with the opportunity to further understand NLP, including how BERT-based transformer models may address the complex problem of semantic search over a specialised corpus of knowledge and how their results compare with other language modelling approaches. Additionally, it allows the researcher to gain experience in designing and deploying cloud-hosted solutions that are secure, performant, and cost-conscious.

2. Literature Review

2.1 Introduction

This chapter reviews prior work relevant to extracting information from unstructured knowledge sources in the biomedical domain space to support clinical research. It will provide insight into the evolution and current state-of-the-art (SOTA) IR methods that strive to deliver improved performance to satisfy the information needs of researchers. It will help synthesise themes relevant to developing a system supporting COVID-19 researchers.

2.2 Review Methodology

We now present our methodology for the execution of our literature review. Qualitative methods select papers on previous work to understand the nature of mining information and answer natural language questions from searching a corpus of unstructured data in a specialised domain. The following search criteria, incorporating concepts of interest, derived an initial list of English-language articles, sourced from research tools including PubMed (PM, 7 results), Semantic Scholar (1370 results), and Google Scholar (2110 results):

```
(embedding OR representation) AND (mining OR retrieval OR extraction OR bioinformatics OR "semantic search" OR indexing) AND (biomedical OR clinical) AND ("natural language" OR nlp OR "artificial intelligence" OR "neural network" OR bert OR transformer OR tf-idf OR bm25) AND (mers OR sars OR sars-cov OR covid19 OR coronavirus)
```

The scope of this review focuses on the application of NLP methods for the surfacing of relevant articles in mining tasks, including IR. Therefore, further screening iterations using the title and abstract discount articles helped assemble a list of article exclusion criteria as follows: subcellular localisation, biological sequencing, clinical notes and text for patient diagnosis, image processing and radiography, drug discovery, drug repurposing, COVID-19 test centers, COVID-19 diagnosis, coronavirus classification, contact tracing, outbreak and incubation forecasting, pneumonia, phenotyping, nursing, convolutional neural networks, genetic algorithms, open information extraction, ontologies, graph mining, social media.

Such an extensive exclusion list significantly reduces the volume of articles to be reviewed in-depth. Importantly, it highlights the broad diversity of academic articles within the domain and demonstrates the typical "*overload*" issue any researcher faces. Furthermore, articles published before 2018 are ignored. For an unbiased approach, we exclude existing literature reviews and surveys.

Following deletion of duplicates, the complete text from the final collection of 36 academic papers is analysed and grouped into themes relevant to developing a COVID-19 domain IR solution. For the final collection of papers to be reviewed, we used Zotero, a literature management application, to organise paper references and metadata, clustering articles by theme with tags. After refinement, a final set of 3 main themes emerged as follows: "a*pplications supporting the fight against COVID-19*", "*language modelling with neural networks for search", and* "*information extraction from literature with alternatives to BERT*", each now discussed in turn.

2.3 Applications supporting the fight against COVID-19

On the 30th January 2020, the WHO declared the COVID-19 outbreak a Public Health Emergency of International Concern (PHEIC). The disease was then characterized as a pandemic on 11th March 2020 (WHO, 2020). The escalating impact COVID-19 was having on humanity sounded a global scientific community call to arms, and the pharmaceutical industry responded spectacularly.

COVID-19 is an inflection point in the trajectory of vaccine development. Historically, a vaccine's initial research phase lasts between one and five years, followed by clinical trials taking between fifteen to twenty years before approved licensing for use (Artaud et al., 2019). On 11th April 2020, the WHO published their first draft landscape of candidate vaccines, which showed 67 already in the preclinical phase of development, testing for safety and protection against the disease in animals, typically mice and primates. This first draft landscape also confirmed three vaccines in clinical evaluation on human subjects. The Moderna (RNA platform) candidate (NCT04283461) entered phase 1 testing for safety just 69 days after the identification of Sars-CoV-2 as the cause of the outbreak (Kim et al., 2020), highlighting the many lessons learned from SARS-CoV that were exploited to expedite the development of the SARS-CoV-2 vaccine.

By comparison, the first vaccine for Mers-CoV (NCT02670187) and Sars-CoV (NCT00099463) took 22 and 24 months respectively to reach clinical trials. In starker contrast, an Ebola vaccine, developed by the Public Health Agency of Canada and sub-licensed to Merck (NCT02344407), took five years to get approved following the 2014 outbreak in West Africa (Beigel et al., 2017, FDA, 2020).

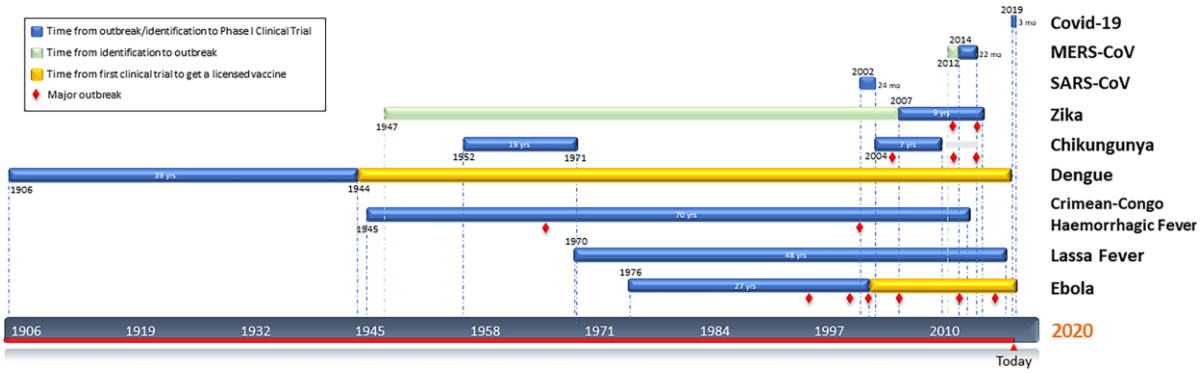

Fig. 1: New emerging diseases vaccine development timeline (Kim et al., 2020)

To support the expedition of COVID-19 research, Microsoft, in collaboration with several partners, heeded the call with the release in March 2020 of the COVID-19 Open Research Dataset (CORD-19, 2020). A scientometric analysis of the dataset claims it provides a complete corpus for research of the coronavirus family, broadly centering on virology, biology, and public health. However, it warns researchers that "*the dataset includes a large number of publications whose relevance for COVID-19 and coronaviruses research needs a more careful assessment, and some of which may be of limited relevance.*" (Colavizza et al., 2021).

The use of CORD-19 is widespread and applied in solving various problems, with systems primarily intended for consumption by two distinct user groups - the biomedical researcher and the non-scientific general public. Several studies develop systems explicitly focused on a question-answering task by responding to questions with relevant extracts from corpus articles. A web-hosted Q&A chatbot guides the non-scientific general public seeking healthcare information (Oniani and Wang, 2020). More common themes are solutions explicitly targeting the expert, both in biomedical research and healthcare communities. CoronaCentral is a system that complements the search of the literature with categorization and metrics (Lever and Altman, 2020). Confirming one problem of the aforementioned scientometric analysis, they estimate 26.2% of dataset articles are neither full papers nor original research, and hence place importance on identifying publication type. To that end, they take a supervised learning approach, using an annotated document subset, for multi-label classification of documents with labels including *Clinical Reports, CDC Weekly Reports (Centers for Disease Control), Book Chapters, Drug Targets, Therapeutics, and Vaccines*.

Understanding both the pandemic and associated scientific research is highly dynamic, a high-level overview of current and historic research trends is insightful. Gupta et al. (2020) propose a solution that identifies the level of research popularity by topic, learning the underlying hidden topic probability of each corpus document by statistical analysis of word distribution. Using publication date, they group articles by week to visually represent temporal trends by topic. They found that using only the abstract was sufficient to convey the articles' underlying essence while also reducing compute resource requirements, allowing for additional optimisation cycles. They determined the optimal number of topics as 50.

Dong et al. (2020) present a similar solution for highlighting hotspots of scientific research interest. As per Gupta et al., only the article abstracts were used. However, in this work, only eight topics were considered optimal: clinical characterization, pathogenesis research, therapeutics research, epidemiological study, virus transmission, vaccines research, virus, diagnostics, and viral genomics. Additionally, they found that by subsetting the corpus into SARS-CoV-2 and other coronavirus types, COVID-19 research places *"more emphasis on clinical characterization, epidemiological study, and virus transmission. In contrast, topics about diagnostics, therapeutics, vaccines, genomics, and pathogenesis only accounted for less than 10% or even 4% of all the COVID-19 publications, much lower than those of other CoV infections*". As this article was published on 30th March 2020, we suggest a significant topic shift has now occurred, especially in the area of vaccines and therapeutics, given that a significant 79 preclinical and 182 clinical candidates vaccines are in development as of 5th March 2021, according to the WHO (Draft Landscape of COVID-19 candidate vaccines, 2021).

Ebadi et al. (2021) also characterise the research landscape by literature topic. Again, their use publishing date to structure topic classification with a temporal granularity by week and month. However, this system differs from the previous approaches in two ways. First, they applied stricter recency constraints, excluding articles published before 2019, which we assume is to filter out coronavirus variants prior to Sars-CoV-2. Second, their topic model includes article title in addition to the abstract, arguing titles contain informative keywords not necessarily present in the abstract. They note that keywords including "*covid*" and "*coronavirus*" add little informational value and exclude them. This work determined the optimal number of topics as seven. As topic modelling generates a collection of top terms per abstract topic, human experts manually labelled each topic for a concise understanding of the underlying meaning. They identified genomics as a continuous focus topic while oncology, personal protective equipment, and high-risk groups attracted growing attention.

In a further development of topic modelling, Bras et al. (2020) provide researchers and research strategists with a view of the landscape by abstracting the literature into a novel 2-layer hierarchical topic model of 400 sub-themes associated with one of 50 main themes based on document similarity. Confirming a trend, the publishing date is again used to visualize the distribution of topics over time. Their work identifies increased research in social distancing and highlights their ability to track pandemic trajectory through territories from the literature alone.

From the reviewed sample of related works, there is a significant difference in opinion regarding what constitutes an optimal number of themes when topic-modelling the same corpus. Although we acknowledge methods including coherence score, principal component analysis (PCA), and the "*elbow*" method for deciding topic number is valid, the decision should not be informed solely by mathematical science. Selecting an "*appropriate*" number should reflect how they add value for the end-user. Fifty topics may be too granular, overlapping, and overwhelming if presented as a multi-value facet list for document filtering. However, fifty topics may be useable and insightful when representing with highly visual bubble maps. We recommend a strategy that includes both an intrinsic, quantitative evaluation of the optimal topic number balanced by extrinsic opinion of domain experts and end-users.

CAiRE-COVID (Su et al., 2020) is a Q&A system that answers questions with highlighted paragraph snippets as evidence. Additionally, it generates an extractive summary by concatenating wholemeal, without any rewrite, the most statistically important paragraphs relevant to the question. (Esteva et al., 2020) also apply summarisation to their NLP pipeline for Q&A; however, their approach differs in two ways. First, in contrast to an extractive summary, the authors' CO-Search solution uses an abstractive method to rewrite a paragraph text into a shorter summary entirely, phrased such that the generated sentences may not appear in the source paragraph. Second, each paragraph summary is not presented to the end-user but instead used to rerank the query result document list generated by BM25 and TF-IDF lexical retrieval models. Such an approach looks to benefit from a fusion of term and semantic search, reducing matched paragraphs to a shorter semantic form more symmetrically aligned in length with a typical question format

Another literature Q&A solution, COVIDScholar (Trewartha et al., 2020), compliments the CORD-19 corpus with additional data sources. They use LitCovid (Chen et al., 2021), a corpus purely focused on COVID-19, complemented with articles from preprint servers such preprints.org, to curate a collection to include works from other domains including economics, psychology, and humanities. Given the volume and frequency of new pandemic research literature, the authors acknowledge the increased importance of preprint servers

with "*widespread usage for the first time in many fields, most prominently biomedical sciences*". However, they share a common concern that bypassing peer review allows for lower quality work with flawed results to be published. They further observe that papers related to virology, published before the COVID-19 outbreak, score highly in search. They argue this is due to 79% similarity between SARS-CoV and SARS-CoV-2 genome sequence identity and believe that "*search results are more likely to contain relevant information, even if it is not directly focused on COVID-19*". This contrasts with the approach of (Ebadi et al., 2021), their topic modelling system excluding documents before 2019.

Beyond the academic research lab, industry giants have leveraged their access to engineering expertise and computing resources to make available commercial products for search and mining of the CORD-19 dataset, with Amazon Comprehend Medical and Google's COVID-19 Research Explorer.

Several studies extract important biomedical named entities (proteins, genes, viruses, DNAs, RNAs) from the corpus to enrich search and discovery. Lee et al. (2020) employ named entity recognition (NER) to build a metadata index of entities associated with each article. A phrase indexing system ingests the complete corpus. Their hybrid engine combines results from a search of both indices when compiling a results list. The authors raise the topic of document recency, proposing a bias towards older articles that document coronavirus research into Sars-CoV and Mers-CoV. They argue that such papers "*do not explicitly mention COVID-19, but may nevertheless provide important clues or information that may help the understanding of COVID-19*". This once more highlights opposing theories to the value of older publications. We are of the opinion that all articles are of value; however, newer documents should be weighted advantageously, given the rapidly changing pandemic landscape leads to critical knowledge discovery found in the very newest of publications, quite possibly preprints that circumvent the lengthy peer-review process.

In a similar NER solution, Sohrab et al. (2020) provide a web-based system for clinical researchers to annotate the corpus text by extracting the biomedical entities and linking them to an established, centralised collection of biomedical vocabularies (Unified Medical Language System). This technique serves to disambiguate terms, for example, by linking the entity SARS-CoV-2 with the type virus and not vaccination, further supporting the correct classification of the document as related to virology and not therapeutics. The researchers behind CORD-NER have used the CORD-19 corpus to create an NER dataset of entities explicitly related to COVID-19 studies that include viral proteins, evolution, materials, and immune responses (Wang et al., 2020). They hope their published work can improve downstream science with a better understanding of the virus through refined annotation of entities within the literature. We argue that this fine-tuned NER solution can only improve entity tagging accuracy given the domain's specialized nature.

Reviewing the literature identifies advances in establishing relationships between biomedical entities with knowledge graphs. Tyagin et al. (2021) introduce the first COVID-19 based hypothesis generator to discover novel implicit connections between papers. Nodes are created for all corpus sentences and extracted entities, with semantic similarity defining edges between them. They show one test predicts potential compounds applicable in treating the disease and found that "*tetherin restricts the secretion of SARS-CoV-2 viral particles and is downregulated by SARS-CoV-2*".

Similarly, Basu et al. (2020) present ERLKG (Entity Representation Learning and Knowledge Graph) to discover associations between proteins, diseases chemicals present in the CORD-19 corpus for clinical decision making. The COVID-19 Knowledge Graph (Wise et al., 2020) identifies biomedical entities, including diseases, genes, mutations, compounds, drugs, and their relationships. However, its additional strength is the inclusion of authors behind the research, so when the knowledge representation is queried heuristically, it allows for the identification of influencing domain experts and articles of high relevance.

Answering questions of a more closed-ended nature, researchers also mine the literature corpus to address narrowly-focused question-like problems that include the correlation between temperature, humidity, and the spread of the disease, extracting text snippets from the literature that directly answer the question (Sastre et al., 2020). Similarly, Ahamed and Samad (2020) present a more topic-focused knowledge graph to reveal relationships between drugs, diseases, pathogens, and molecules within the specific topic areas of transmission, drug types, and genome research. Their work identifies animal hosts worthy of zoonotic pathogen study and varying degrees to which various drugs play a role in fighting the disease.

2.4 Language modelling with neural networks for search

NLP is the science of building systems to process and understand human language. However, computers do not understand text directly, so an initial transformation into real numbers allows search and analysis with mathematical tools and machine learning. Various methods are available to transform words, sentences, paragraphs, or complete documents into vectors of numbers, known as embeddings, including Bag of Words (BoW) and transformer neural network architectures. While Chapter 5 covers embedding in more detail, this section's primary focus is how different transformer embedding approaches vary in their ability to capture the complex semantic properties of biomedical literature to serve downstream NLP tasks, including IR.

A neural network is loosely inspired by a biological model that combines multiple inputs from senses to generate output signals. In 2003 Bengio et al. introduced the first language model using a neural network to learn the distributed representation of words. Their experiments with a feed-forward architecture addressed the curse of dimensionality by demonstrating that a vocabulary of 17,000 words could be represented by as few as 30 features. They also showed that their sentence vector representation method effective in comparing semantic similarity and predicting word sequence probability. In 2010 a Recurrent Neural Network (RNN) model with superior performance in speech recognition tasks was presented (Mikolov et al.). In 2012, a Long-Short Term Memory network architecture trained on an English and French corpus was shown to surpass a standard RNN in speech recognition tasks (Sundermeyer et al.).

More recently, the transformer, a simplified neural network without recurrence and convolution operations, demonstrated performance exceeding that of RNNs and LSTMs in NLP tasks, specifically English to German and French translation, according to the seminal paper by Vaswani et al., 2017. When embedding text for dense vector representation, a self-attention layer supports the encoder by looking at other words around the word being encoded, in this way learning context and understanding.

In 2019, Google disrupted the field of NLP by open-sourcing BERT (Bidirectional Encoder Representations from Transformers), a pre-trained neural network architecture based on transformers that achieved state-of-the-art results in eleven benchmarking language tasks (Devlin et al., 2019). Two key characteristics of BERT facilitated leading performance. The first is model pre-training on vast amounts of data to learn the nuances of natural language; BERT-base using the BooksCorpus of 800M words, BERT-large on English Wikipedia (2500M words). The second breakthrough characteristic is the model's ability to learn word context based on words to both the left and right of a given word, contrasting with traditional English-text learning fixed sequentially left to right. In the context of a

search task, this feature enables the model to understand a query's intent to generate more relevant results. BERT models have become a popular choice for solving NLP tasks relevant to this thesis, including semantic search, which goes beyond traditional search that only finds documents based on keyword match by locating synonyms of the query within documents. For greater detail on the BERT architecture see Chapter 5.

Searching the literature for applications supporting the fight against COVID-19 uncovers many using BERT-based language modelling. Examples include a system for extracting named entities from text and linking them to the unified medical language system (UMLS) knowledge base (Sohrab et al., 2020), entity relation extraction (Basu et al., 2020), and several tackling the question and answering task (He and Bakhtiari, 2020, Nguyen et al., 2020, MacAvaney et al., 2020, Farokhnejad et al., 2021). These examples use the CORD-19 corpus of biomedical literature, which naturally exhibits different linguistics characteristics than texts available in the general domain, such as Wikipedia articles and Google News. Given that the vanilla pre-trained BERT models learn the nuances of language from such generic corpora, and biomedical semantics can incorporate explicit naming of genes, proteins, drugs, and other entities, the following subsections consider fundamental questions around using BERT-based models for modelling biomedical domain unstructured text.

2.4.1 How does BERT perform in the specialised biomedical domain?

There exists a point of view that "*word embeddings trained from biomedical domain corpora do not necessarily have better performance than those trained on other general domain corpora. That is, there might be no significant difference when word embeddings trained from an out-domain corpus are employed for a biomedical NLP application*." (Wang et al., 2018). However, most reviewed work contradicts this position, suggesting that despite the reported state-of-the-art performance in NLP tasks, vanilla BERT, which comes pretrained on generic corpora, underperforms when tasked with mining biomedical literature due to domain shift (Wise et al., 2020). In support, Wang et al. (2018) show embeddings trained on biomedical literature can better capture the semantics of medical terms than embeddings trained on generic texts.

We suggest there remain gaps in understanding the nuances of biomedical semantics and how types and structures of questions asked by clinical researchers relate to the underlying corpus queried for knowledge. This disconnect is not limited to technologists within academia, as an evaluation of systems searching the CORD-19 corpus from Amazon and Google concluded poor performance attributed to lack of domain-tuning and warned that "*deploying state-of-the-art methods without event-specific data may be dangerous, and in the face of uncertainty simple may still be best*" (Soni and Roberts, 2020).

Pretraining or fine-tuning transformers with specialized domain corpora is therefore generally accepted to improve performance in domain-specific NLP tasks. Pretraining builds transformers entirely from scratch, while fine-tuning transfers existing learning from a pre-trained transformer model and refines it further using additional corpora, typically associated with the problem domain. For example, Wang et al., 2018 emphasizes the benefits of transformer pre-training closer to the problem domain, finding the EHR (Electronic Health Record) dataset better captures clinical language, while MedLit, a dataset of medical research, is better able to find semantically relevant medical terms.

Several scientific variations of the original BERT model are available. Lee et al. (2019) introduced BioBert (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining), taking the BERT architecture building blocks (weights), then pre-training on over one million PM abstracts and PubMed Central (PMC) full-text articles for a model aimed at clinical research. Claiming pre-training on biomedical corpora "*helps it to understand complex biomedical texts*", they show 12% improvement over vanilla BERT on biomedical Q&A. The COVID-19 knowledge graph application Wise et al. (2020) leverages the biomedical pre-training of BioBERT and further fine-tunes it with the NCBI (National Center for Biotechnology Information) disease dataset. He and Bakhtiari (2020) propose a system using the CORD-19 corpus to answer biomedical questions, using BioBERT further fine-tuned with the SQuAD and COVID-QA datasets for Q&A language tasks.

Another scientific variant is SciBERT (Beltagy et al., 2019), pre-trained on over one million articles from Semantic Scholar, a corpus of full-text papers primarily from the biomedical domain (82%) but includes 18% from the computer sciences which, we propose, would support modelling and knowledge extraction on research at the union of these two domains, for example where computational modelling advances biomedical research. The authors claim superiority over vanilla BERT on biomedical tasks. COVIDScholar, another search application, explicitly selected off-the-shelf SciBERT for classification of papers by discipline, appreciating its "*broad, multidisciplinary training corpus*" and "*state-of-the-art performance on the task of paper domain classification*" (Trewartha et al., 2020).

In using a model pre-trained on generic corpora such as BERT, Trewartha et al. (2020) are concerned that "*the majority of general domain text is substantively different from biomedical text, raising the prospect of negative transfer that actually hinders the target performance*". They argue that pre-training from scratch is a superior strategy to fine-tuning when assembling a model for NLP tasks in the biomedical domain that does not suffer from the unavailability of training data and therefore does not need to lean on language learning support from large-scale generic corpora.

However, neural model selection and choice of pre-training or fine-tuning approach for the engineering practitioner is not clear-cut when considering that training BioBERT from scratch took 23 days on eight NVIDIA V100 GPUs (Lee et al., 2019). Generally, what is missing from the reviewed literature is an appreciation of what it takes to productionise such models, including the required compute resources for model training and inference, and the storage to persist large vector representations. Such requirements are often beyond the reach of the student, academia, and small enterprise. Furthermore, one should always consider how fast indexing and query response needs to be. Understanding these potential pitfalls can help determine BERT-based building block size and application of neural vector spaces for representation and comparison in search strategies.

2.4.2 How are BERT models applied in biomedical domain search tasks?

Creating a dense-vector embedded representation of all text (sentence, paragraph, document) from a large corpus like CORD-19 can take several hours. It is critical to minimise latency between asking the scientific question and presenting a ranked result list to provide the best possible user experience. We propose taking the approach used in CovidAsk, which answers natural questions related to COVID-19 (Lee et al., 2020) with semantic models. The authors encode the complete corpus once then persist the generated vector representation for real-time availability in later Q&A. Only the question requires encoding at runtime inference, using the same model vector space for corpus embedding, with a metric such as cosine similarity applied to find similar passages. It is timely to highlight that none of the reviewed articles explicitly considered whether the article abstract contains sufficient semantic information to answer most key terms and semantic-based questions posed by end-users. Using only the abstract might offer an opportunity to optimize systems by reducing the overall vector representation footprint during indexing and inference.

In reviewing transformer models applied to biomedical corpus search strategies, there is a trend for ensemble approaches whereby a domain-expert BERT-based model trained on relevant literature and a more traditional lexical-based key term retrieval model such as BM25 are combined. The review finds hybrids are of two types:

1. **a "*fusion*" type** where the question is asked over both the lexical and embedding vector index, with the final list the highest-ranking results from either index

2. **"*re-rank*" type** where the question asked once over the lexical index, and retrieved documents re-ranked for similarity with the question using the semantic model

An example application of the fusion type is Co-Search for Q&A (Esteva et al., 2020). This system represents title and document paragraph tuples with three vectorizors: TF-IDF, BM25, and SBERT (Siamese-BERT). Upon asking a question, documents are retrieved using all three models, the result list consolidated, and documents ranked by weighted score. The solution won the first and second rounds of the TREC-COVID information retrieval challenge. The fusion strategy of Nguyen et al. (2020) combines BM25 scores with normalized cosine similarity from the neural representation of sentences embedded with various models (BioBERT-NLI, Covid-NLI, ClinicalCovid-NLI, BioBERT 1.1 STS, BioBERT MS MARCO). They claim that "*neural scoring is beneficial in alleviating some of the shortcomings of the keyword-based retrieval. A balanced scoring function combines the strengths of the inverted and neural indices*".

In recent work applying re-ranking is the CAiRE-COVID application (Su et al., 2020) that first uses the generalization capabilities of their HLTC-MRQA model, then re-ranks the extracted relevant snippets with BioBERT fine-tuned on SQuAD. In this Q&A system design, text indexing is at the paragraph level, returning top *n* paragraphs matching the query. By winning tasks in the Kaggle COVID-19 Open Research Dataset Challenge, their system demonstrates the value of a two-pass approach, first matching passages by key terms then filtering further by semantic similarity.

If we consider questions are broadly of the shorter keyword, longer natural language, or extended narrative format, the introduced lines of research are of particular relevance to this thesis. They allow us to hypothesize that such a hybrid search mechanism provides two benefits: key term matching specific biomedical nomenclature, and leveraging semantic similarity to achieve good results with questions of a natural language format. The solutions provide insight into the strength of ensemble models favoring lexical or semantic search according to the question being asked. We propose that the re-ranker type is a robust strategy. An initial lexical search with BM25 or TF-IDF is likely to have a low false-positive rate, meaning only documents containing key terms including named entities such as proteins and genes are included. This results in a first pass high-quality candidate list re-ranked with the semantic model to order by most similar meaning. In effect, this strategy empowers lexical search with semantic capabilities while also reducing the compute cost and associated latency penalty of a full semantic search.

In conclusion, what remains unanswered from the literature is the importance attached to balancing the structure of typical questions being asked with selecting an appropriate transformer embedding mechanism optimised for the question format. We hypothesize paragraph-level embedding performs poorly for comparison with short key term questions where little similarity exists. As CAiRE-COVID applied BioBERT, trained at sentence level, for asymmetric search where the question looks to find a longer paragraph, we suggest fine-tuning on SQuAD effectively redresses the imbalance.

The review of existing studies does not yield insight into the compute cost of comparing an embedded question with a large collection of passage vector representations. With retrieval speed key to the end-user experience, how much does response latency increase with increasing complexity of semantic model strategies, and how well do they scale with increasing vector representations? Given the varying strategies, an evaluation of lexical, neural, and re-ranking approaches to search is justified. Particular attention will be paid to the effectiveness of lexical-only, what observable improvements a neural re-rank offers, and the cost of a full neural search.

2.4.3 How are biomedical search tasks evaluated?

The theme of challenges in intrinsically evaluating models for NLP tasks in the biomedical domain emerged, as existing benchmark suites, trained from generic corpora, have an inherent selection bias towards terms in the source text collection. Models working well in answering questions from Wikipedia Q&A evaluation datasets are unlikely to do well on the focused thesis task of COVID-19 Q&A.

To resolve the lack of evaluation datasets in the domain, (Lee et al., 2020) released their own called COVID-19 Questions. Similarly, Tang et al. (2020) take their learnings from exploring the CORD-19 dataset to assemble CovidQA, a small dataset of 489 question-article pairs designed to evaluate models. Despite these two evaluation datasets being valuable research contributions, there will always be selection bias otherwise a vast suite incorporating all question possibilities would be necessary.

While intrinsic evaluation of systems with benchmark suites is valuable in isolation, they are insufficient. Indeed, SOTA performance in synthetic testing is irrelevant and of academic interest only if results are considered inferior by human domain experts. Extrinsic evaluation is expensive yet critical for any Q&A system to be widely adopted by demanding clinical researchers. In evaluating language models applied to COVID-19 Q&A, Oniani and Wang (2020) ask two medical researchers to rate responses. In collaborating with human

domain experts in the design of Q&A evaluation suites and preparing deployment of Q&A systems for end-user consumption, we advise sharing with the user community an understanding of how types of models incorporated into the selected search strategy work. Such knowledge transfer provides an awareness that keywords help lexical models, and an extended, more elaborate question better captures meaning for semantic-based models.

2.5 Information extraction from literature with alternatives to BERT

In reviewing other research lines proposing alternatives to deep-learning for knowledge extraction from biomedical literature, a system extracting the relationship between temperature, humidity, and the spread of COVID-19 by Sastre et al. (2020) shows particular promise. Articles are first indexed then searched for keyword terms in Apache Lucene. With Unitex and GrapeNLP, lexical, syntactic, and semantic grammar restrictions are hand-crafted for second-pass filtering of paper selection by fuzzy, non-exact match. The system was selected as the best proposal for creating key insight summary tables by the Kaggle CORD-19 competition jury panel. While the authors acknowledge the current popularity of deep learning approaches such as BERT transformer models for complex information retrieval tasks, they claim an advantage in two key areas: *1)* reduced training times and *2)* clearer explainability with the defined grammar self-documenting the retrieval mechanism. We recognise the value these characteristics support continuous model re-training in highly dynamic data drift areas like biomedicine research while providing insight, transparency, and trust in models used in the decision-making process.

BioSentVec (Chen et al., 2019), a vector embedding model using a continuous bag-of-words (CBOW) approach as an alternative to a neural network, is pre-trained using PubMed articles and anonymized intensive care unit clinical notes from the MIMIC-III database (Johnson et al., 2016). The authors claim their system is "*robust across multiple text genres in biomedicine text*", exemplified by deployment within LitSense (Allot et al., 2019), a sentence-level search application querying half a billion statements from PubMed biomedical literature. They confirm the current trend in hybrid search strategies that fuse vector embeddings with lexical indexing, claiming that "*combining traditional term matching based methods and BioSentVec can significantly improve the search effectiveness of sentence retrieval*".

While we applaud the solution deployed at such scale, we argue there remains room for improvement by replacing CBOW with BERT. Although both approaches create word embeddings based on word context during training, the Word2Vec CBOW model takes the average of all vector representations for each word into a final vector, while BERT retains all word representation vectors, meaning its final embeddings are context-dependant. This allows BERT to capture the nuances of different semantic meanings of the same word.

Zhang et al. (2019) introduce an approach to fine-tuning the embedding of biomedical text by combining the vector representation of words from PubMed article titles and abstracts with vector representation of related metadata from the MESH medical subject headings ontology. They show their embedding technique captures cosine similarity between "*mycosis*" and "*histoplasmosis*", and similar words to deltaproteobacterial (deltaproteobacterium, betaproteobacteria, zetaproteobacteria, delta-proteobacteria) better than other methods. This work is a clear demonstration of the value in annotating text with additional in-domain knowledge to improve the performance of downstream NLP tasks.

2.6 Conclusion

As reported by many reviewed works, the key learning relates to SOTA performance of BERT-based models in capturing complex domain semantics for downstream tasks such as IR. The reviewed papers indicate additional performance from training and fine-tuning to calibrate models to the target domain. On this basis, we will proceed with experiments evaluating a variety of IR strategies using base and tuned BERT models, comparing their ability to surface the most relevant evidence to directly answer information needs and measure them against baseline traditional search models such as BM25.

3. Datasets

3.1 Introduction

The proposed COVID-LEAP research system relies on a collection of open-source datasets to provide foundational knowledge for answering questions related to the highly dynamic COVID-19 pandemic research domain and vaccine landscape. This chapter presents each dataset, providing an overview, its source, purpose, and format, discusses initial interpretations, and justifies its value to the solution.

3.2 COVID-19 Open Research Dataset for biomedical literature

Compiled by a technology partnership that includes Microsoft Research and partners, CORD-19 is a corpus of coronavirus research articles updated weekly. When first released on 20th March 2020, the collection consisted of 27K articles from PMC and 2K pre-prints from archive sites medRxiv2 and bioRxiv. Note that only PMC is peer-reviewed. As of 8th March 2021, the dataset we analyse in the remainder of this chapter has seen tremendous growth, with 170K peer-reviewed texts, and significantly, 129K pre-prints.

In compressed format, the dataset is 7.8GB, and 32GB when extracted. An open-access license allows reuse for raw data analysis, the content helpfully provided in machine-readable JSON format. A metadata file summarizes all articles with their associated title, authors, abstract, affiliations, publish date, unique digital object identifier (DOI), and link to the full body text. Sub-folders *PMC_JSON* and *PDF_JSON* store the full articles with sections including body text, abstract, results, conclusion, and bibliography. The full body text preserves paragraph breaks to allow tokenization into smaller units.

CORD-19 provides COVID-LEAP with open-source domain knowledge to answer pandemic-related biomedical questions. Initial exploration after cleansing (Chapter 5) provides some insight into the remaining 127,434 articles. English is the dominant written language (124,671), followed by German (1255), Spanish (770), French (507), Dutch (127), and Italian (42) (Fig. 2). With the initial COVID-LEAP NLP pipeline designed only for English texts, other language articles are excluded. The average paragraph count per article is 47 (Fig. 3), insightful for extrapolating the volume of expensive semantic vector embeddings in fusion search strategies.

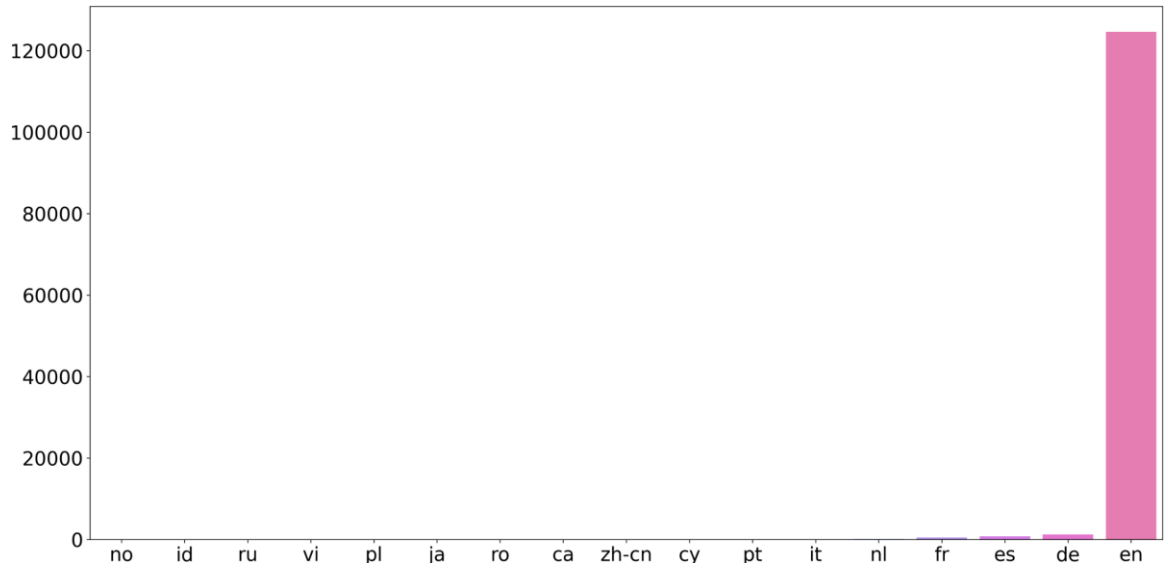

Fig. 2: CORD-19 articles by language

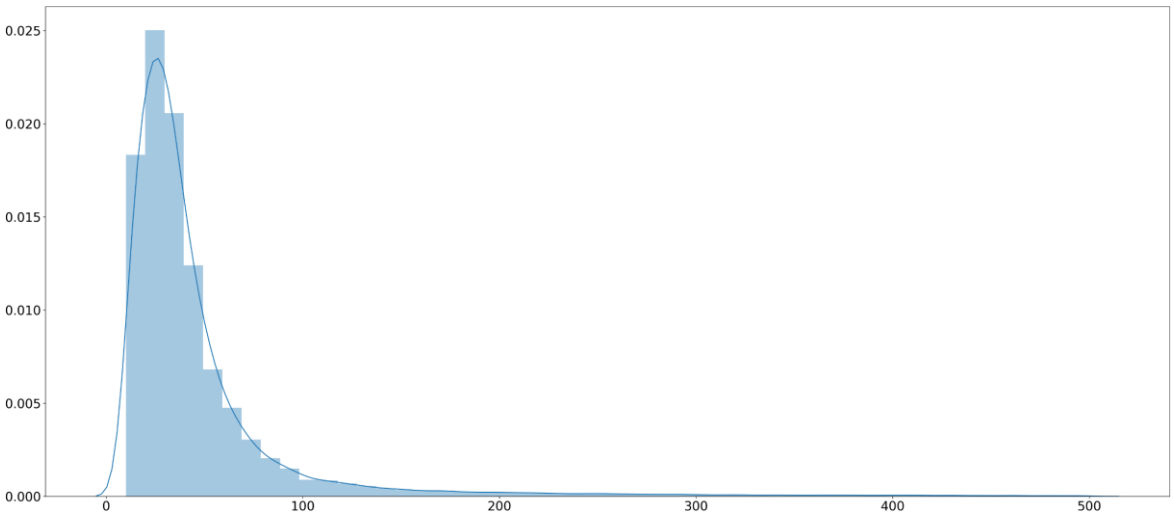

Fig. 3: Density distribution of CORD-19 article paragraph count

Inspired by the work of Colavizza et al., we characterize each article's full text according to a case-insensitive match on the terms listed below. However, we perform a second, far stricter search using only the article title, for a total of eight flags per article, indicating two levels of relevance to specific virus variants and the coronavirus family as a whole.

- **Coronavirus**
  `covid OR "covid 19" OR covid-19 OR 2019-nCoV OR coronavirus OR "sars cov" OR sars-cov OR sars-cov-2 OR mers OR mers-cov OR "middle east respiratory syndrome" OR "severe acute respiratory syndrome" OR hcov`

- **Sars-CoV-2**
  `"covid 19" OR covid-19 OR 2019-nCoV OR sars-cov-2`

- **Sars-CoV**
  `sars-cov`

- **Mers-CoV**
  `mers OR mers-cov OR "middle east respiratory syndrome"`

A small but growing number of coronavirus-related papers were published following the 2002 Sars-CoV and 2012 Mers outbreaks, with a tremendous increase in research with Sars-CoV-2 in 2020 (Fig. 4). We hypothesize the reduced disease article subset obtained when characterizing the corpora with the stricter title search includes research directly focused on coronaviruses or specific variants (Fig. 5), compared with categorisation using the full-body text that might include works of less or historical relevance. With title alone, 238 articles are categorised as Mers-relevant, compared with 13,127 using the full body text. Similarly, 38,604 and 56,878 articles for Sars-CoV-2 with title and full body text, respectively. Using these flags to subset the corpora during search has the potential to improve relevancy and will be evaluated.

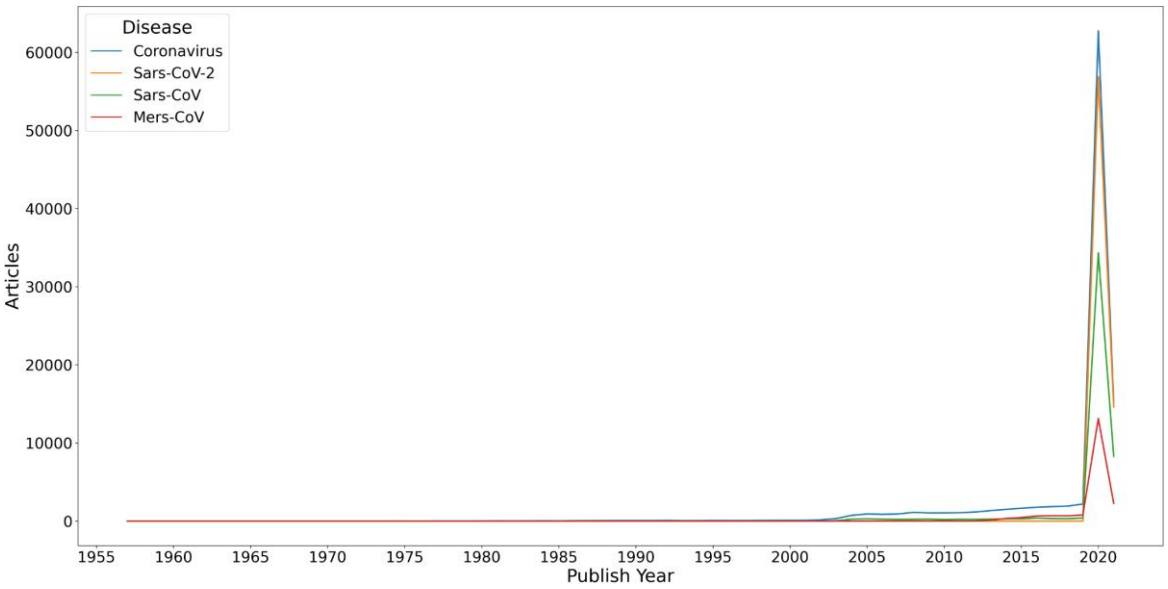

Fig. 4: CORD-19 article body text by disease category

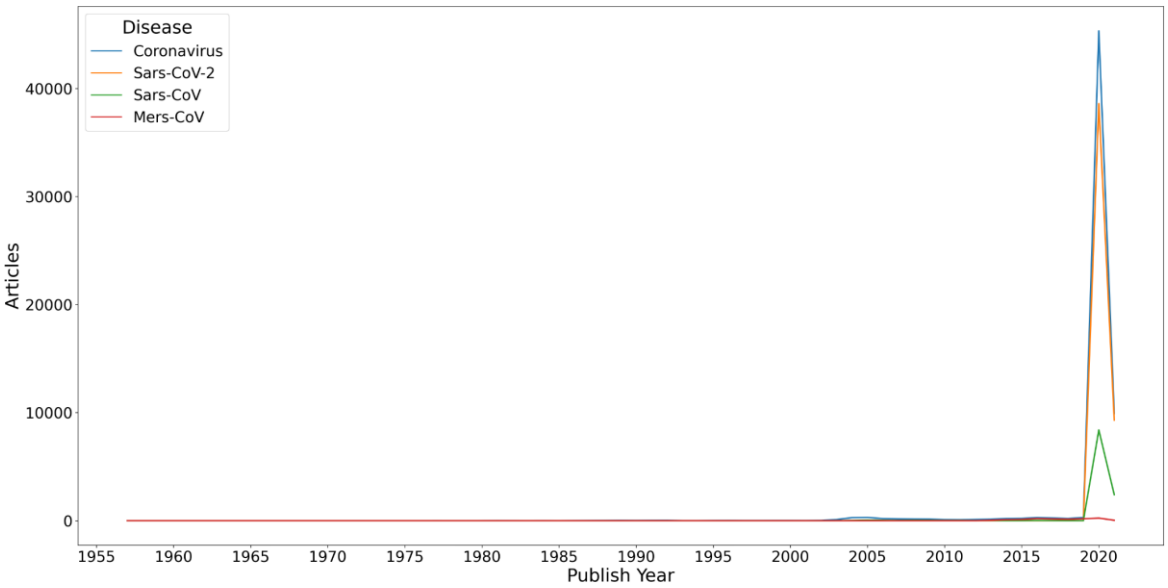

Fig. 5: CORD-19 article title by disease category

In fighting the global crisis that is COVID-19, the speed of research and expediency in sharing it is critical. Having identified the explosion in pandemic literature, we also show the increased importance of alternative channels of knowledge dissemination to the scientific community. PLoS One (Public Library of Science), an online journal specialising in the acceleration of peer-reviewed papers, dominates by the number of articles contributed, followed by bioRxiv, a pre-print server of biology papers (Fig. 6). Considered highly reputable journals, Nature and The Lancet publish significantly fewer papers (Fig. 7). It is reasonable to question whether the more traditional review methods, often taking months, are too restrictive and out of step with pandemic dynamics.

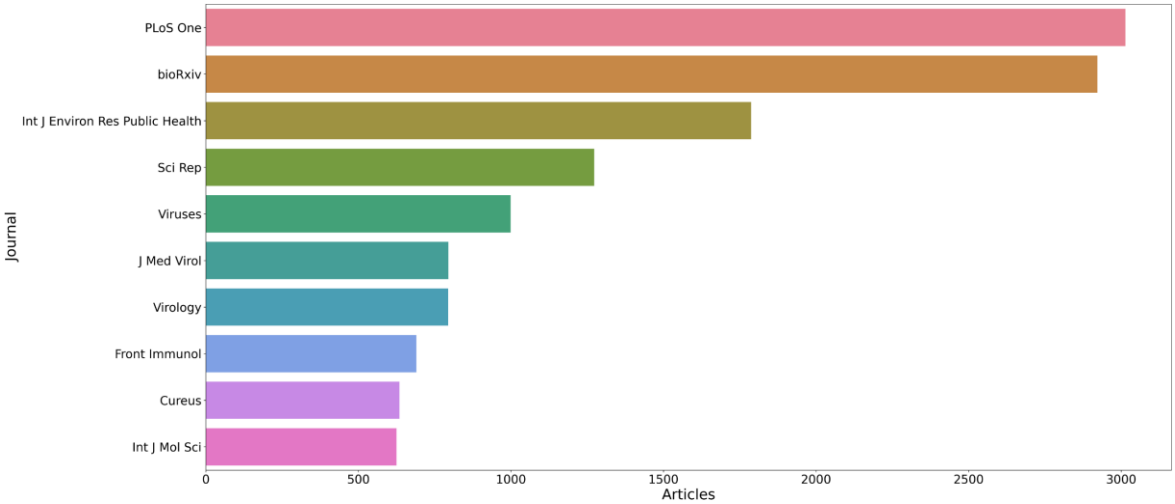

Fig. 6: CORD-19 top 10 journals by the number of articles

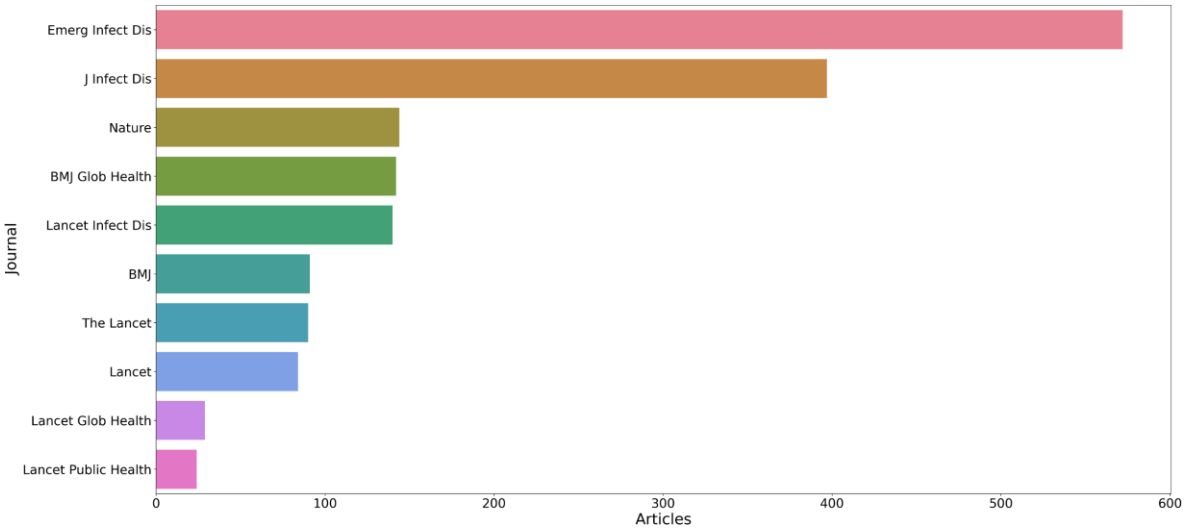

Fig. 7: CORD-19 top 10 reputable journals by the number of articles

3.3 WHO candidate vaccine landscape database

The World Health Organisation (WHO) COVID-19 candidate vaccine landscape database compiles detailed information on vaccine candidates in development (Draft landscape of COVID-19 candidate vaccines, 2021). First published on 11th April, it tracks the progress of pre-clinical and clinical vaccine trials, with updates published bi-weekly. Information includes developer (e.g., Moderna) and platform (e.g., non-replicating viral vector). For vaccines in clinical trials, the current phase (e.g. phase 1 evaluating safety and identifying dose), identifiers for associated trial registries (NCTn for U.S. ClinicalTrials.Gov, ChiCTR for Chinese chictr.org.cn), and number/timing of doses are given.

Unfortunately, the intended audience is the human reader, with the "*database*" provided as one of two Portable Document Format (PDF) formats, or more recently as a Microsoft Excel file that changes in both layout and formatting over publications. There is no inherent formal structure in the shared information, with manual extraction, evaluation, and summarization an expensive and error-prone endeavour, often performed by clinical analysts whose time is best utilised elsewhere.

While a lack of structure poses a data mining challenge, this project aims to automate the capture of this essential clinical trial information with high accuracy to eliminate manual effort and reduce the turnaround time between update and availability to key decision-makers. COVID-LEAP uses this dataset for high-value insight to answer questions including: *Which private companies and government institutions are funding research? Where are their vaccines being registered? What are the characteristics of their clinical*

study? Which of the vaccines have reached human trials? How fast are their vaccine programmes moving, and how close are they to market readiness? Which research papers reference COVID-19 vaccine clinical trials? What platforms are proving most effective? Lastly, what platforms dominate the research landscape, as seen in Fig. 8, an example from COVID-LEAP which extrapolates the vaccine platform temporal evolution from the clinical subset of this dataset.

As the WHO only provides the latest publication via a dedicated web page and does not maintain an archive of historical publications, older publications for time-series analysis are captured with the internet archive. In total, the landscape database collection comprises 44 publications released between 11th April 2020 and 22nd March 2021.

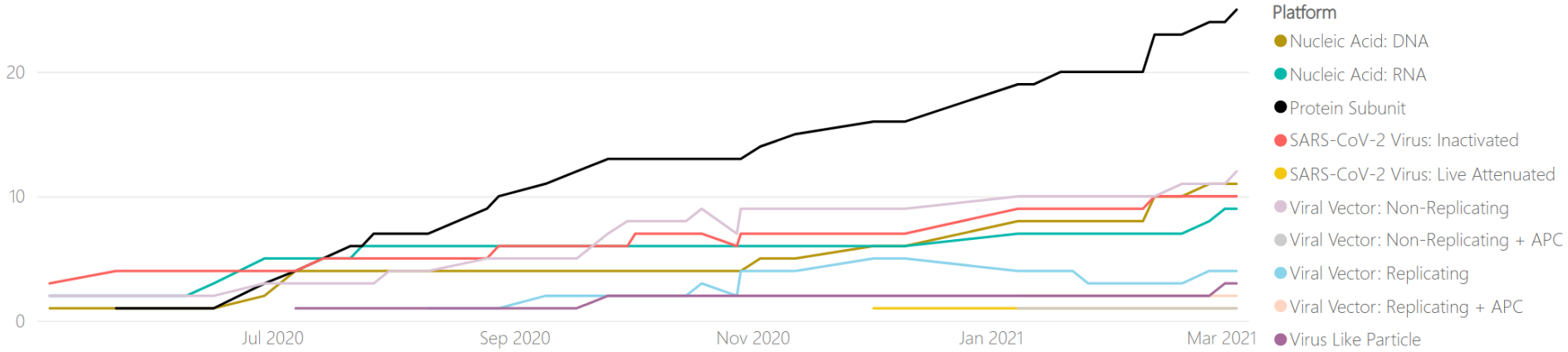

Fig. 8: WHO landscape clinical candidates by platform/type over time

3.4 ClinicalTrials Registry

Provided by the United States National Library of Medicine at the National Institutes of Health, ClinicalTrails.Gov is a database of clinical trials registered in the U.S. From close collaboration with clinical researchers, we understand it is the go-to information resource for clinical studies on a wide range of diseases and conditions. Providing a search engine, using the term "*covid OR covid-19 OR covid19 OR covid 19 OR coronavirus OR sars-cov-2 OR 2019-nCoV OR wuhan OR china virus*" in the "*Condition or disease*" field lists pandemic relevant clinical studies with associated information, including unique registration identifier (NCTn), study title, status, and interventions. Selecting a listed trial navigates to a view detailing study design, participants, eligibility criteria, and more.

It is important to note that only a subset of trial information is included in a comma-separated machine-readable download of search results. As of 22nd March 2021, the earlier mentioned search term finds 5289 studies. COVID-LEAP uses this trusted source dataset to provide a detailed view of clinical trials referenced the WHO vaccine landscape.

3.5 Aggregated Analysis of ClinicalTrials.Gov database

While valuable, extracts from ClinicalTrials.Gov are incomplete, missing essential study information such as interventions and design. COVID-LEAP resolves this issue by supplementing CTG extracts with the AACT (Aggregated Analysis of ClinicalTrails.Gov, 2016) database developed by the Clinical Trials Transformation Initiative (CTTI). The purpose of the AACT is to enhance the usability of the registered trials data by restructuring study attributes into discrete fields and enriching by "*Annotating studies by clinical speciality, using a custom taxonomy employing MeSH terms applied by an NLM algorithm, as well as MeSH terms and other disease condition terms provided by study sponsors*" (Tasneem et al., 2012).

Updated daily, it is available as a collection of pipe-delimited files extracted from a normalised data model of 46 relational tables. Trials unrelated to COVID-19 research are included, so content is further filtered with `"covid 19" OR covid-19 OR 2019-nCoV OR sars-cov-2".`

3.6 Conclusion

COVID-LEAP brings disparate silos of biomedical information into one collection for a rich view of the domain landscape for clinical researchers and decision-makers. Although not perfect, requiring several cleansing steps in preparation for optimal semantic modelling and indexing of content, we acknowledge the publishers of CORD-19, led by the Allen Institute of A.I., at least understand the needs of the scientific data research community by providing each release of the dataset in a consistent, machine-readable JSON format.

Unfortunately, the WHO vaccine landscape database does not adopt the same desirable characteristics, proposing they did not anticipate a section of the wider audience consuming their vaccine candidate information with machine-based analytics. Particularly for releases in PDF e-document format, it is not straightforward to process. Several challenges include processing each page individually, capturing information with rules searching at string level, colour formatting untranslatable by machine, inconsistency in what a column represents, and mixed units of information. Furthermore, publication has not always been timely, on occasion delayed by days.

We recommend the WHO implement some data best practices and standards to produce robust, trusted machine-readable datasets, ideally in a more expressive format like JSON, allowing the focus to be on insight and knowledge discovery rather than preprocessing maintenance. From a data-sharing perspective, there is further opportunity for the WHO to learn from the pandemic's velocity by advancing their practices to align with the needs of the biomedical and data sciences for a more efficient sharing of knowledge. We would be better placed and more effective in supporting human domain experts to analyze the trials and vaccine landscape if valuable data sources are in a consistent format easily processed by the machine.

4. Solution Architecture

4.1 Introduction

The vision is to develop an application that makes the process of identifying and extracting relevant COVID-19 knowledge more efficient for the researcher. This chapter documents the three stages of the solution architecture design process, organised as follows: first, a review of the problem context is discussed, followed by a translation of the problem into a high-level process flow, and then finally, a detailed architectural blueprint that presents specific technology components and their interaction.

4.2 Problem Discovery Canvas

Before assigning technology to solve problems within our project domain, we consolidate key discoveries and learnings from the literature review (Chapter 2) and dataset evaluation (Chapter 3) into a problem discovery canvas (Fig. 9) that lays out the landscape in understanding the existing situation. Several domain challenges are identified:

1. The corpus of knowledge is large, noisy, unstructured, disparate, complex, varying quality, and highly dynamic.

2. There is a broad range of users, each with their own information needs and approach to forming questions.

3. There are many technical challenges in implementing a solution, from the robust and timely ingestion of source data, to modelling it in acceptable timeframes for search at reasonable compute and monetary expense.

If success criteria include reduced wasted time by delivering highly relevant answers to biomedical questions at low latency and provisioning the latest insight into the fast-evolving vaccine landscape, we consider these problems worth solving.

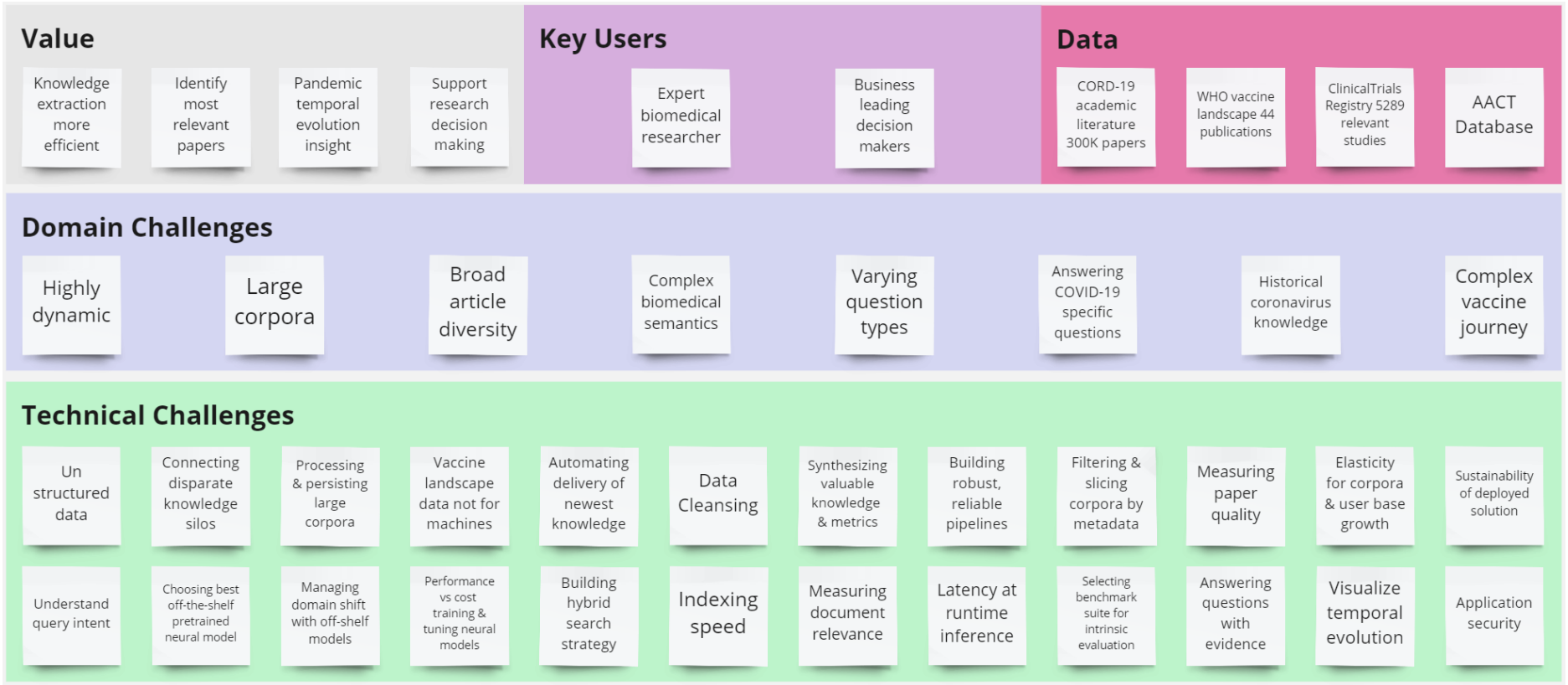

Fig.9: Problem discovery canvas

4.3 High-Level Solution Process Flow

As we now understand the problem domain, what data sources of knowledge are available, who our data consumers are, what problems need to be solved, and are aware of the technical challenges before us, we can begin building out a vision of the solution. Before any technology choices are made, we deconstruct the complex problem into a simple directed flow of principal process steps (Fig. 10). It begins with the extraction of the four publicly available datasets, their contents filtered and cleansed. Additional features to support subsetting, such as by publishing year or document topic, or linking literature papers to clinical trials, are engineered before the knowledge is persisted in a relational database.

Text from the literature collection is consumed by a fundamental sub-group of tasks related to the assembly of the semantic neural model, with pre-training and fine-tuning tasks responsible for incorporating COVID-19 specialised domain knowledge. This model encodes a numeric representation of each literature paper paragraph which is stored in a search engine. The relational database can provision COVID-19 landscape knowledge to both business intelligence services and web-based apps. Via an application programming interface (API), a question passed to the inference engine is encoded with the previously built model for similarity comparison with the indexed vector representation of the complete corpus.

4.4 COVID-LEAP Solution Architecture

Following a review of our problem discovery and mapping out the solution process flow, we can now translate the problems into a cohesive solution, presenting the detailed solution architecture, including developer tooling alongside selected hardware, software, and cloud infrastructure (Fig. 11). Note blue titles denote cloud-hosted layers, with arrows indicating component integration and the direction of information flow.

The problem demands the elasticity and agility of cloud services. As an innovating R&D project, experiments with solution components follow a development pipeline process with provisioning, prototyping, test, refine, tear-down, rescale and rebuild steps that are easier, quicker, and cleaner to execute with cloud services than on bare-metal servers. The solution is realized in Azure, a cloud computing platform from Microsoft, selected for several reasons. Microsoft has a long history as a trusted enterprise partner. Authentication to services is often simplified due to the integration of identity management with Azure Active Directory (AAD). The author finds navigating the plethora of services more user-friendly than cloud portals from other vendors.

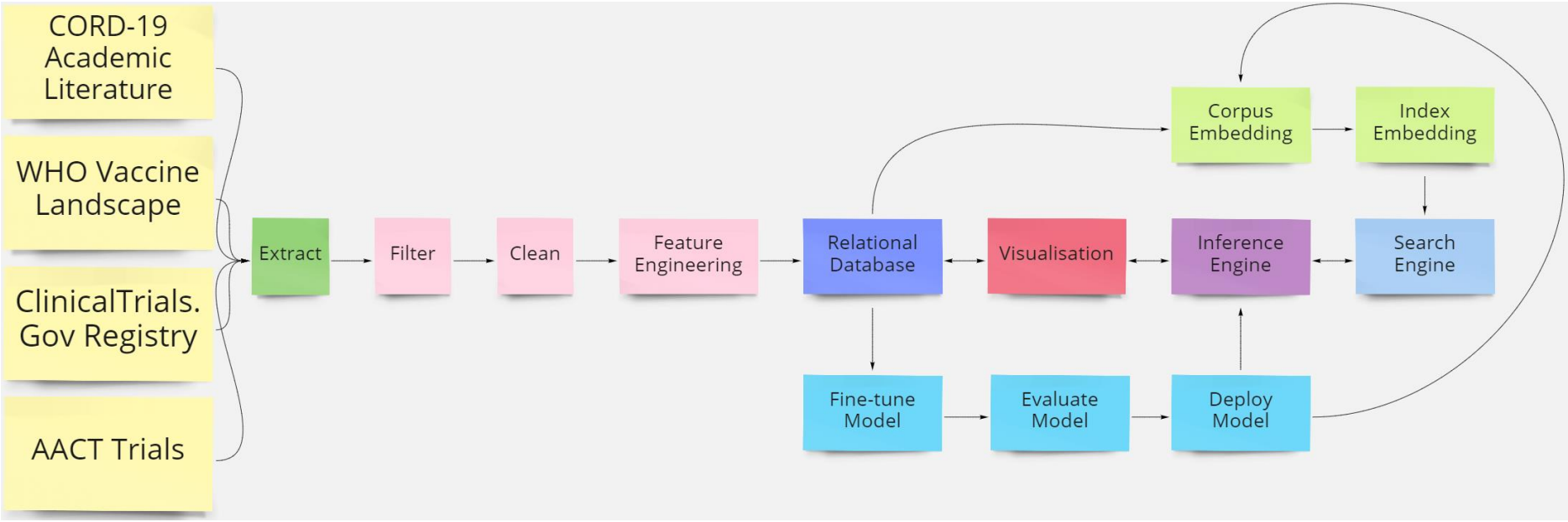

Fig. 10: COVID-LEAP high-level solution process flow. **Yellow** shows incoming knowledge sources. **Dark green** indicates extraction of sources and uploads into Azure. **Pink** covers data pre-processing. **Blue** represents data storage of prepared data. **Light blue** indicates dense model lifecycle from tuning to deployment. **Light green** reads article data from the relational database and, using the dense model, embeds paragraphs for vector indexing in ElasticSearch. **Purple** represents the inference script that takes a query and executes the retriever model to return a list of results to the user in **red**.

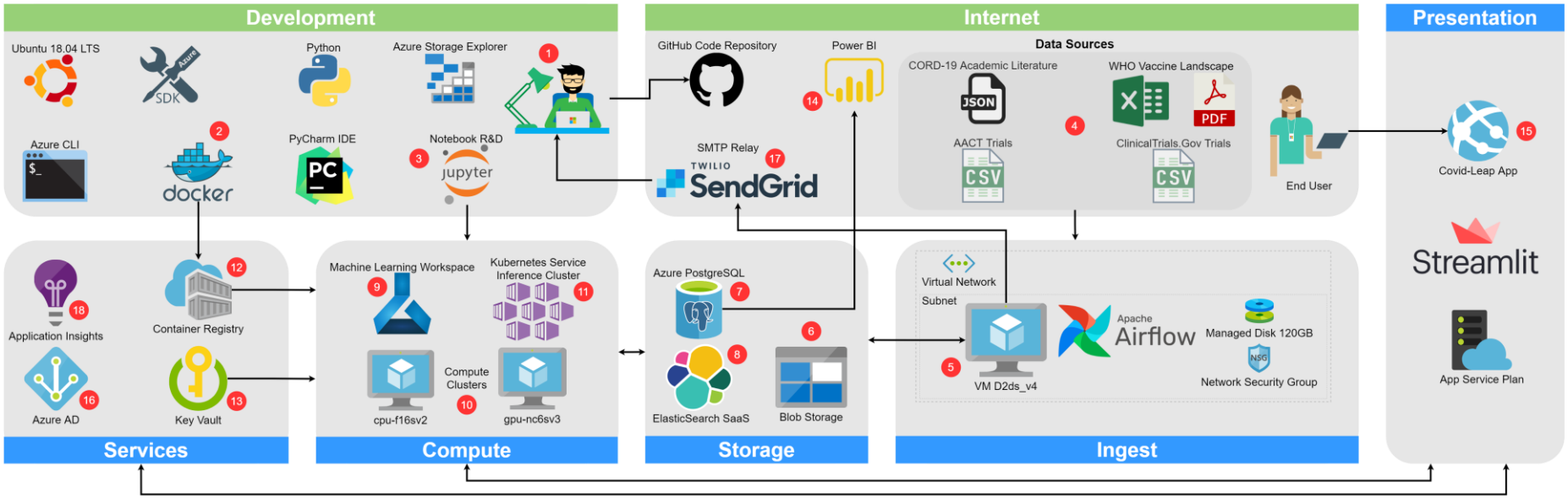

Fig. 11: COVID-LEAP solution architecture. Each main box is an architectural layer. Blue titles denote Microsoft Azure-hosted layers. Within the layers are explicitly named technical components

A set of guiding principles governs specific service and component inclusion. These are modularity, consistency, sustainability, scalability, performance, automation, and security. The remainder of this chapter covers each of the architecture blueprint layers, such as those responsible for data ingestion and storage, and where appropriate, underlines the association selected components have with these principles.

4.5 Development

This sub-section describes the primary development environment of COVID-LEAP. The local development system specification includes an AMD 5950X CPU, Nvidia 3090FE GPU, with 64GB RAM running Ubuntu 18.04 LTS (Fig. 11-1). The codebase of the COVID-LEAP clinical research system is developed in Python, a procedural, object-orientated programming language with an extensive array of packages for NLP and cloud service interaction that support the expedition of a complex application build.

Each technical challenge highlighted on the problem discovery canvas (Fig. 9) is first prototyped as a standalone Jupyter notebook experiment (Fig. 11-3) for data analysis, experimentation, and to prove the technical feasibility of multiple methods to delivery of proposed main features, for example, IR with BERT and topic modelling, validating performance and end-user value. Taking this approach assembled a collection of building blocks, helping gauge the possibility of a minimum viable product (MVP). In PyCharm, a popular python IDE (Integrated Development Environment), approved notebook experiment code is refactored into a structured solution framework following an object-orientated paradigm for modularity and reusability.

A docker container (Fig. 11-2) encapsulates OS runtime version and Python package dependencies. This docker is used by PyCharm as its Python interpreter and shipped to an Azure container registry (Fig. 11-12) to ensure the same consistent application environment when developing locally and running in the cloud. Adopting containerisation brings several benefits. As Infrastructure as Code (IaC), docker files explicitly describe system configuration and hence are self-documenting. As scripts, version control with rollback is possible. Looking to the future, it facilitates multiple developer support with the same environment and solution mobility for straightforward deployment to cloud providers other than Azure. The Azure CLI supports command line management of Azure resources, while Azure Storage Explorer is a graphical tool used to upload, explore and delete the Azure-hosted dataset collection during development.

4.6 Internet

This layer provides access to our open-source datasets and cloud services. Throughout the development lifecycle, software assets including experimentation notebooks, Python scripts for data cleansing and model building, docker files describing runtime environments, and images for the web application front-end are pushed to a GitHub code repository for archiving and version control.

The internet layer also hosts the publicly available datasets we consume to provision our solution with domain knowledge (Fig. 11-4), as well as provide connectivity to data stores for the Power BI business intelligence suite (Fig. 11-14) and end-user access to the web application (Fig. 11-15). SendGrid (Fig. 11-17), a web-hosted email service available free of charge for a maximum of 100 emails per day and used by large customers, including Uber and Airbnb, is used as an SMTP relay for data ingestion status notifications.

4.7 Ingest

Our datasets of domain knowledge are at the heart of our solution. With the COVID-19 knowledge landscape changing rapidly, it is critical for the latest published versions of these datasets to be brought into our solution environment for trust that answers and insight are with the newest information available. The ingest layer of the COVID-LEAP architecture for enterprise-grade automated loading our multiple sources of knowledge into storage uses Airflow (v.2.0.1), originally developed internally at Airbnb before being open-sourced to the Apache foundation.

Automating the consistent deployment of our Airflow platform is a docker container, configured beyond vanilla to reference environment variables persisting "*secret*" credentials for authenticated connectivity to cloud storage and a custom configuration specifying the SendGrid email backend. Following the modularity principle by separating the ETL process from the data processing and compute pipelines, the docker container runs on an independent virtual machine (Fig. 11-5, SKU-type D2ds_v4, two vCPU, 8GB ram) with Ubuntu 18.04 LTS and resized OS disk. Security is incorporated into this layer with a network security group limiting SSH (secure shell, port 22) and HTTP (for web server front-end, port 8080) connectivity to white-listed IP addresses.

The orchestration of dataset extract and load is managed with the programmatic describing, in Python code, of workflows called Directed Acyclic Graphs (DAG), which are just a collection of related tasks. Each graph node is known as an operator, able to execute Python functions or OS-level commands. In addition to declaring tasks, their flow, and

dependencies, DAGs also specify a runtime schedule, for example, to be executed once daily at 01:00 UTC. Fig. 12 presents the DAG for the CORD-19 dataset, showing a positive check for newer published version branching to the extract and load process, followed by the triggering of Azure ML pipelines, then finally a success confirmation email relayed by SendGrid.

Datasets are uploaded in their raw, unprocessed form to Azure blob storage. Airflow has a web user interface for monitoring and troubleshooting workflow pipelines. Note Airflow is responsible for extracting the internet-hosted datasets and loading them into cloud storage, with data transformation executed by subsequent cloud services; hence COVID-LEAP adopts an ELT (extract, load, and transform) pipeline design rather than ETL.

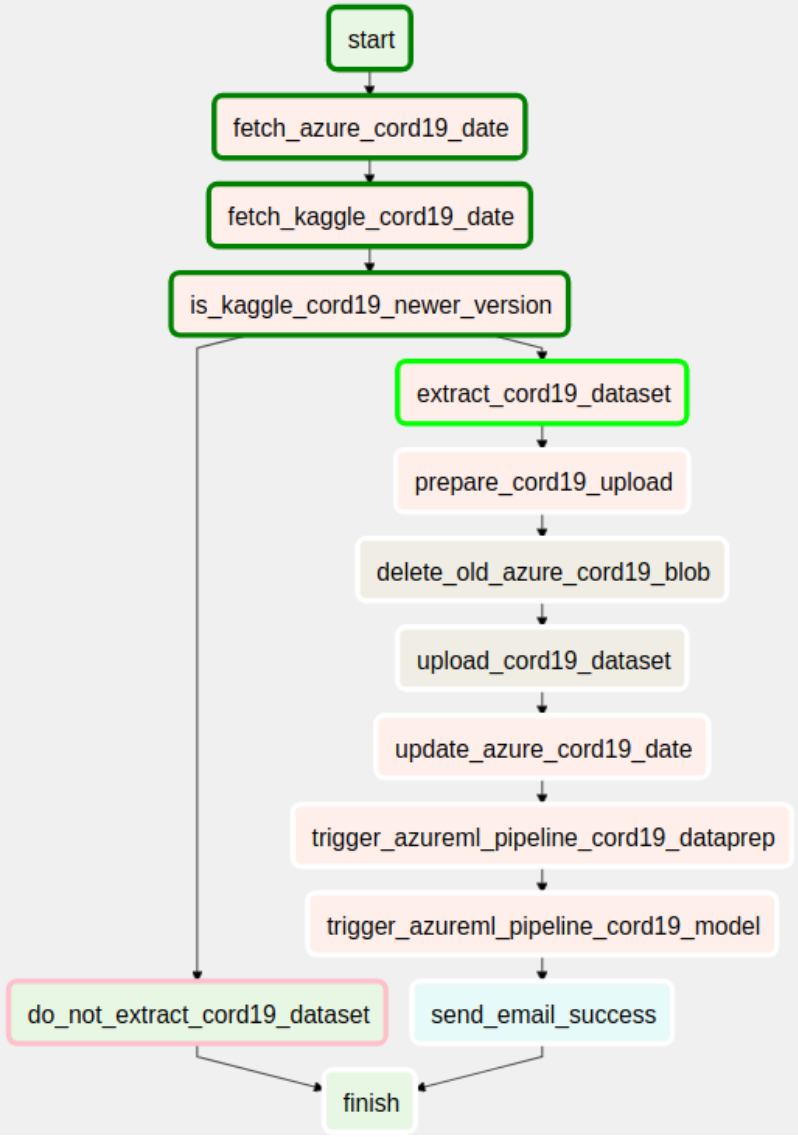

Fig. 12: Directed Acyclic graph for automated CORD-19 dataset extract from web and load into Azure blob storage

4.8 Data Storage

This layer is responsible for storing our raw datasets, persisting relational models, and hosting the search engine in which academic texts are indexed ready for information retrieval. The ingest layer uploads raw domain datasets into blob storage (Fig. 11-6). A set of Azure ML pipelines filter, cleanse, and further engineer the data before persisting it to a digital repository of academic literature and clinical trials data with an Azure Database for PostgreSQL server (Fig. 11-7).

Despite our research knowledge originating from a corpus of unstructured academic literature, the content has an inherent structure, for example, papers having authors (*paper > authors*), authors having authored papers (*author > papers*), authors having co-authors (*author > authors*), and journals publishing papers (*journal > papers*). This validates the selection of a SQL relational database over a NoSQL alternative such as MongoDB. With corpus entities normalised into individual tables, multiple sources of knowledge can be brought into one cohesive collection by connection at the database level, for example connecting literature papers with clinical trial data via trial reference.

The database provides data to several consumers, including the PowerBI business intelligence platform for clinical landscape dashboarding and the AzureML pipeline encoding paper paragraphs for neural semantic search. For information, the PostgreSQL server is configured as the minimum SKU within the general-purpose tier (two vCores, 100Gb storage) to allow room for up-scaling, as changing to and from the basic compute tier is not supported.

A proven solution for enterprise-ready, scaleable search, we selected a managed ElasticSearch SaaS offering as the basis of our search engine. ElasticSearch provides the ability to index the complete literature corpus at paragraph-level, including a distributed vector representation of each paragraph encoded by a neural model. Additionally, metadata such as publishing year, authors, and journal are included during indexing. With this design selection, we address the technical challenge of architecting a hybrid search strategy fusing lexical and semantic search results to handle a range of question formats while also allowing users to filter by metadata facet. The web front-end app (Fig. 11-15) passes a question to the compute layer inference cluster (Fig. 11-11), which encodes it with the deployed neural model and uses ElasticSearch to retrieve vector representations of paper paragraphs for similarity comparison.

4.9 Compute

The compute layer provides the CPU and GPU computing resources in addition to the managed Azure Machine Learning (AzureML) service framework (Fig. 11-9). Within AzureML, a subset of features is used to explore datasets, author, publish, orchestrate and monitor enterprise-grade scalable ML pipelines in addition to registration and deployment of models. With COVID-LEAP, we take a code-first approach with custom scripts implementing pipelines for data transformation and specialised, in-domain deep learning (DL) model lifecycle tasks.

Provisioning pipelines with data, the four knowledge sources (CORD-19, WHO, ClinicalTrials.Gov, AACT), uploaded by Airflow into blob storage, are registered as AzureML datasets, making them seamlessly available without the need for access keys, connection strings, or passwords. An example transformation pipeline responsible for processing vaccine landscape and clinical trials data is presented in Fig. 13. The process is modularised with an individual Python script for each step function. Typically, inputs and outputs of steps define dependencies and flow order, while step parameters allow pipeline execution locally or in the cloud without code change. While each pipeline step is a Python script, we currently author pipelines with Jupyter notebooks, allowing decoupling of the internal step functionality from workflow orchestration. An example benefit of this separation is testing to find a good balance between VM compute performance versus cost without amending pipeline step code.

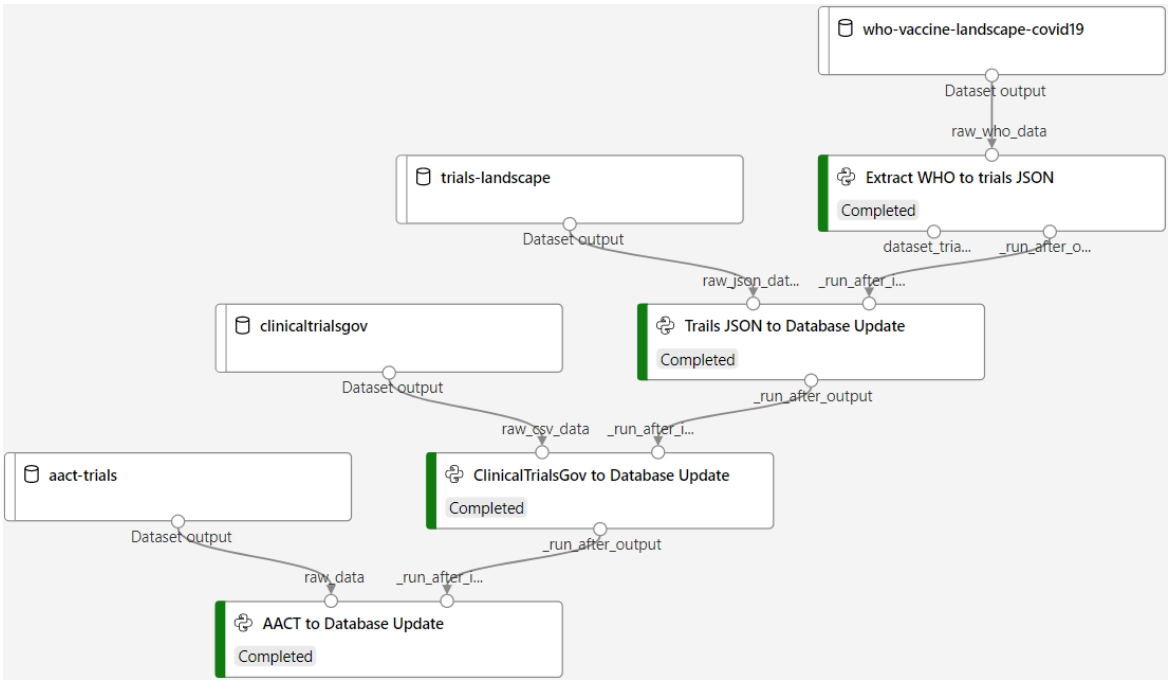

Fig. 13: AzureML pipeline processing clinical trials datasets. Steps with barrel icons, e.g., *who-vaccine-landscape-covid19*, represent a dataset in Azure that is read and processed by the following Python script step

Pipelines are published as web services and triggered from Airflow by calling service endpoints. The pipeline execution environment with all the necessary Python package dependencies is the same docker container (Fig. 11-12) used locally, ensuring consistency between off-line development and cloud runtime. We solve the challenge of preprocessing and encoding a large corpus by taking advantage of a key feature of cloud services: creating auto-scaling, multi-node virtual machines with size matched to workload requirement.

For initial cleansing, language detection, mining literature for clinical trials references, identifying authors and papers of influence, topic modelling, and engineering derived features such as published year and association of paper with specific coronavirus variant, we opted for the compute-optimized VM SKU-type F16sv2 with 16 vCPU (virtual CPUs), 32GB memory and 128GB SSD storage (Fig. 3-10). Tensor operations in the BERT-based neural model embedding the complete literature collection are greatly accelerated with GPU resources, hence using the more expensive GPU-enabled VM nc6sv3 with 6 vCPU, 112GB memory, 336GB SSD, and 1 Nvidia Tesla V100 card with 16GB VRAM. The pipeline definition creates the compute clusters, with the AzureML service scaling down inactive clusters, so costs are only incurred for this resource when consumed by active pipelines.

Note that pipeline steps can run in parallel if there are no remaining step dependencies and unallocated computing nodes are available. For example, the step searching for clinical trial references can run independently and simultaneously as the step computing derived features such as publishing year. This decreases the overall pipeline execution time and maximises the use of available resources.

Our proposed solution's literature search feature is provided by a managed Azure Kubernetes cluster inference service (Fig. 11-11), published as a web endpoint. It is configured to use VM type nc6sv3 to provide the high performance necessary for real-time, low-latency surfacing of articles most relevant to end-user questions passed by the COVID-LEAP app (Fig. 11-15). A Python script encodes the question using the deployed neural model before search execution on ElasticSearch, similarity comparison, and relevance scoring. Once again uses our configured docker container provides a consistent runtime environment.

The author appreciated the short learning curve of AzureML, the ease with which bespoke algorithms can be orchestrated with robust, modular pipelines to solve multiple technical challenges, while also finding the Studio web application valuable in monitoring pipeline runs and tracking metrics.

4.10 Presentation

The presentation layer facilitates the interactive exploration of the literature and clinical development data by end-users. It was initially envisaged that the Microsoft Power BI business intelligence platform (Fig. 11-14) would satisfy all presentation layer needs, having successfully prototyped visualisation features tracking the COVID-19 vaccine landscape. However, the requirement to expose an input field for search phrase entry for web service inferencing and article relevancy response was impossible.

A solution was found with Streamlit, a framework to develop interactive web apps with a minimal Python code footprint. We leverage it here for the ease with which Python can call RESTful API web services. A fully managed Azure app service takes care of deploying and scaling the webserver for the COVID-LEAP app (Fig. 11-15). A custom docker container image provides environment dependencies, including the Streamlit Python package and image assets.

4.11 Services

As a collection of ancillary services, this layer supports the development, ingest, compute, and presentation layers. The Azure container registry (Fig. 11-12) stores Docker container images for the development, compute runtime, Airflow, and Streamlit environments. A key vault (Fig. 11-13) securely stores secrets for server credentials to avoid direct reference in code archived in GitHub. COVID-LEAP app uses the built-in security of AAD (Fig. 11-16) to authenticate users. Application insights (Fig. 11-18) logs AzureML pipeline execution metrics for monitoring.

5. Natural Language Processing

5.1 Introduction

Humans use language to share information. Natural Language Processing (NLP) concerns the processing and understanding of human language by machines. At its core, COVID-LEAP is a semantic information retrieval system. It is tasked with finding the most relevant academic paper snippets within a large open-source knowledge base of biomedical literature to answer human natural language form questions. We decompose the problem of extracting knowledge to support clinical research into a step-by-step process. Language texts and trials data flow in, are processed through several data engineering and modelling tasks, and prepared data flows out to storage for applications to consume via an inference web service and database connection. Fig. 14 illustrates this pipeline of numbered tasks, with the green labels indicating the pipeline layer and service component, while blue labels specify the consumed dataset.

As may be observed, some tasks are not particularly related to NLP but are critical preprocessing steps for optimising more complex language modelling activities, such as corpus cleansing to remove incomplete or duplicate titles, and computation of metrics to bias influencing authors and papers. Other tasks with a greater NLP focus include language detection for document filtering and topic modelling for document classification. Extracting vaccine landscape knowledge from PDF e-document and XLS Excel spreadsheet document formats is a pure data engineering activity.

The lifecycle of our neural network-based semantic models, from here now referred to as dense models, is covered. We include an example encoding biomedical text into numerical representation for query encoding and paragraph indexing in our ElasticSearch engine. We detail our approach to dense model fine-tuned to capture better our complex problem-domain semantics and context. We describe our information retrieval inference service that provides ranked search results from the lexical, semantic, and ensemble search strategies being evaluated.

Finally, design considerations and features of the application front-end to optimize the user experience are presented, including filters for result refinement and adjusting ranking order by selecting paper quality metrics calculated in earlier pipeline tasks to favour opinion leaders and important papers. The remainder of this chapter now discusses each of the numbered tasks grouped by data preprocessing pipeline and search.

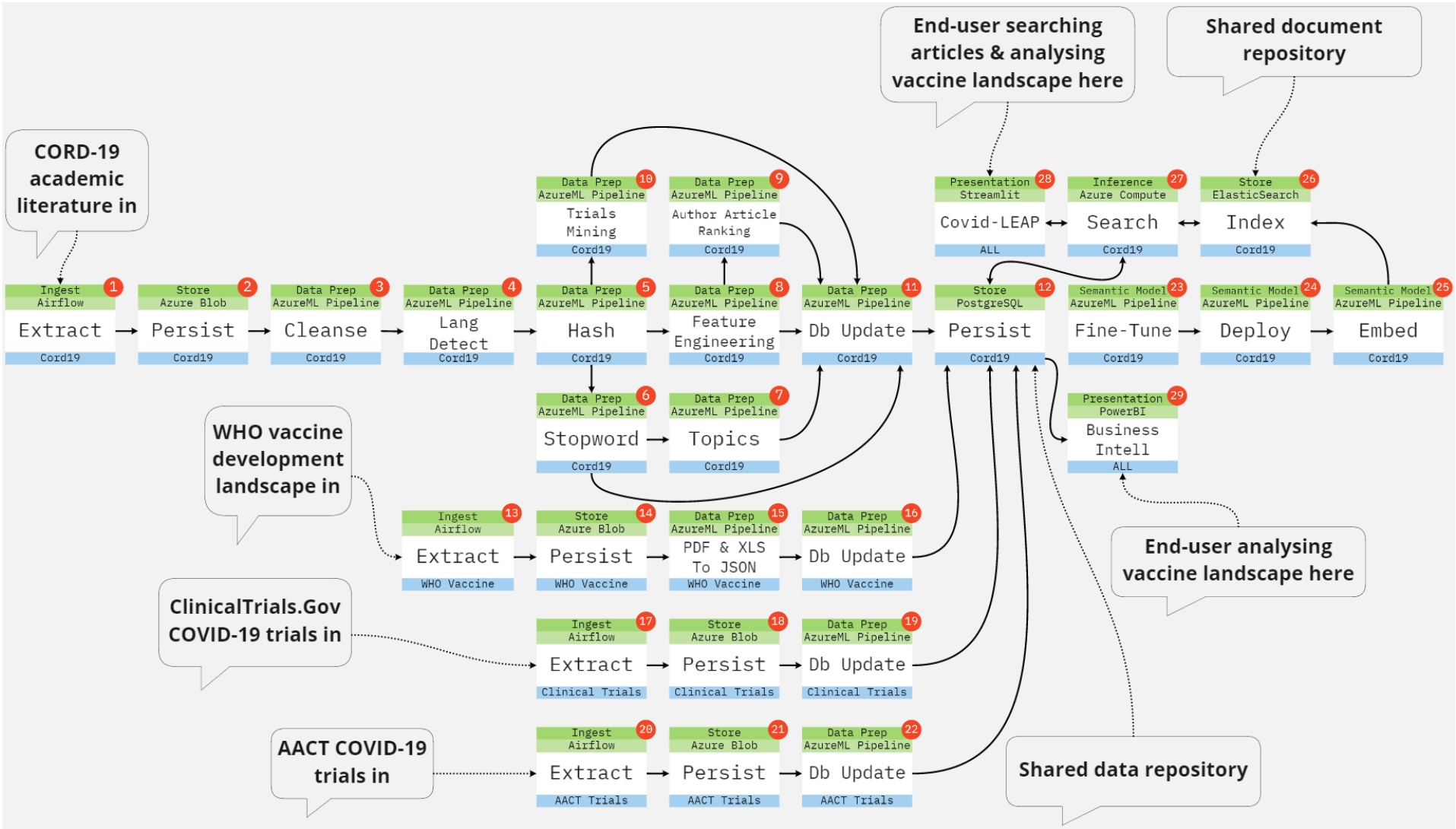

Fig 14: COVID-LEAP NLP pipeline. The central white box names each pipeline step task. The green labels indicate the pipeline layer and service component. Blue labels specify the consumed dataset.

5.2 The Academic Literature Preprocessing Pipeline

Following the collection of a new CORD-19 dataset version by the Airflow component (Fig. 14-1) and persistence into Azure blob storage (Fig. 14-2), a fully automated AzureML pipeline of Python scripted steps is triggered to ingest, prepare and store the academic literature before any consumption by dense modelling and search engine embedding tasks.

Authored and published within a jupyter notebook, the pipeline is defined to take advantage of the availability of multiple processing nodes in Azure compute clusters by executing steps in parallel where there are no direct dependencies. For example, the algorithm for removing stopwords (Fig. 14-6) can run simultaneously to the code mining trial references (Fig. 14-10). Each pipeline step can be considered a transform function configured with input and output parameters to ingest data from proceeding tasks and generate output for subsequent tasks. This enables pipelines to be restarted from failed steps without having to rerun the complete pipeline.

During step execution, logs detail instantiation of the compute cluster runtime with specified docker container runtime environment and information written to the Python console by pipeline steps to support progress tracking and debugging. Plots and runtime statistics coded in the Python script support step validation. Steps of the CORD-19 preprocessing pipeline are now described in the following sub-sections.

5.2.1 Dataset Cleanse

This step (Fig. 14-3) loads data for papers from the PMC (129,700), and preprint server (170,883) JSON files stored in AzureML registered dataset *kaggle-cord19* (Fig. 14-2), creating the new column *subset source* to identify the article's origin. Merging the two file collections is by unique paper ID as assigned by the Allen Institute of A.I.

Despite the dataset authors compiling Cord-19 with steps to harmonize, deduplicate and filter content (L. L. Wang et al., 2020), additional cleansing (Fig. 14-3) prepares the dataset for downstream tasks by execution of the following actions:

- **Resolve abstracts** If no paper abstract from the preferred PMC source, use the preprint version

- **Missing metadata** Drop papers with no title, abstract, body text, or publication date

- **Paragraph irregularities** As papers will be indexed by paragraph for search, drop papers with less than nine or greater than 501 paragraphs to handle overly concise papers or sentences encoded with carriage returns

- **Duplicates** There are many versions of the same paper, especially preprints. Papers are sorted in descending order by paper ID, subset source, and publish time, and then duplicates are removed based on title or paper ID. The first occurrence is retained, which preferences peer-reviewed PMC articles before preprints

- **Removing latex markup commands** Remove commands such as *usepackage* from article title, abstract, and body text with the regular expression list `[r'\\usepackage.*}', r'\\setlength.*}', '{.*}', '{.*pt}', r'\\begin{.*}', r'\\document.*}', 'left(.*)']`

Due mainly to deduplication, 135,092 articles remain from the original 300,583 and are output as a single consolidated comma-separated file for the next pipeline step.

5.2.2 Language Detection

This step (Fig. 14-4) constrains COVID-LEAP language modelling and search to English-only articles. SciSpacy model *en_core_sci_lg*, an NLP pipeline model pretrained on biomedical data, provides the language detection capability. An input parameter specifies the first *n*-word tokens of each paper to be inspected. Testing finds that checking the first 500 words provides the optimal balance between detection accuracy and step runtime. 3257 non-English articles are filtered out.

5.2.3 Hashing

Following article cleansing and filtering, we consider our corpus of academic literature suitably prepared for downstream tasks. A reliable 20-bit SHA (Secure Hash Algorithm) digital signature is generated for each paper using a combination of original paper ID and title (Fig. 14-5) for unique paper identification in tasks including paper ranking by citations and search retrieval.

5.2.4 *Stopwords and Lemmatisation*

In preparing the corpus for topic modelling, another SciSpacy pipeline model strips very frequently used words such as "*a*", "*an*", "*the*", "*for*", and "*of*" from copies of article title and abstract (Fig. 14-6). These words, known as stopwords, only create noise in this article categorisation task, and with their removal, more weight is attributed to keywords of value. An additional custom list that includes "*author*", "*et*", "*table*", "*doi*", "*preprint*", "*figure*", and "*license"* define problem-specific stopwords to cleanse the cord19 corpus further, removing additional tokens of no interest in topic modelling. The article title and abstract are then lowercased and punctuation removed before remaining words normalised to their base *lemma* word, for example, "*children*" to "*child*", "*deaths*" to "*death*", and "*associated*" to "*associate*". This lemmatisation process in effect groups inflected forms of words into a single word to reduce features and streamline topic modelling analysis.

5.2.5 *Topic Modelling*

We propose article title and abstract contain sufficient information to express the overall theme of any given paper. Following stopword removal and lemmatisation, we now use these prepared texts to categorise our extensive academic literature collection into topics (Fig. 14-7). Tagging each paper with a topic metadata label offers the potential to:

- support the subsetting of search results into fields of study for relevancy filtering
- visualise current COVID-19 research trends

COVID-LEAP concatenates article title and abstract for unsupervised topic modelling with the widely used LDA (Latent Dirichlet Allocation) algorithm, implemented in the Python package Gensim. LDA creates two Dirichlet or probability distributions. The first distribution associates the probability of articles belonging to each abstract topic. The second distribution associates each topic with a collection of words, each word weighted to indicate significance. When training an LDA model, the number of topics t is specified. We build models ranging from three to twelve topics, performing an intrinsic evaluation for each using the coherence score metric for quantitative measurement of semantic similarity of top words within a topic. The higher the coherence score, the more cohesive the collection of topic words.

As discussed in the literature review, prior work suggested the optimal number of underlying topics varied widely between eight and fifty. Our testing indicates six topics, with a coherence score of 0.5280, is optimal before flattening out then degrading (Fig. 15).

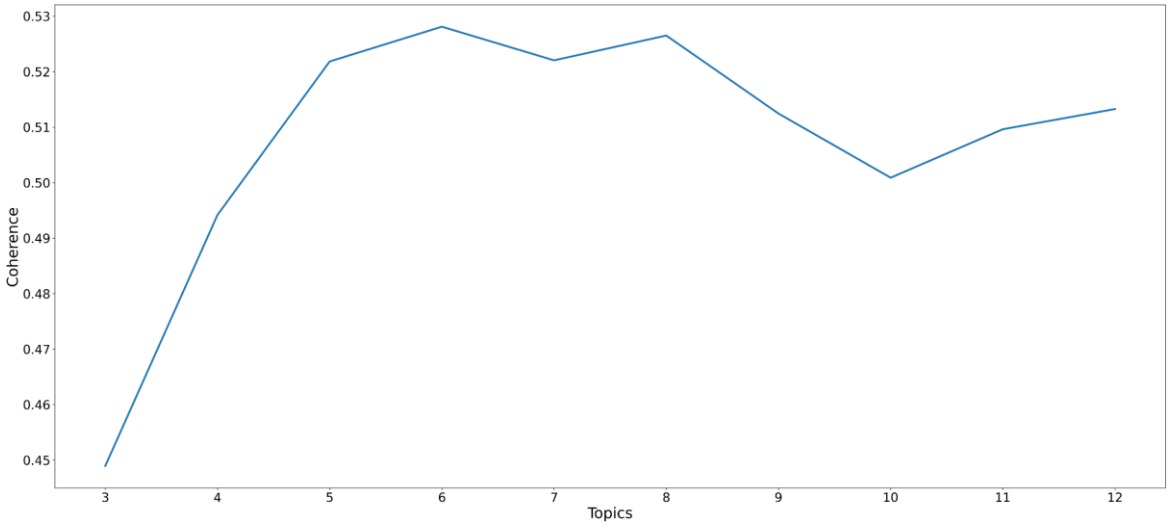

Fig. 15: Topic modelling with LDA - model coherence score by the number of topics

Our model with *t=6* is visualised in two dimensions using Python package pyLDAvis, each topic represented with a numbered bubble (Fig. 16). Bubble size indicates topic dominance within the corpus, and the distance between bubbles the level of topic similarity. A review of their word collection confirms the overlapping of topics four and five, where *"infection*" is a similarly dominant keyword.

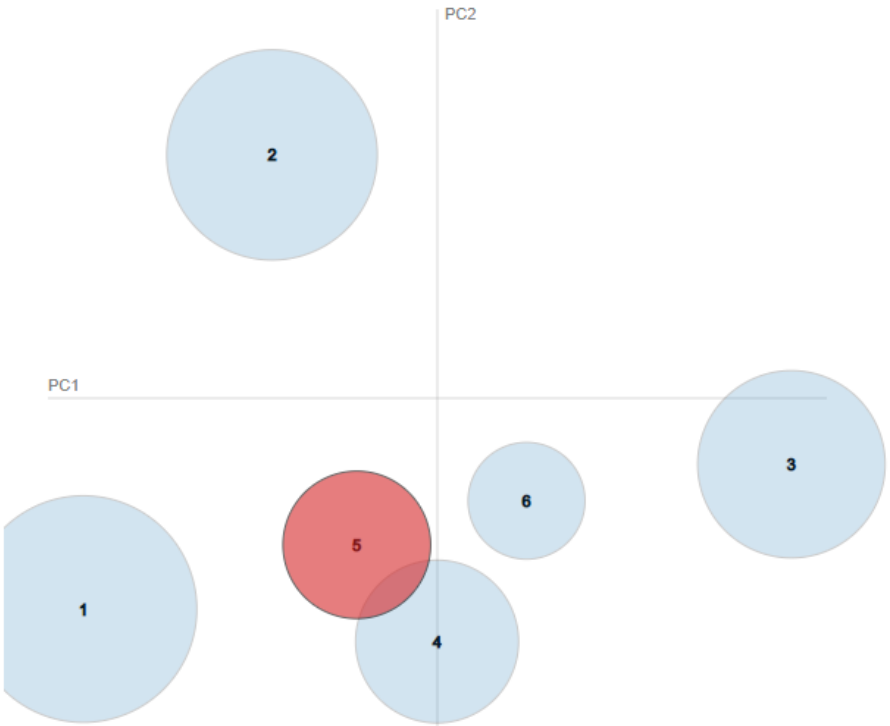

Fig. 16: Principal Component Analysis 2D representation of LDA model topics *t*=6

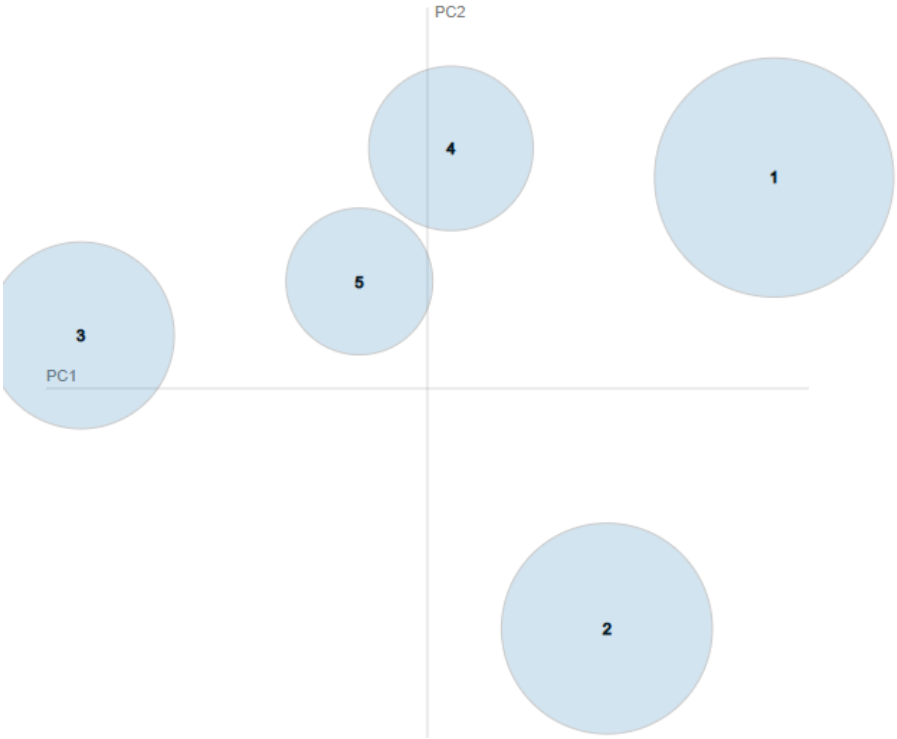

Fig. 17: Principal Component Analysis 2D representation of LDA model topics *t*=5

In this project, we use topic coherence as a starting point to guide the selection of optimal topic numbers, guided by our opinion stated in the literature review that metrics alone should not determine the number of themes. Despite *t=6* resulting in a better coherence score, we opted for the model trained on *t=5* after reviewing its PCA representation (Fig. 17), observing improved intra-cluster topic separation with no overlap. Table 1 presents the topic keyword collection ordered by word significance, which for *t=5* resulted in a more variance of keywords and their weights than for *t=6*. This collection guides the inference of meaningful single field of study labels for each abstract topic ID. Feedback from a medical expert following discussion of these assigned labels is shared in Chapter 6 detailing experiment results.

Topic 1 Clinical		Topic 2 Virology		Topic 3 Public Health		Topic 4 Genomics		Topic 5 Healthcare	
Word	**Weight**	**Word**	**Weight**	**Word**	**Weight**	**Word**	**Weight**	**Word**	**Weight**
disease	0.006	virus	0.019	health	0.013	cell	0.019	patient	0.034
method	0.006	infection	0.010	pandemic	0.010	protein	0.012	disease	0.011
sample	0.005	vaccine	0.009	model	0.007	virus	0.011	infection	0.008
drug	0.005	human	0.008	disease	0.006	infection	0.010	clinical	0.008
virus	0.005	respiratory	0.007	social	0.005	viral	0.008	severe	0.007
detection	0.004	antibody	0.007	risk	0.004	response	0.006	risk	0.006
based	0.004	viral	0.006	public	0.004	gene	0.005	respiratory	0.006
system	0.004	disease	0.006	care	0.004	expression	0.005	treatment	0.006
model	0.004	influenza	0.005	country	0.004	human	0.005	hospital	0.006
clinical	0.004	strain	0.004	impact	0.004	immune	0.005	care	0.005
assay	0.004	analysis	0.004	measure	0.004	disease	0.005	mortality	0.005
infection	0.004	patient	0.004	outbreak	0.004	host	0.004	acute	0.005
review	0.003	sequence	0.004	population	0.004	rna	0.004	associate	0.005
analysis	0.003	cause	0.004	analysis	0.004	role	0.004	outcome	0.005
pcr	0.003	protein	0.004	epidemic	0.003	receptor	0.004	pandemic	0.005
potential	0.003	cell	0.004	infection	0.003	target	0.004	age	0.004
treatment	0.003	sample	0.003	research	0.003	mechanism	0.003	symptom	0.004
pandemic	0.003	pathogen	0.003	spread	0.003	mouse	0.003	analysis	0.004
vaccine	0.003	species	0.003	system	0.003	model	0.003	results	0.004
present	0.003	transmission	0.003	control	0.003	ace	0.003	pneumonia	0.004

Table 1: LDA model t=5 with assigned label and weighted word collection

As suggested, having a topic assigned to each article allows for statistical analysis of the topic distribution over time using publication dates, showing the increased popularity of research into public health (Fig. 18).

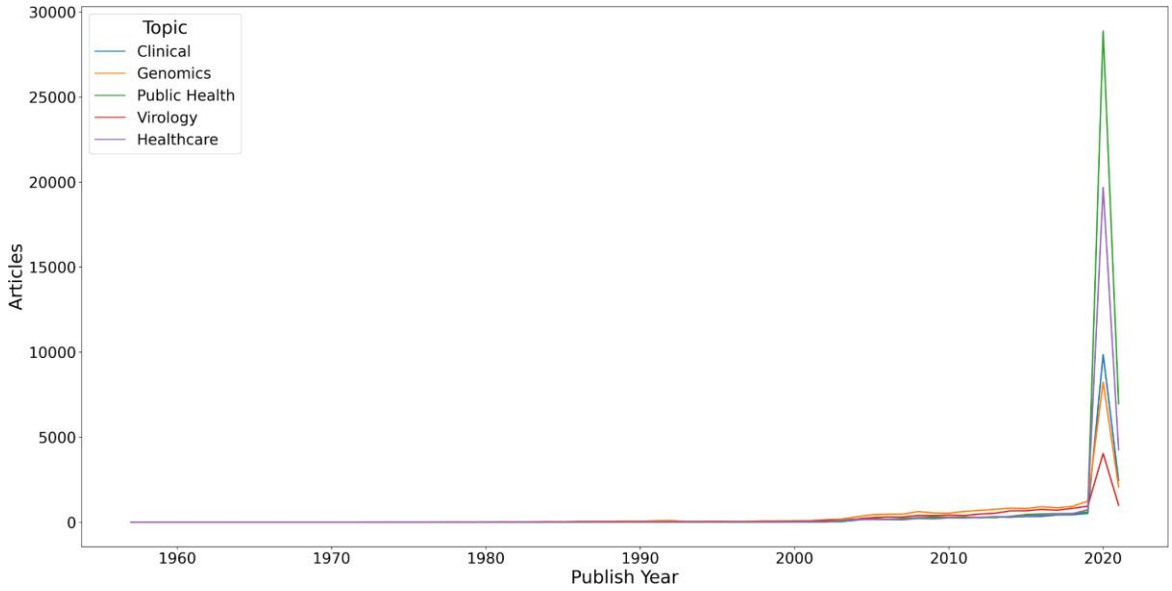

Fig. 18: CORD-19 article topic classification – topic temporal evolution

This custom pipeline step for topic modelling can be configured to build a model with a predefined topic number or optimize by selecting the highest coherence scoring variant from a topic number range. Note that given LDA modelling is a probabilistic method relying on random number generation when creating Dirichlet distributions, any top-scoring models are saved and reloaded for final topic allocation in auto-tuning mode. Persisting the model ensures the same word collection, weighting, and coherence score characteristics for both model auto-tuning evaluation and final deployment. Furthermore, this auto-tuning feature manages corpus drift as the research landscape evolves. However, we stress the importance of domain experts reviewing changes to the optimal topic number with re-validation and consensus of topic labelling.

5.2.6 Feature Engineering

This step (Fig. 14-8) synthesizes additional helpful information from the existing data. By taking the first four characters of column *publish_time*, the new column *publish_year* simplifies subsetting the corpus by year in visualisation and filtering tasks. Each article title and full text are searched with a set of regular expression heuristic rules to characterise the significance of their association with coronavirus family variants. As an example, the following regex searches for keywords related explicitly to the Sars-CoV-2 virus. Note that casing is ignored. The escape code *\b* defines the bounds of a whole string, preventing "*19*" flagging as false positive. In total, eight additional columns enrich each document with additional metadata that can be used to subset results.

```
r'\bcovid 19\b|\bcovid-19\b|\b2019-nCoV\b|\bsars-cov-2'
```

5.2.7 *Author and Article Ranking*

Preprint servers as channels of COVID-19 research sharing have seen a massive increase in popularity. However, there is concern that the circumventing of the strict peer-review process adopted by traditional journals allow the dissemination of lower quality knowledge. COVID-LEAP introduces quality indicators (Fig. 14-9) with computation of the three metric groups listed below. These bibliometrics are integrated into a paper ranking component of our search strategy to weight paper scores.

1. **Citation counts** measuring paper popularity and surfacing the research of trusted, well-funded authors, beneficial to researchers new to the domain

2. **PageRank algorithm scores** evaluating the influence of authors and papers in their research network

3. **Recency** to bias newer papers

Illustrated by Fig. 19, this is one of the more complex pipeline step scripts with over 700 lines of code, primarily using the Python package Pandas to wrangle author and paper metrics by subsetting and aggregating our tabular literature dataset. We break down this ranking problem into several sub-tasks now described in turn.

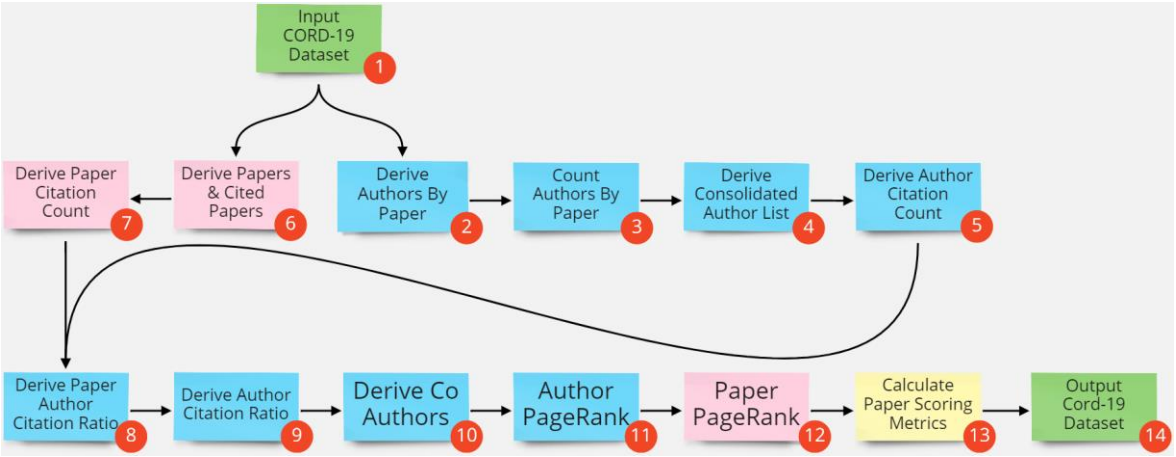

Fig. 19: Process flow for computation of author & article ranking metrics. **Green** indicates CORD-19 data in/out. **Pink** processes bibliographies & compiles paper-based statistics. **Blue** processes authors & compiles author-specific statistics. **Yellow** consolidates corpus statistics to compute quality metrics.

Following the load of the literature corpus (Fig. 19-1), the raw paper author string is transformed into a vertical list of authors indexed by paper hash, allowing easier aggregation counts of papers by an author (Fig. 19-2). Names are cleansed by lowercasing, stripping whitespace and punctuation, and removing duplicate authors from single papers. A hash is now required to uniquely identify the author, presenting us with the problem of author disambiguation, which attempts to deal with multiple instances of the same name, for

example, "*Joe Brown*". This is a highly challenging problem, primarily due to the lack of additional identifying information such as individual email addresses or institute affiliation. However, if left untouched, authors will be accredited with citations for papers they did not write, introducing a significant and flawed paper ranking bias.

With COVID-LEAP, we take a pragmatic approach to identifying unique individuals by creating the author hash using a combination of author name and journal. While we fully appreciate that many authors diversify their target journals to broaden their impact and audience, in this first attempt, we follow the guidance of (K. Wang et al., 2020) with their design choices for author disambiguation within the Microsoft Academic Graph (MAG) solution for literature search. They err on the side of caution, only assigning papers to authors if the association exceeds a high confidence threshold. They state their "*design choice leads to the fact that the publication count of an author node in MAG can only be lower than the actual number of publications by the real-world author*". We acknowledge that combining the author's name with the journal is rudimentary and overly strict; however, the resulting author's underconflation is preferred.

We now aggregate the authors associated with each article (Fig. 19-3), providing additional author count metadata for the weighting of citations based on the authors' contribution level. After dropping author duplicates by hash (Fig. 19-4), a paper citation count is assigned to each author entry by aggregating the authors by paper list generated in Fig. 19-2.

Deriving paper citations begins with the processing of article bibliography metadata (Fig. 19-6) to compile an index of referenced and referencing papers. In other words, what papers are citing other papers in the CORD-19 corpus. The same hash algorithm used for unique indexing of paper titles in Fig. 11-5 now hashes cited titles to facilitate linking bibliography entries with corresponding main articles, with a simple cite count by main paper informing paper popularity (Fig. 19-7). As a side note, due to the corpus size and compute resource memory limitations, it was necessary to parse the bibliography in chunks of 10,000 papers. Next, a "*paper author citation ratio*" is calculated (Fig. 19-8), per paper, by dividing the paper citation count by the paper author count. An "*author citation ratio*" then aggregates this paper author citation ratio per author that characteristically favours single authors and small collaborative teams with an extended publication history.

An index of authors and co-authors generates the core information for an author network (Fig. 19-10), complementing the index of referenced and referencing papers. A standard undirected graph of author nodes with edges between them defining co-authorship is

assembled for scoring author collaboration with the PageRank algorithm (Brin and Page, 1998), invented by the Google founders to measure the link popularity of web pages. PageRank then consumes a second graph of directed edges from citing paper to cited paper to infer papers of influence.

We experiment with the following four functions (mf*n*) using metrics including PageRank scores and citation ratios as measures of paper quality, with terms *p* denoting paper, *a* author, *c* citations, *r* PageRank, and explicit w_β, w_γ *and* $w\varepsilon$ weightings assigned to the associated custom metric term. Adjusting rank score purely on classical paper citation count is computed as follows:

$$mf1 = c_p w_\beta$$

Our next metric integrates the popularity and contribution level of authors in the Cord-19 research space by aggregating the sum of each author's paper citation ratio irrespective of paper:

$$mf2 = \left(\sum_a \frac{c_p}{\sum a_p}\right) w_\beta$$

Incorporating the influence of papers and authors via their PageRank scores, we compute *mf3* as:

$$mf3 = \left(r_p w_\beta\right) + \bar{x}(\sum r_a) w_\gamma$$

Finally, and similar to *mf3* with PageRank scoring but adding the concept of paper recency to bias newer papers, *mf4* is computed as follows, where y_{curr} denotes current year and y_{publ} a paper's published year:

$$mf4 = \left(r_p w_\beta\right) + \bar{x}(\sum r_a) w_\gamma + \left(\left(\frac{1}{\sqrt{y_{curr} - y_{publ}}} + 1\right) w\varepsilon\right)$$

The paper-level metrics are normalised with the *MinMax* scaler from Python package Sklearn to transform ranges between 0 and 1 then persisted as output from this pipeline step, indexed into the search engine (Fig. 11-26), and consumed by the COVID-LEAP application (Fig. 1-28) to weight model relevancy scores with a quality adjustment.

Reviewing the calculated metrics, we share selected observations. Astrophysics dominates the top five papers by the highest number of contributing authors, including "*Diving below the spin-down limit: Constraints on gravitational waves from the energetic young pulsar PSR J0537-6910*" with 1591 authors. Clearly, there is an opportunity for improving the curation of the CORD-19 corpus. While the dominance of significant collaboration efforts is dampened by taking the average and not the sum of author PageRank in functions *mf3* and *mf4*, derived features such as *is_sars_cov2* to associate papers with coronavirus family variants should filter these out in search.

"*Clinical features of patients infected with 2019 novel coronavirus in Wuhan China"*, published in *The Lancet* in February 2020, is the most cited paper with 7477 bibliographic references, followed by "*Clinical Characteristics of Coronavirus Disease 2019 in China*" published in *The New England Journal of Medicine* with 5619 citations (Table II). Notably, these are peer-review articles shared in well-respected journals. Author *Zhang, Y.*, whose work includes "*The Incubation Period of Severe Acute Respiratory Syndrome Coronavirus 2:A Systematic Review*", makes the largest corpus contribution with 53 papers. Another paper by the same-named *Zhang, Y.*, titled "*Characteristics of air pollutant dispersion around a high-rise building*" is hashed differently due to publication in the journal "*Environmental Pollution*". Given the research domain of this second paper, the use of journals to underconflate paper attribution is justified.

Unsurprisingly, the top 5 papers by citation count match PageRank and associated *mf1* scores (Table 2), all recent peer-reviewed PubMedCentral articles published in 2020. Metric *mf2* introduces author influence (Table 3). Favouring well-established authors with a long history of publication and those collaborating within smaller research groups, the top-scoring paper is "*Mechanisms of transmissible gastroenteritis coronavirus neutralization*", first published in August 1990 in Virology. The paper's authors include Mariaje Bullido and Carlos Sune, cited 14570 and 1947 times respectively, according to Google Scholar. They have well-established track records for coronavirus research dating back to the 1990s, with other work including "*Location of antigenic sites of the S-glycoprotein of transmissible gastroenteritis virus and their conservation in coronaviruses*" and "*Induction of an immune response to transmissible gastroenteritis coronavirus using vectors with enteric tropism*". The second highest paper is "*Subcellular location and topology of severe acute respiratory syndrome coronavirus envelope protein*", published in Virology in 2011, with author ML DeDiego attributed 3814 citations.

Metric *mf3* combines PageRank for paper and author link popularity, averaging the author score to combat outliers with many authors (Table 4). While the paper "*Clinical features of patients infected with 2019 novel coronavirus in Wuhan China*" again scores highest, as per top citations and paper PageRank, what is of interest here are papers scoring 4th and 5th positions. Both works have zero citations, inferring a corresponding paper PageRank score of zero. However, upon verification, one co-author of both these papers is *J. Li*, the third-highest contributing author by paper count with 48 publications, primarily pre-preprints involving ten or more authors. An article's publication date is considered by metric *mf4* that coerces PageRank scores with a bias towards more recent research, with the top 5 papers all preprints published in 2021 (Table 5).

With 123K papers, 601K authors, and a bibliography of 5.3M citations, compilation of the discussed quality metrics over such a large corpus is compute-intensive. Considerable effort was made to optimise this pipeline step to maximise the efficiency of available cloud resources. We implemented *Modin*, a Python package to parallelize vector computations over all available CPU cores. In initial testing of paper and author aggregation for metric calculations, we exceeded the 64GB local development system memory. Deleting temporary dataframes directly after aggregations, followed by explicit execution of the Python garbage collector, enforced continuous memory management, released memory to other calculations, and resolved the memory allocation issues.

Title / (Ref)	Year	# Citations	# Authors	MF1
Clinical features of patients infected with 2019 novel coronavirus in Wuhan China (PMC7159299)	2020	7477	29	1
Clinical Characteristics of Coronavirus Disease 2019 in China (PMC7092819)	2020	5619	19	0.7826
Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study (PMC7270627)	2020	5303	14	0.7092
Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study (PMC7135076)	2020	4590	29	0.6138
A pneumonia outbreak associated with a new coronavirus of probable bat origin (PMC7095418)	2020	4070	18	0.5443

Table 2: Top 5 papers as measured by quality metric MF1 citation count

Title / (Ref)	Year	# Citations	# Authors	MF2
Mechanisms of transmissible gastroenteritis coronavirus neutralization (PMC7131644)	1990	16	7	1.0
Subcellular location and topology of severe acute respiratory syndrome coronavirus envelope protein (PMC4726981)	2011	66	9	1.0
Transmissible gastroenteritis coronavirus gene 7 is not essential but influences in vivo virus replication and virulence (PMC7126239)	2003	47	8	1.0
Transmissible gastroenteritis virus (TGEV)-based vectors with engineered murine tropism express the rotavirus VP7 protein and immunize mice against rotavirus (PMC7111951)	2011	5	6	1.0
Antigenic homology among coronaviruses related to transmissible gastroenteritis virus (PMC7130632)	1990	42	11	1.0

Table 3: Top 5 papers as measured by quality metric MF2 author citation ratio

Title / (Ref)	Year	# Citations	# Authors	MF3
Clinical features of patients infected with 2019 novel coronavirus in Wuhan China (PMC7159299)	2020	7477	29	1
Disposable N95 Masks Pass Qualitative Fit-Test But Have Decreased Filtration Efficiency after Cobalt-60 Gamma Irradiation (10.1101/2020.03.28.20043471)	2020	5	8	0.9799
Identifying Synergistic Interventions to Address COVID-19 Using a Large Scale Agent-Based Model (10.1101/2020.12.11.20247825)	2020	2	2	0.9797
Dynamic changes in human single cell transcriptional signatures during fatal sepsis (10.1101/2021.03.01.21252411)	2021	0	7	0.9795
Preparing For the Next Pandemic: Learning Wild Mutational Patterns At Scale For For Analyzing Sequence Divergence In Novel Pathogens (10.1101/2020.07.17.20156364)	2020	0	3	0.9795

Table 4: Top 5 papers as measured by quality metric MF3 combined author and paper PageRank

Title / (Ref)	Year	# Citations	# Authors	MF4
Dynamic changes in human single cell transcriptional signatures during fatal sepsis (10.1101/2021.03.01.21252411)	2021	0	7	1
Numerical study of COVID-19 spatial–temporal spreading in London (10.1063/5.0048472)	2021	0	10	1
Returning to a Normal Life via COVID-19 Vaccines in the United States: A Large-scale Agent-Based Simulation Study (10.2196/27419)	2021	0	2	1
Phased implementation of COVID-19 vaccination: rapid assessment of policy adoption, reach and effectiveness to protect the most vulnerable in the US (10.1101/2021.02.19.21252118)	2021	0	5	0.9621
Home stay reflects symptoms severity in major depressive disorder: A multicenter observational study using geolocation data from smartphones (10.1101/2021.02.10.21251512)	2021	0	24	0.8757

Table 5: Top 5 papers as measured by quality metric MF4 combined author/paper PageRank with recency

5.2.8 *Extracting trials*

Searching the CORD-19 literature collection for trial identifiers (Fig. 14-10) allows us to answer questions that require knowledge beyond a single dataset by connecting papers with the WHO COVID-19 vaccine trials landscape and detailed clinical trial data from ClinicalTrials.Gov. In COVID-LEAP, regular expression (regex) pattern matching mines the unstructured academic literature for referenced clinical trial identifiers.

For example, the regex `NCT[0-9]{8}` deterministically searches for a National Clinical Trial in text beginning "*NCT*" followed by eight characters in the range 0 to 9. A more complex regex example is `EUCTR[0-9]{4}-[0-9]{6}-[0-9]{2}-[A-Z]{2}` searches for references to the E.U. clinical trials register. Output is a two-column index of related paper hash IDs and trial IDs.

5.2.9 *Database Update*

This final step in the academic corpus processing pipeline is responsible for persisting the output from multiple prior steps, including feature engineering and topic modelling, into an Azure-hosted PostgreSQL database (Fig. 14-11). First, the database connection is authenticated with credentials from docker container runtime OS environment variables. Next, all CORD-19 relational model tables such as *pub_article* and *pub_document_topic* are dropped, as each pipeline run is triggered by a full dataset refresh. Due to the corpus size, the database cannot be updated with the full corpus in one single execution of the update command. Updates are done in blocks to prevent performance issues, with the first table chunk executed in *refresh* mode and subsequent chunks in *append* mode.

5.3 The COVID-19 Development Landscape Preprocessing Pipeline

Steps of the preprocessing pipeline to provide COVID-LEAP with the data required to provide the researcher with an interactive, visual, and insightful exploration of the COVID-19 development landscape are now described in the following sub-sections.

5.3.1 WHO Vaccine Landscape Extract

As discussed in Chapter 3, every week the WHO updates its view of the COVID-19 candidate vaccine landscape. This collection of publications is information-rich, detailing the enterprises funding vaccine research and how advanced their clinical developments are. From it, we can visualise the temporal evolution of specific trials and any dominance of vaccine platforms. Unfortunately, with release in a mix of PDF and XLS formats intended only for the human reader, it is challenging to robustly consume by machine.

After extraction by Airflow (Fig. 14-13) and persistence into Azure blob storage (Fig. 14-14), the two-step ingestion process first transforms the landscape knowledge into a consistent, more generic, and source-independent, machine-readable JSON format (Fig. 14-15). Features include:

- **Parses PDF tables** to extract vaccine information, using Python package Camelot
- **Detects PDF and XLS structure** as the layout of vaccine information in both file types is inconsistent and evolving
- **Identifies pre-clinical vs. clinical layout** as structure and associated metadata differs between trial study types
- **Cleanses** with string processing to remove carriage returns, tabs, white space, and Unicode characters
- **Handles inconsistent notation** such as multiple trial identifiers in the same PDF table cell, "*not yet recruiting*" and "*Interim Report*" in phase columns intended for trial identifiers, "*" characters after trial identifiers which refer to footnotes, non-numeric information such as dates in the "*number of doses*" column

A hash is generated using a combination of trial type, manufacturer, platform, and vaccine type to uniquely identify and index each trial. The generated JSON file per publication is uploaded to Azure blob storage.

5.4 WHO Vaccine Landscape Database Update

We now have a cleansed JSON representation of WHO vaccine landscape publications in a generic trial format. After database connection, each file is processed (Fig. 14-16). Observing inconsistencies in naming and capitalisation of vaccine platforms, we again leverage a set of heuristic regular expressions to consolidate, an example given below:

```
(.*?)live attenuated virus(.*?),(.*?)live attenuated
bacterial(.*?) > SARS-CoV-2 Virus: Live Attenuated
```

We observed one instance of trial duplication from the WHO publication dated 2nd March 2021. Safeguarding against unique key violation, duplicates are dropped before database update in two steps: first, all *candidates* that are clinical trials irrespective of study type are indexed into PostgreSQL using generated hash and publication date. Second, *clinical stages*, with phase information explicitly related to clinical trials, are indexed.

5.5 Clinical Trials Database Update

Trials knowledge extracted from the clinical trials registry *ClinicalTrials.Gov* (Fig. 14-17) and persisted into Azure blob storage (Fig. 14-18) is now indexed into PostgreSQL table *clinicaltrialsgov* (Fig. 14-19). Minor cleansing tasks include dropping entries without a valid NCT number, replacing null values, and correct data typing of columns containing dates to support downstream filtering and time-series analytics.

5.6 AACT Trials Database Update

To be fully informed on the current state of the COVID-19 vaccine landscape, researchers require access to key characteristics of clinical studies, including eligibility criteria for study group participation and technical vaccine names. Unfortunately, *ClinicalTrials.Gov* extracts contain only a subset of trials information, so with COVID-LEAP, we supplement it with ingestion of the *ACCT* database (Fig. 14-22). As an extensive collection of 46 text files, we process only a subset targeting our additional information needs. The remainder of this subsection reviews a sample of tasks in this pipeline step related to string processing.

As AACT contains trials not exclusively related to COVID-19, we assemble a master index of trials of interest by filtering fields *brief title* and *official title* with the following pattern to constrain selected information from other files:

```
COVID-19|covid|sar cov 2|SARS-CoV-2|2019-nCov|2019 ncov|SARS
Coronavirus 2|2019 Novel Coronavirus|coronavirus 2019| Wuhan
coronavirus|wuhan pneumonia|wuhan virus|China virus
```

Of interest is the qualifying age range of study participants. First, potentially useful records mentioning characteristics of study individuals are selected from the *design_groups* file with wildcards in the *LIKE* clause of the example SQL expression below:

```
title like lower('%%aged%%') or title like
        lower('%%participants%%')
```

A second pass with regex pattern matching further filters for eligibility information, an example snippet below:

```
                    '\d\d-
\d\d\syears','adults(.*?)up','between(.*?)old','aged(.*?)old'
                    ,'\((.*?)older\)'
```

Technical vaccine names are mined from the *interventions* file, using regular expressions to omit any drugs and biological entries matching specific terms, as shown below. Multiple drug names are concatenated (Fig. 7).

```
'(.*?)placebo(.*?)','(.*?)control(.*?)','(.*?)blood(.*?)','(.
                    *?)control(.*?)'
```

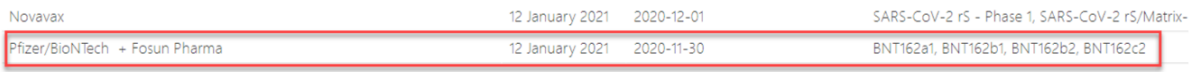

Fig. 20: Extraction and concatenation of technical drug names from AACT interventions file

Other AACT files consumed include *eligibilities* for study cohort gender, *custom values* for trial primary completion date, and *brief summaries* for indexing trials information in our search engine. Before persisting to our PostgreSQL database, specific data typing of columns containing dates supports downstream analytics.

5.7 Search

5.7.1 Introduction

Following the preparation of our domain knowledge, we now address the main task of fulfilling a biomedical information need. The document collection is texts of academic research queried with a scientific question to retrieve the most relevant text to satisfy the information need. With over 137K papers, we not only need to locate matching papers but also rank results in order of relevancy to the question.

Our work experiments with the three main retrieval strategies identified in the literature review. We first establish a baseline approach with the industrial-standard BM25 (Best Matching) lexical keyword matching method. As we identified a trend towards deep-learning neural models for language modelling, we next implement a transformer-based encoder to evaluate if models trained to capture semantics deliver better retrieval performance than traditional surface-level keyword matching. Our final strategy experiments with variations of ensemble approach in which candidate document selections from a first model are re-scored by a second model.

First, however, we momentarily step back from referring to our NLP pipeline (Fig. 14), as we first provide context by describing retriever model architecture, illustrate text embedding essentials, and outline selected dense models for experimentation before jumping back to the pipeline with model fine-tuning.

5.7.2 Lexical Model

The foundational component of our information retrieval architecture is ElasticSearch ("Elasticsearch," n.d.), an engine designed to store documents for search and used by companies including Audi, P&G, Netflix, Uber, and Microsoft. By default, ElasticSearch indexes documents with BM25 (Robertson and Zaragoza, 2009), a distributional bag-of-words method that maintains frequency counts in indexes of document-word (term) frequency pairs. During search, these indexes are referenced to subset the corpus in relevant and irrelevant documents according to term-matching with the question. Documents are scored by the sum of their matching term frequencies, with greater weighting given to "*elite*" words considered highly relevant to documents and less weighting to common stopwords such as "*the*" and "*in*".

A benefit of BM25 is low latency between search execution and response, typically sub-second, as the pre-calculated document-term frequency indexes are used for relevancy scoring, negating the need for any additional read of the text collection. Unlike state-of-the-art dense models, BM25 does not consider word order and, therefore, context. However, given its adoption in enterprise search and strong performance in several information retrieval tasks (Guo et al., 2020, Yang et al., 2019, Rosa et al., 2021), we use this model to establish a baseline.

5.7.3 *The Transformer Model*

The literature review highlighted the popularity and effectiveness of dense models in representing specialised domain text for natural language tasks, including information retrieval. This project evaluates if the ability of these model architectures based on transformers to go beyond keyword matching by capturing semantic meaning offers better information retrieval performance.

The seminal paper "*Attention is all you need*" (Vaswani et al., 2017) by researchers at Google first introduced the transformer for language modelling as a simplified neural architecture, with two main components: the encoder stack, which accepts input, for example, a source English language phrase, and the decoder stack to generate the output probabilities for a task, such as target German language translations (Fig. 7). With a bi-directional attention mechanism in the encoder for improved word embedding by contextualising with surrounding words, state-of-the-art performance was demonstrated in several NLP tasks.

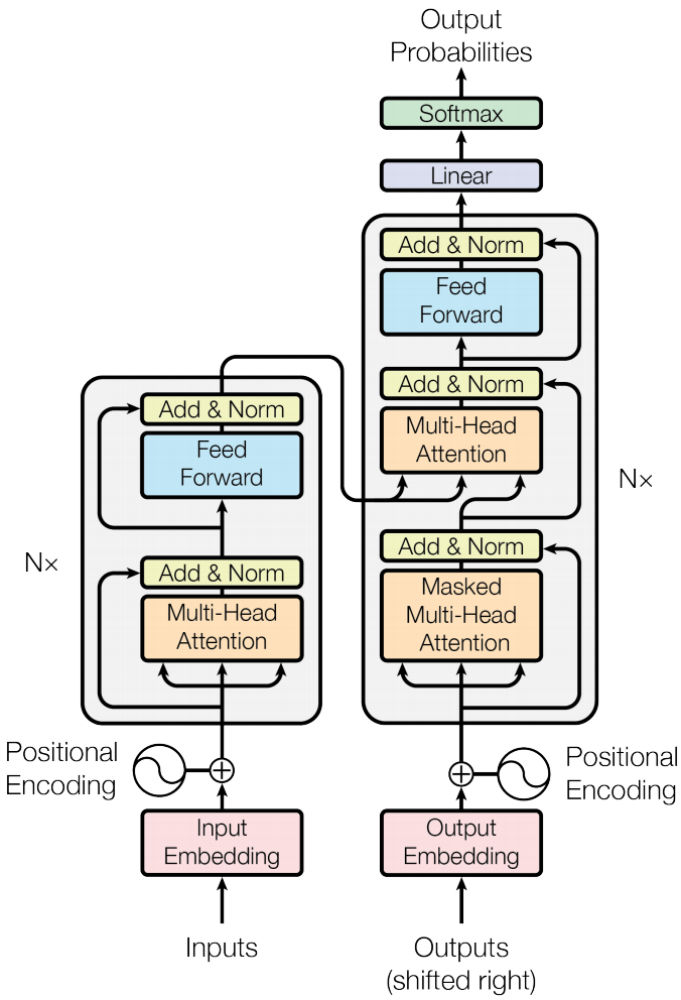

Fig. 21: The Transformer – model architecture (Vaswani et al., 2017)

Transformer models learn language representation by training on large text datasets such as Wikipedia. A phrase is split into unique terms by a tokenizer, assigned an identifier, and stored in a model vocabulary. A word embedding matrix represents each token with a vector of real numbers initially random but refined during model training with the attention mechanism. It is attention that gives words their context by calculating the scaler dot product between all combinations of tokens within a sequence and passing them through a SoftMax activation function. This amplifies strong word relationships with a higher dot product combination while dampening weaker word relationships with a lower dot product. The vector representation of each sequence token is adjusted by the SoftMax output of its pairings, in effect contextualising it by its relationship with every other word in the sequence, and adding positional encoding to preserve word order.

Through many training iterations, words of similar meaning move in the same direction towards shared concepts such as healthcare or genetics within an *n*-dimensional vector space. Vaswani et al. reported that the training of their transformer models on 4.5 million sentence pairs for English to German translation took between 12 hours and 3.5 days on 8 Nvidia P100 GPUs, depending on model *n*-dimension size and hyperparameter configuration.

With COVID-LEAP, we use three variants of the transformer. We primarily focus on high-quality in-domain semantic representation with an acceptable resource cost to compute embeddings for questions and a large corpus of article paragraphs. We have no requirement to decode vectors back to any target text. Therefore, we use an encoder-only transformer model to aggregate word vectors in a sequence with a mean pooling strategy to average each word dimension to get a final representation of the title-paragraph passage combination. By encoding the biomedical question with the same model, we can make a mathematical relevance comparison between the question vector and all passage vectors, with representations of semantically similar questions and article passages closer in vector space. Our second applied transformer variant uses the complete encoder-decoder dual-stack to generate query-answer pairs for dense model fine-tuning. Our third variant is a cross-encoder for second-stage dense re-ranking with only an encoder stack.

5.7.4 *BERT-based Transformer Models*

Further transformer-related research at Google produced BERT (Bi-Directional Encoder Representation from Transformers) (Devlin et al., 2019), a model with only an encoder stack that comes pre-trained on two unsupervised tasks. The first, Masked Language Model (MLM), helps BERT learn in-sentence context by replacing random words in a sequence with a mask which the model tries to predict using words to the left and right of the mask. For example, with the *bert-base-uncased* model, the following input gives output probabilities for the work tokens "*viruses*" (0.401), "to" (0.190), "*they*" (0.127), "*mutations*" (0.057), and "*genes*" (0.031).

```
[['CLS'], 'knowledge', 'of', 'how', '[MASK]', 'impact',
'the', 'SARS-CoV-2', 'RBD', 'would', 'aid', 'efforts', 'to',
'understand', 'the', 'evolution', 'of', 'this', 'virus', '.',
'[SEP]']
```

The second training task, Next Sentence Prediction (NSP), has the model predicting the probability of a second sentence following the first in a sentence pair, helping the model understand the context between sentences. To reduce model vocabulary size yet still cover the English language, the BERT tokenizer uses the WordPiece method to split words into subunits which "*naturally handles rare words, and ultimately improves the overall accuracy*" (Wu et al., 2016). Consider the following phrase:

"*What T-cell epitopes been identified in the Receptor Binding Motif (RBM) region of the S-glycoprotein Receptor Binding Domain (RBD) of the SARS-CoV-2 virus?*"

Tokenized by WordPiece, it is split as below, with subunits indicated by the # symbol:

```
'what','t','-
','cell','ep','##ito','##pes','been','identified','in','the',
'receptor','binding','motif','(','rb','##m',')','region','of'
,'the','s','-',
'g','##ly','##co','##pro','##tein','receptor','binding','doma
in','(','rb', '##d',')','of','the','sar','##s','-
','co','##v','-','2','virus','?'
```

To facilitate downstream NLP tasks, including similarity comparison, we know dense models represent text as vectors of real numbers. To briefly demonstrate and highlight some characteristics of dense models, we begin by creating a model and tokenizer:

```
model = AutoModel.from_pretrained("distilbert-base-uncased")

tokenizer = AutoTokenizer.from_pretrained("distilbert-base-
uncased")
```

Retrieving the model vocabulary, we can verify its length as 30522 tokens:

```
v = tokenizer.get_vocab()

len(v) > 30522
```

We can get a word vocabulary vector back from the BERT tokenizer for our phrase:

```
phrase_tokens = tokenizer("What T-
cell epitopes been identified in the Receptor Binding Motif (
RBM) region of the S-
glycoprotein Receptor Binding Domain (RBD) of the SARS-CoV-
2 virus?", return_tensors="pt")
```

The vector below is returned, where id 101 is the token [CLS] to represent the sequence start, and 102 the [SEP] token at the sequence end:

```
[101,2054,1056,1011,3526,4958,9956,10374,2042,4453,1999,1996,
10769,8031,16226,1006,21144,2213,1007,2555,1997,1996,1055,
1011,1043,2135,3597,21572,9589,10769,8031,5884,1006,21144,
2094,1007,1997,1996,18906,2015,1011,2522,2615,1011,1016,
7865,1029,102]
```

To confirm word vocabulary IDs, we can directly look up the word cell, which returns ID 3526, the 5th vector dimension above:

```
v['cell'] > 3526
```

We can use the dense model to encode a text sequence:

```
model.encode("cell")
```

Which returns the vector representation of it below, the number of dimensions determined by the shape of hidden nodes in the model, in this example 768 for *bert-base-uncased*.

```
[2.75568098e-01, -1.55317977e-01, -1.87793612e-01,…]
```

Table 6 presents the main architectural characteristics of dense models used in COVID-LEAP, including the shape of hidden nodes that directly determines the size of paragraph vectors and ElasticSearch index sizing requirements.

Base Model	Hidden Nodes	Layers	Params	Attention Heads	Vocab Size
Query & Passage Embedding					
distilbert-base-uncased	768	6	66M	12	30K
sentence-transformers/msmarco-distilbert-base-v3	768	6	66M	12	30K
sentence-transformers/ce-ms-marco-TinyBERT-L-4	312	4	14M	12	30K
castorini/ance-msmarco-passage	768	12	124M	12	50K
Query Generation For In-Domain Training					
BeIR/query-gen-msmarco-t5-large-v1	1024	24	737M	16	32K
Cross-encoder Reranking					
cross-encoder/ms-marco-MiniLM-L-12-v2	384	12	33M	12	30K

Table 6: COVID-LEAP experimental dense model architecture

We now briefly describe each dense model and its utility within COVID-LEAP:

5.7.5 distilbert-base-uncased

A smaller version of the original BERT model, this one is trained using a knowledge distillation technique to reproduce the behavior of the larger "*teacher*" model with only half the layers, claiming to "*retain 97% of its language understanding capabilities and being 60% faster*" (Sanh et al., 2020). We are particularly interested in the evaluation of smaller models to verify several potential benefits. First, the total encoding time of all article passages should be lower than with larger dense models primarily due to fewer layers and parameters. Second, latency between real-time query encoding and similar document retrieval should be lower for a fully semantic search strategy.

As DistilBERT inherits the same BookCorpus (Zhu et al., 2015) and English Wikipedia training as BERT, it provides an opportunity to benchmark retrieval performance on specialised domain literature with a model trained in the general domain.

5.7.6 *sentence-transformers/msmarco-distilbert-base-v3*

With COVID-LEAP, we seek to retrieve relevant evidence to satisfy information needs such as "*Which vaccines are approved?*" and *"Which of the current vaccines in the clinic have reported the highest levels of neutralizing Abs after a single vaccination?"*. Proposing the quality of answers is likely to be higher within a paragraph rather than a sentence, our information retrieval problem can be considered one of asymmetric semantic search, with the question typically shorter than the text best suited to answering it.

Therefore, we now introduce the first of several models available that take BERT variants and further fine-tune on MS MARCO (Microsoft MAchine Reading COmprehension). This dataset of over 1M real-world questions sampled from Bing search engine logs and 8M web-page passages retrieved by Bing that are used to guide curation of human rewritten answers to the question set (Nguyen et al., 2016). Using models trained to encode and rank longer passages is assumed more likely to give better search results.

5.7.7 *sentence-transformers/ce-ms-marco-TinyBERT-L-4*

Another BERT variant tuned with MS MARCO, we include this model to evaluate performance impact with an architecture that has fewer hidden nodes, layers, and parameters. A reduced vector space representation and lower query inference latency while offering semantic search capabilities are potentially attractive characteristics for reducing GPU compute and large corpus embedding storage, especially for the smaller enterprise.

5.7.8 *castorini/ance-msmarco-passage*

Researchers at Microsoft introduce a BERT variant with Approximate nearest neighbor Negative Contrastive Estimation (ANCE) (Xiong et al., 2020). This method improves text representation by encoding the training query and document pairs together and using Approximate Nearest Neighbour (ANN) proximity selection to asynchronously build a global collection of negative training samples available to the local batch learning mechanism. Fine-tuned on MS MARCO, the authors claim "*the advantage of ANCE in web search, OpenQA, and the production system of a commercial search engine*".

5.7.9 *BeIR/query-gen-msmarco-t5-large-v1*

The largest of all pre-trained models we leverage in COVID-LEAP, the base "*T5*" model (**T**ext-**t**o-**T**ext **T**ransfer **T**ransformer) released by Google in 2019 (Roberts et al., 2019), uses both the encoder and decoder stacks to generate new output text from input text. Trained on cleansed, English-language only web scraping data from the Common Crawl repository and fine-tuned with MS MARCO, we use it to generate queries from COVID paper paragraphs for in-domain fine-tuning.

5.7.10 *cross-encoder/ms-marco-MiniLM-L-12-v2*

With the ability to attend to, or contextualise, query-article paragraph pairs simultaneously, we evaluate this cross-encoder transformer architecture fine-tuned on MS MARCO as a second-stage re-ranker to improve the ordering of first-stage retriever model search results. Due to this cross-attention feature, they are prohibitively slow, therefore we limit the candidate subset to a re-rank depth that defaults to the top 100 hits from the first stage.

5.8 Semantic Model Fune Tuning

One conclusion of our literature review was a generally held understanding that models trained in-domain typically deliver higher performance in NLP tasks. However, no large training dataset specialised in training models for COVID-19 information retrieval is available. To evaluate the potential for improving off-the-shelf models, we develop a new model dedicated to our downstream search task (Fig. 11-23).

First, a subset of 1K academic articles selected from the CORD-19 corpus using Bernoulli random sampling are split into paragraphs then cleansed with punctuation and URLs removed. Then, each of our article paragraphs is pushed through UKPLab's query generation function (UKPLab sentence-transformers, 2021) based on the sequence-to-sequence task of a T5 encoder-decoder transformer model fine-tuned on MS MARCO. The output is *n* synthetic queries per paragraph. Typically, the output queries are prefixed with interrogative phrases such as "*what is*", "*which of the*", and "*how to*", derived from the distribution of these phrases in MS MARCO. Three shorter generated queries from sampled input paragraphs are shown in Table 7:

Generated Query	Input Paragraph
which antiviral medication is effective for hospitalized patients with coronavirus disease	What are the effectiveness and harms of remdesivir in hospitalized patients with coronavirus disease 2019 COVID19
what is calcium phosphate used for	Calcium phosphate CaP in the form of microparticle or nanoparticle has been used as potential adjuvants or vaccine carrier delivery system for DNA and peptide vaccines for humans and mammals CaP nanoparticles have a number of advantages over other inorganic particles Their biodegradability and biocompatibility are excellent 103 they are native to the body welltolerated and absorbed in the body nontoxic costeffective and easily manufactured and have a high affinity to protein DNA antigens and chemotherapy drugs
which mutation increases affinity of sarscov	Three key mutations present in P1 N501Y K417T and E484K are located in the spike protein RBD The former two interact with human angiotensinconverting enzyme 2 hACE2 11 whilst E484K is located in a loop region outside the direct hACE2 interface fig S11 Notably the same three residues are mutated with the B1351 variant of concern and N501Y is also present in the B117 lineage The independent emergence of the same constellation of mutations in geographicallydistinct lineages suggests a process of convergent molecular adaptation Similar to what was observed for SARSCoV1 4244 mutations in the RBD may increase affinity of the virus for host ACE2 and consequently impact host cell entry and virus transmission Recent molecular analysis of B1351 45 suggests that the three P1 RBD mutations may similarly enhance hACE2 engagement providing a plausible hypothesis for an increase in transmissibility of the P1 lineage

Table 7: Examples of synthetic query generation from CORD-19 articles for dense model training

As the T5 model is large and consumes significant GPU memory resources, we limit *n* queries per paragraph to two. Further, we set the maximum input length of paragraphs to the mean paragraph length of the CORD-19 corpus (800 characters) and fix the generated output query length to 80 tokens. To optimise query quality, they are parsed with the removal of punctuation, numeric-only, and duplicate queries before upload to Azure blob storage as a comma-delimited file.

The second step uploads the generated queries and fits them to a base model such as *castorini/ance-msmarco-passage* to create a model calibrated to our literature corpus. During development, CUDA (Compute Unified Device Architecture, the API for GPU compute) memory errors related to resource constraints of affordable Azure GPU compute options were resolved by setting batch size to 16, slowing model training. Table 8 indicates

the time taken over one epoch to fine-tune off-the-shelf dense models to CORD-19 using generated synthetic queries.

Base Model	Fine-Tuning Time	~ Iterations Per Second
distilbert-base-uncased started	6m 57s	10
sentence-transformers/msmarco-distilbert-base-v3	6m 54s	10
sentence-transformers/ce-ms-marco-TinyBERT-L-4	3m 20s	25
castorini/ance-msmarco-passage	13m 27s	6

Table 8: Time over one epoch to fine-tune off-the-shelf dense models to CORD-19 using generated synthetic queries

5.9 Semantic Model Deployment

Upon completion of fine-tuning, the model is saved to the runtime working directory then registered within the Azure Machine Learning workspace (Fig. 11-24). Naming uses the prefix *c19gq* to represent a specialised COVID-19 model fine-tuned with synthetically generated queries while tags identify the originating base model and training dataset (Fig. 22). The models are now available for consumption by authenticated indexing and inference services.

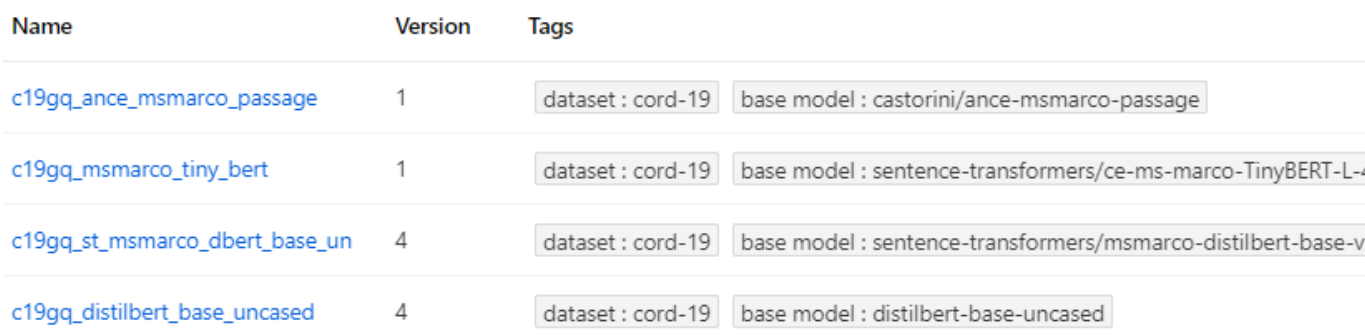

Fig. 22: Azure machine learning model registry of COVID-LEAP fine-tuned dense models

5.9.1 *Corpus Embedding and Indexing*

Our literature corpus has been extracted, cleansed, filtered for English-only texts, new features including coronavirus variant flags and paper quality metrics derived, and papers classified by topic. Dense models have been calibrated to COVID-19 IR. We can now proceed to upload article knowledge into ElasticSearch for query by our downstream IR task.

Getting our corpus into ElasticSearch is a two-step process. "*Model_embed_corpus*" (Fig. 11-25) downloads the dense model of choice, for example, *c19gq_ance_msmarco_passage*, then establishes a connection to the PostgreSQL database persisting our prepared corpus relational model. Title, text, and other metadata are retrieved in chunks of 5000 articles to accommodate memory constraints. For each article, the full body text is split into paragraphs then combined with the title before encoded as a single vector representation by our fine-tuned model. Unlike with topic modelling, no stopword removal or lemmatisation preprocessing of paragraph texts is done, as we wish to preserve meaning and context by retaining meaningful negations like "*not*", verbs like "*show*", and avoid reducing "*vaccinated*" to "*vaccinate*". The expanded chunks by paragraph with associated encoding are saved to the pipeline runtime working directory as serialised byte stream files.

The second step, "*Model_corpus_to_es*" (Fig. 11-26), compiles a list of paragraph encoded files generated by the first step then makes a connection to the cloud-hosted ElasticSearch instance. Each file is read in and deserialised. Transformation of columns to ElasticSearch format, for example, lowercasing column *is_coronavirus* containing Python representations of boolean values. Next, as we currently take an "*all or nothing*" approach to our search engine update, we delete the ES index *pub_text* then recreate it with a field specification that includes *title* as type *text*, *publish_year* as type *integer*, *text_processed* as type *text* for a UTF-8 character representation of the paragraph, and *text_processed_vector* as type *dense_vector* for the embedded title-paragraph combination.

By computing and pre-indexing all article title-paragraph embeddings in advance, only the query or any semantic reranking of a result subset needs to be calculated during search inference. Indexing the title and paragraph as text in the same ES document as the vector representation facilitates a BM25 lexical search across both text fields and a semantic search comparing encoded query with the vector form of title and paragraph. Updates to ES are done in bulk to increase indexing performance. Indexing all 139K corpus articles at paragraph level results in an index of 4.7M documents consuming 71GB of storage.

5.10 Search Strategy & Inference

Our objective is to satisfy the researcher's information needs by searching the indexed CORD-19 corpus with their query and returning the most relevant candidate paragraphs as direct answers. Our literature review uncovered two main approaches to biomedical IR. The first is a single-model strategy applying a semantic model as a dense retriever for similarity comparison between embedded query and all paragraphs in the vector space. The second is an ensemble architecture, typically a first-stage key term matching model for retrieving candidates with a top subset re-scored for similarity by a dense model to improve in-domain capturing of semantics.

There is a trade-off between these strategies. A full pairwise comparison of paragraph vector embeddings has the potential to locate better-quality candidates at the expense of high computation cost and latency. In contrast, strategies with a lexical first stage and dense re-ranker can benefit from sub-second retrieval and limited semantic comparison but are constrained by first stage result quality. With COVID-LEAP, our testing of search models includes a third way. We re-rank best matches from first stage dense retrieval with a dense cross-encoder model for second-stage semantic fine-grain scoring to evaluate if this most expensive strategy is justified.

An inference script (Fig. 11-27) is developed with input parameters for the question and any corpus pre-filtering constraints such as publication year and coronavirus variant. Output is an answer list of relevant documents containing relevancy scored paragraph snippets and associated indexed metadata, including journal and DOI. For the baseline lexical search, no additional encoding is required as ES is instructed to perform a multi-match keyword search on both title and paragraph text using the input question, returning a ranked list of documents according to BM25 scoring.

For any first-pass search with dense model strategies, only vector encoding of the question is necessary, passed as a *queryVector* argument to ES search for cosine similarity comparison with all embedded paragraph vectors. This similarity measurement indicates if the two documents are heading in the same direction within our dense vector space, and calculated (Fig. 23) by taking the dot product sum of the query to paragraph multiplication of each dimension value which is then divided by the product of the square root of the magnitudes (sum of each dimension multiplied by itself) of each vector.

$$\cos(\mathbf{q}, \mathbf{p}) = \frac{\mathbf{q}\mathbf{p}}{\|\mathbf{q}\|\|\mathbf{p}\|} = \frac{\sum_{i=1}^{n} \mathbf{q}_i \mathbf{p}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{q}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{p}_i)^2}}$$

Fig. 23: Calculation of cosine similarity of two text sequences, where *q* denotes query and *p* paragraph

No additional paragraph encoding is necessary when reranking BM25 results with our fine-tuned neural models. As the lexical search candidate hitlist includes the pre-indexed embedded vector representation of each paragraph, only a pairwise cosine similarity measurement between question and paragraph representations is necessary. This reranking strategy adds minimal computational cost to the lexical search (~0.05s).

Conversely, we limit the top *n* results reranked with the cross-encoder strategy by setting a hard rerank depth (default 100) given the overhead of this model's full attention mechanism re-encoding both the query and paragraph for all candidate documents. The output of all neural model strategies is a similarity measurement for each paragraph snippet between 0 and 1, easily weighted by the paper quality metrics we introduced previously. By default, results are ranked in descending order according to the relevance to the question being asked.

Making the inference script available as a highly scalable web service for consumption by our application front-end is achieved by deployment to an Azure Kubernetes Service (AKS) cluster, which provides the computational resources for our GPU-enabled neural models.

5.11 Application front-end

A fundamental principle guiding the application front-end design for COVID-LEAP (Fig. 11-28) is the ease of use for both the developer and biomedical end-user. The application segregates analysis of the vaccine development landscape and surfacing of relevant academic articles into clearly defined, selectable pages.

Four highly visual and interactive pages present overviews and detail of both clinical and preclinical trials (Figs. 24-27). For literature search (Fig. 28), a prominent yet straightforward search bar allows entry of a free-text question. Paragraphs attempting to answer the question directly are ordered by most to least relevant as default and tagged with authors, journal source, topic label, and publishing year. Trials referenced in papers are shown, linked the vaccine development knowledge to the academic papers. In such a highly dynamic research domain, researchers want to focus only on the most recent papers, so default selection for each of the last three years is a feature in addition to setting a custom publication year timeframe. Facilitating the refinement of results, filtering mechanisms search and list only documents associated with selected journals, categorised as specific to a virus variant, or classified as an explicit topic. Drop-down options for ordering search results by different criteria, including paper citation count, are available. Further drop-downs allow selecting explicit search and bibliometric scoring strategies, particularly helpful during solution development and evaluation.

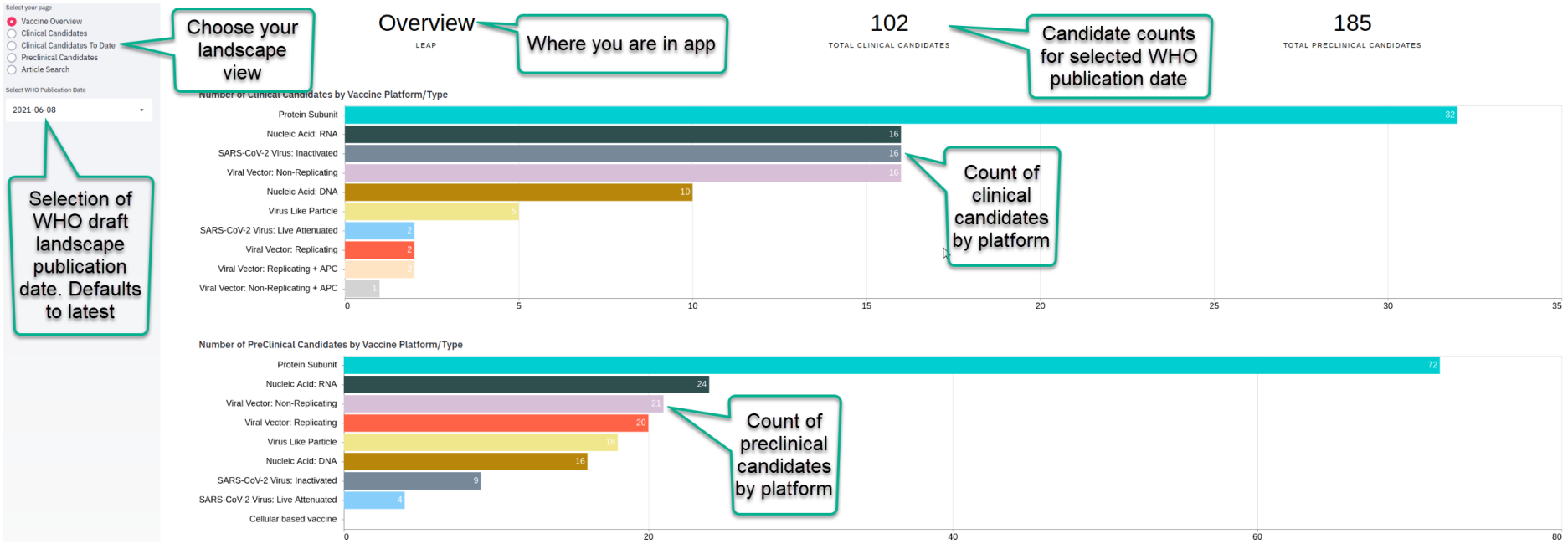

Fig. 24: COVID-LEAP vaccine development overview page. Provides a high-level view of clinical and preclinical vaccine candidates by date with KPIs.

Fig. 25: COVID-LEAP interactive clinical vaccine candidates detail view, explorable by latest development stage or cumulative "*over time*" view

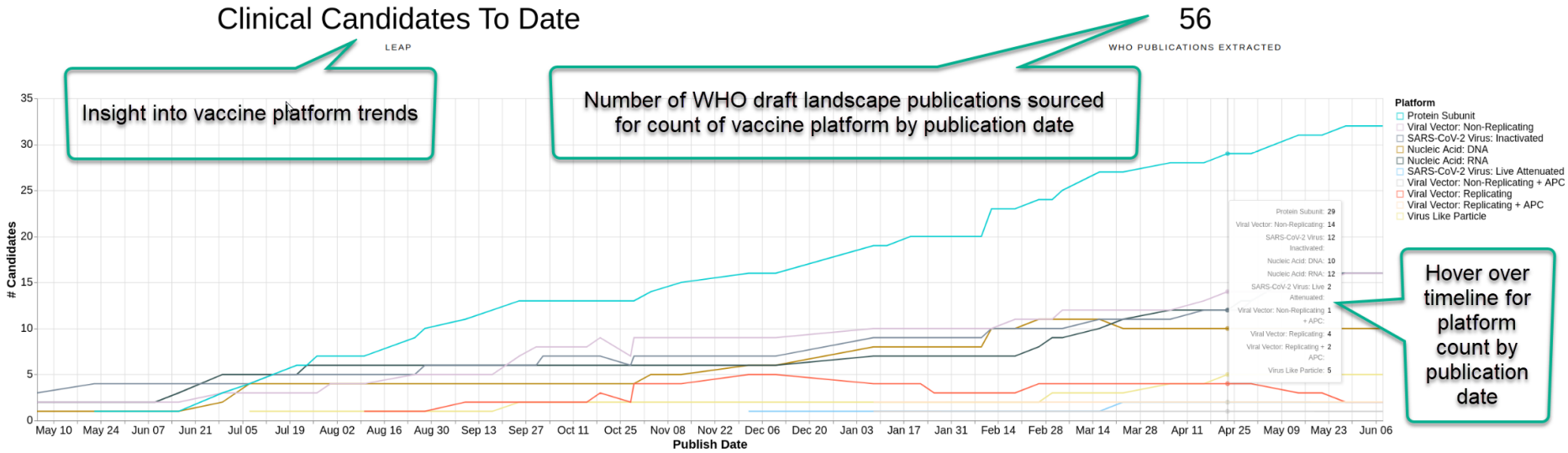

Fig. 26: COVID-LEAP clinical vaccine candidates to date view, by platform

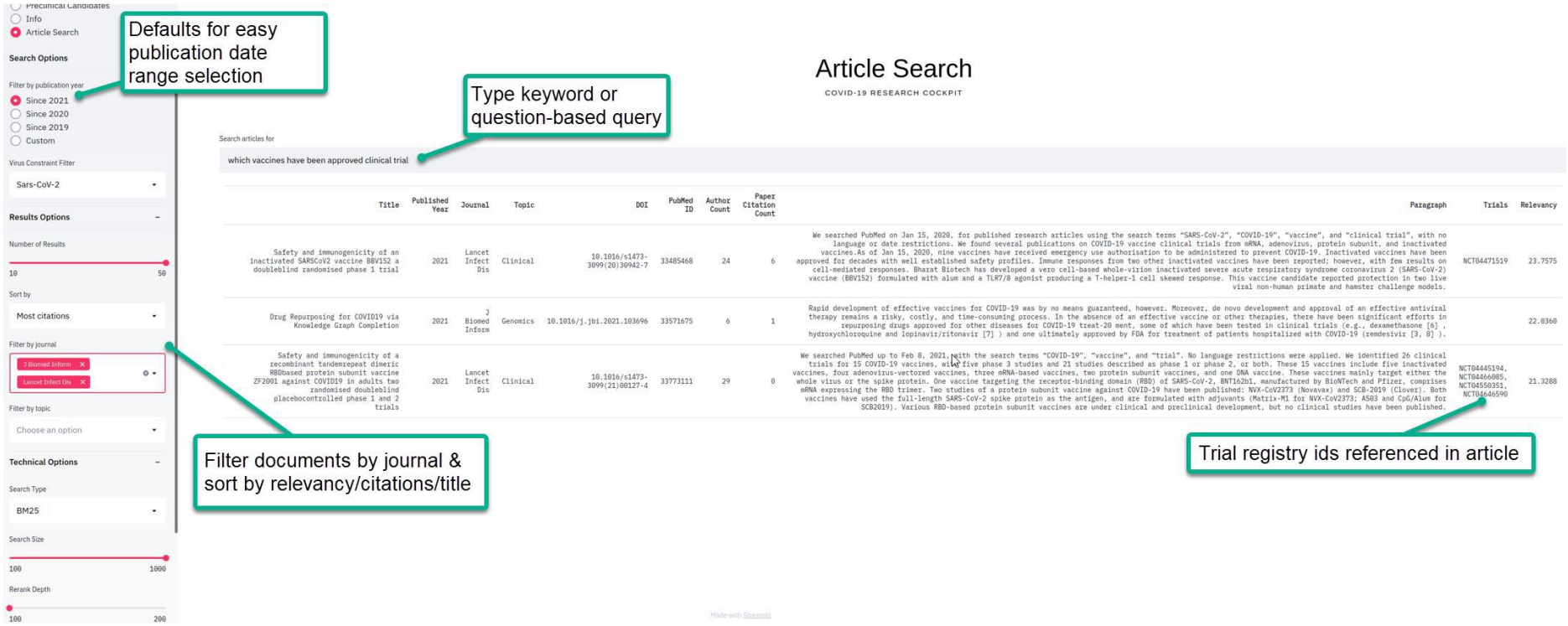

Fig. 27: COVID-LEAP academic literature search. Each result includes title, paragraph text as a direct answer to the question, and links to trials referenced in the paper. For ease of use, the left sidebar contains pre-configured filters for publication year and coronavirus type. Results can be sorted by relevancy, citations, or title. Results filterable by journal and topic. Technical options to set IR strategy (BM25, dense, two-stage rerank), search size, and rerank depth.

6. Experiment Design & Results

We have proposed IR strategies incorporating lexical (e.g., *BM25*), dense (e.g., *msmarco-distilbert-base-v3*), and re-ranking models (e.g., lexical first stage *BM25* reranked by second stage dense *c19gq-ance-msmarco-passage*), with and without applied paper quality metrics, for the location and ranking of relevant candidate paragraphs from a large corpus of academic articles. We now perform a set of experiments to empirically evaluate our approaches for their ability to surface the best evidence for answering biomedical questions.

Given the specialised nature of the domain, the complexity of the information need and questions asked, and the volume of potentially useful and irrelevant knowledge persisted in our ElasticSearch engine, "*best*" can be a subjective term. We seek to determine any measurable benefit dense models offer over traditional lexical search to justify the additional compute costs they incur in tuning and deployment. We want to understand how off-the-shelf and fine-tuned models perform IR tasks across general and specialised domains. Do models excel or fail when answering particular types of questions, or is their performance consistent? For a balanced assessment of models and strategies, we incorporate two types of evaluation into our experimental design: intrinsic and extrinsic.

We first perform an intrinsic evaluation to test the technical performance of each search approach, benchmarking results against a set of known query-answer pairs. Such tests can be automated, require no human expert feedback, and provide insight into model ability and support iterative hyperparameter optimisation at low cost. However, we consider extrinsic evaluation far more valuable and realistic. Human-defined test cases and result feedback based on real-world operational expertise are often more challenging, highly critical, and the best indicator of how close the solution is to solving the information problem.

We now present the experimental design and results for each evaluation type. All experiments were conducted on the local development system featuring an AMD 5950X CPU, Nvidia RTX3090FE GPU, and 64Gb RAM.

6.1 Intrinisic Evaluation

6.1.1 Experimental Design

Our intrinsic offline evaluation, isolated from the indexed CORD-19 literature, selects five benchmarking datasets designed to test the specialist and out-of-domain capabilities of each search strategy. The collected metrics aim to provide a first impression of performance before progressing to next-stage evaluation by human experts.

Our initial experiment with several dense models (e.g., off-the-shelf *distilbert-base-uncased*, fine-tuned *c19gq-ance-msmarco-passage*) benchmarks their ability to make the correct similarity prediction for each of the 489 labelled domain-specific question pairs in the CovidQA dataset (Tang et al., 2020), introduced in the literature review. For each pair, a human-annotated label indicates binary question similarity (1=similar, 0=dissimilar), with examples presented in Table 9. Each model encodes both questions as individual vector representations then passed as a pair for predicting cosine similarity. Predictions exceeding a probability threshold of 0.5 indicate similarity.

Question 1	**Question 2**	**Similarity**
What are the clinical features of COVID-19?	How does COVID-19 present?	1
Who is at risk for severe disease from COVID-19?	Who will COVID-19 affect most?	1
Are the symptoms of COVID-19 different in children than in adults?	Does warmer temperature stop the outbreak of COVID-19?	0
What are the clinical features of COVID-19?	Is the use of facemask recommended to prevent COVID-19?	0

Table 9: Examples of Covid-QA evaluation dataset labelled question pairs for similarity prediction

For the second round of intrinsic evaluation, we select four benchmarking datasets from UKPLab's BEIR (BEnchmarking IR) (Thakur et al., 2021) framework:

- **TREC-COVID** (Voorhees et al., 2020) is a benchmark based on a collection of 171K biomedical articles drawn from the CORD-19 dataset. A large-scale pool of human experts judges the most relevant documents answering a set of 50 COVID-specific questions. Articles surfaced by search strategies are graded as relevant, partially relevant, or not relevant to the question and scored accordingly. This benchmark is selected to evaluate COVID domain performance.

- **SciFact** is designed to verify COVID-19 claims. Experts have compiled a dataset of 1.4K claims complete with "*evidence containing abstracts annotated with veracity labels and rationales*" (allenai/scifact, 2021). Again, this benchmark is selected to evaluate COVID domain performance.

- **NFCorpus** (NFCorpus, 2021) is a benchmark is designed to evaluate medical IR, containing 3K questions with 170K automatically extracted relevance judgments for 9K medical documents. It is included to test model performance in capturing representation from a non-COVID scientific domain.

- **FiQA-2018** (FiQA, 2018) is a question-answering dataset of 57K opinion-based answers sourced from news channels and blog posts to 648 finance-specific questions. For cross-domain IR performance evaluation of models with distinctly different semantics.

6.1.2 Results Analysis

Performance for our experimental collection of IR models when benchmarked with Covid-QA (table 10) and BEIR 4 set suite (Table 11) are presented below. While the F1 metric measures the accuracy of dense models on the Covid-QA test, we present BEIR results using nDCG@10 (normalised Discounted Cumulative Gain, Wang et al., 2013). This rank-aware metric is commonly used to evaluate document preference in retriever systems, including web search engines. The *@10* denotes the top 10 candidate documents are evaluated.

Model (↓)	**F1 Score**
Base Dense	
distilbert-base-uncased	0.667
msmarco-distilbert-base-v3	0.819
Fine-Tuned Dense	
c19gq-distilbert-base-uncased	<u>0.862</u>
c19gq-st-msmarco-dbert-base-uncased	0.852
c19gq-ance-msmarco-passage	**0.911**
c19gq-msmarco-tiny-bert	0.822

Table 10: IR model Covid-QA benchmark F1 performance. Higher is better. Best score in **bold**, 2nd best <u>underlined</u>.

NLP Task (→)	**IR**		**Q&A**	**Fact-Checking**	**Avg**
Domain (→)	**Biomedical**		**Finance**	**Scientific**	
Benchmark Dataset (→)	**TREC-COVID**	**NFCorpus**	**FiQA-2018**	**SciFact**	
Model (↓)					
Lexical					
BM25	0.616	**0.305**	0.239	<u>0.644</u>	<u>0.451</u>
Base Dense					
distilbert-base-uncased	0.149	0.047	0.03	0.096	0.080
msmarco-distilbert-base-v3	0.477	0.256	0.257	0.538	0.382
Fine-Tuned Dense					
c19gq-distilbert-base-uncased	0.481	0.231	0.194	0.522	0.357
c19gq-st-msmarco-dbert-base-uncased	0.451	0.247	0.229	0.546	0.368
c19gq-ance-msmarco-passage	0.58	0.209	0.261	0.514	0.391
c19gq-msmarco-tiny-bert	0.328	0.126	0.085	0.263	0.200
Lexical + Fine-Tuned Dense Re-ranker					
BM25 + c19gq-distilbert-base-uncased	0.653	0.23	0.2	0.523	0.401
BM25 + c19gq-st-msmarco-dbert-base-uncased	0.636	0.246	0.232	0.547	0.415
BM25 + c19gq-ance-msmarco-passage	<u>0.664</u>	0.209	<u>0.262</u>	0.515	0.412
BM25 + c19gq-msmarco-tiny-bert	0.506	0.127	0.089	0.264	0.246
Fine-Tuned Dense + Cross-Encoder Re-ranker					
c19gq-ance-msmarco-passage + cross-encoder/ms-marco-MiniLM-L-12-v2	**0.753**	<u>0.287</u>	**0.352**	**0.671**	**0.516**

Table 11: IR model BEIR benchmark set performance measured with nDCG@10, higher is better. Best score in **bold**, 2nd best <u>underlined</u>.

In COVID-QA benchmarking, we first observe how the base distilbert model benefits from fine-tuning on MS MARCO, up +0.152 on the F1 score. On BEIR TREC-COVID benchmarking, tuning realises a +0.328 gain to 0.477 from the base model's 0.149. Average BEIR benchmarking performance is up ~fivefold, jumping from 0.080 to 0.382. We propose the MS MARCO tuning helps the model deal with the asymmetric nature of academic Q&A, with longer passages required to address the information needs of shorter questions. Moreover, it is clear how base models benefit from unsupervised fine-tuning on synthetic queries generated from sampled CORD-19 literature. All tuned models outperform their base counterparts with better capture of domain semantics when tasked with in-domain IR.

With only four layers and a vector representation less than half the dimensions of other models, the tiny BERT variant is the poorest performing tuned model. However, it still outperforms baseline versions, another verification of the benefits of in-domain tuning. This demonstrates the potential utility of smaller semantic models if the latency response of real-time full similarity comparison with larger, higher-performing dense models is prohibitively expensive.

By quite some margin, the strongest model is the ANCE variant. Albeit a complex and expensive model, it is an early indication of the compound effectiveness of several layers of vanilla BERT refinement to tune for asymmetric passage retrieval, cross-encoding training with query-document pairs for better representation, an improved learning mechanism using globally sourced negative samples, and a final calibration with COVID literature.

Reviewing results from the BEIR benchmark suite, the baseline BM25 lexical model is the strongest on the NFCorpus task, in contrast to the poor performance of all dense models. BM25 generalises well, ranking 2nd in the SciFact task and outperforming all single-stage dense strategies in TREC-COVID. NFCorpus results of re-ranking strategies based on a BM25 first stage are not improved by any second stage dense re-score. However, an interesting observation is how the dense models can capitalise on BM25 in TREC-COVID and fine-grain the lexical model results with semantic capture for even better performance. This demonstrates how the lexical baseline can be improved with minimal additional computational expense.

Cross-encoding already good results from the ANCE model re-orders document relevance for the best performance in TREC-COVID. In all tasks but especially SciFact, considerable improvements are realised from tuning the baseline *distilbert-base-uncased* with synthetic queries. Surprisingly, the opposite is true of *msmarco-distilbert-base-v3*, where all four tasks show a minor performance degradation with the fine-tuned variant. We

assume fine-tuning with a small synthetic query set for only three epochs is introducing instability. Reflecting the Covid-QA results, the tiny BERT model performs poorly across the board, significantly impacted by its reduced layer and vector dimensions counts.

The dense models calibrated to COVID perform poorly out of domain. Only the ANCE model with cross-encoder re-ranking generalises well. It shows it can cope with domain shift as the best performer in the FiQA-2018 finance dataset despite its COVID fine-tuning heritage, and 2nd second best in the non-COVID medical NFCorpus task behind BM25. It achieves the best results for the in-domain TREC-COVID and SciFact tasks. However, such performance is achieved only with the highest computational cost. Based solely on intrinsic evaluation results, if a one-minute response is acceptable, we recommend this strategy. Otherwise, BM25 re-ranked by ANCE for competitive results that include a semantic search capability with low latency.

6.2 Extrinsic Evaluation

6.2.1 Experimental Design

Ultimately the relevance quality of documents surfaced to medical researchers must sufficiently contribute to their information need for a solution to be trusted, adopted, and considered a success. We now introduce a second quality gate to measure the performance of our search strategies. A medical expert with 30 years of experience in biotech, pharma, and academia devised a fixed test set of 5 questions (Table 12). Observe the diversity and varying difficulty of the question types, from the short *Q5*, the open-ended, multiple entity, and temporal nature of *Q4* (vaccines, countries, when), to the longer, more precise, and elaborate format of *Q2* and *Q3*.

Number	Question
1	Which of the current vaccines in the clinic have reported the highest levels of neutralizing Abs after a single vaccination?
2	What T-cell epitopes have been identified in the Receptor Binding Motif (RBM) region of the S-glycoprotein Receptor Binding Domain (RBD) of the SARS-CoV-2 virus?
3	What mutations have been identified in the Receptor Binding Motif (RBM) region of the S-glycoprotein Receptor Binding Domain (RBD) of the SARS-CoV-2 virus?
4	Which vaccine(s) are approved? In which countries? When did they receive approval?
5	How effective are these vaccines?

Table 12: Question test set devised by a medical expert for extrinsic evaluation of COVID-LEAP

All questions for each search strategy are pushed through the web application for testing with the same service endpoint conditions as actual end-users of COVID-LEAP. The top 3 results for each question, including article title, paragraph, PMC ID, D.O.I., and link to the full article, are collected and passed to the medical expert for a relevance assessment. Each question result, or "*hit*", is marked by the expert using the TREC (Text Retrieval Conference) grade labelling system presented in Table 13. We add a points system to empirically measure the strategy results. For example, the best hits are awarded 3 points, while irrelevant hits are awarded -1 points.

To add rank awareness, the points awarded to the top result are multiplied by three and the 2nd result by two. A perfectly relevant document ranked 1st is awarded 9 GRA (Grade Rank Aware) points, while an irrelevant document ranked 1st is awarded -3 GRA points. This scheme recognises strategies able to place the best document at the top of their hit list and penalises highly ranked irrelevant documents. The test was blind, with the expert unaware of the explicit strategy providing answer sets. As the question set is focused on COVID-19, we subset the corpus being searched by including only those indexed documents tagged with coronavirus variant "*is_sars_cov_2*" as true.

Grading Label	**Description**	**Points**
Perfectly relevant	The document content is dedicated to the query and worthy of a top result	3
Highly relevant	The document content provides substantial information on the query	2
Relevant	The document provides some information relevant to the query, which may be minimal	1
Irrelevant	The document does not provide any useful information about the query	-1

Table 13: Extrinsic search result grading label and points

For brevity, each search strategy is assigned an identifier. Given the high intrinsic evaluation performance of ANCE when used both as a single-stage retriever and re-ranked by a cross-encoder, we select this model as our base dense model of choice in methods *B* to *I*. Method *C* limits the re-rank depth to only the top 3 BM25 results (RD3), whereas method *D* permits ANCE broader access to the top 100 BM25 results. Methods *F* to *I* weights retriever relevancy by citation count (MF1), author citation ratio (MF2), combined author and paper PageRank (MF3), and PageRank with recency (MF4). With PMC the "*go-to*" for biomedical literature search, we include method *J*, which applies the same question set to a PMC advanced search to constrain the search to title and body content containing the same Sars-CoV-2 keywords used in engineering our "*is_sars_cov_2*" document tag.

6.3 Results Analysis

Table 14 presents the expert graded evaluation of 10 search strategy results. Columns marked *G* indicate the grading label point awarded. Columns marked *GRA* adjust the label point to add rank awareness. *Total GRA* (TGRA) sums GRA by question to reflect each strategy's overall retriever ranking ability to generalise across the 5 question set. Winning strategies by question are marked **bold** with 2nd place underlined. Average response times are given for each strategy.

While BM25 showed good results across multiple synthetic benchmarking tasks, it ranks 6th here. Reviewing TGRA suggests BM25 offers generally stable but unremarkable performance except for Q5 with TGRA comparable with several other strategies surfacing the same document. Consistent with intrinsic results, strategies using ANCE deliver the best performance. Only one negative TGRA for Q3 and winning results in Q2 and Q5 push method B into 1st place alongside method E, also using ANCE but with an additional cross-encoding stage. Method *E* outperforms *B* in Q4 but penalised by poor Q2 performance. However, this method scores the only "*perfectly relevant*" of the test.

In contrast to the intrinsic evaluation, where dense re-ranking strategies benefitted from solid BM25 results, we now see the effectiveness of dense models severely impacted by BM25 limitations. Method *D* increasing the re-rank depth helped marginally. This suggests the necessity of dense retrievers to capture and match the intended meaning of complex and sometimes ambiguous real-world questions. Of the four strategies applying paper quality metrics, only method *I* combining PageRank for author and paper influence with recency for bias towards newer papers does well, ranking 2nd overall. All other metrics show unsatisfactory performance, appearing overly aggressive and noisy, weighting good ANCE results (methods *B*, *E*) to detrimental effect. We now share some additional thoughts on each of the questions:

- **Q1** stressing "*which of the current vaccines in the clinic have reported the highest…*", seeking recent vaccine comparison, proved very challenging. Terms similar to "*antibody*" and "*neutralising*" raise document relevance without attending to the plural interrogative for "*which vaccines*", with suggested papers only discussing one vaccine or not at all. This question is also difficult to resolve with models unable to infer "*highest*" from multiple reports of clinical numerical statistics.

- Similarly, for **Q2** "*what T-cell epitopes have been identified in the Receptor Binding Motif (RBM) region of the S-glycoprotein…*", models commonly identify the presence of "*receptor*" and "*S-glycoprotein*" but either miss the interrogative on entity "*epitopes*", or discuss epitopes but not the T-cell types. We suggest this last oversight due to the WordPiece tokenizer splitting "*T-cell*" into two separate terms.

- **Q3** "*what mutations have been identified in the Receptor Binding Motif…*", without any hyphenation of entity broadly appeared the most straightforward question for the dense models to extract useful information, with cross-encoding significantly improving ANCE results.

- **Q4** "*which vaccine(s) are approved? In which countries? When did they receive approval?*" seeks combined knowledge on multiple entity types. The information need infers a preference for the most recent papers. Here, methods E and I based on ANCE cross-encoding suggest highly relevant papers in their top 2 hits, but in a different order, with *I* and its quality metric preferring the article having double the accesses of *E* on the springer article archive.

- **Q5** poses the shortest question in the set, "*how effective are these vaccines?*". Most strategies retrieve the same document titled "*How effective are the Covid-19 vaccines – A Bayesian analysis*". With both the title and paragraph embedded in each vector, it demonstrates the ability of dense models to represent and extract word order and context . Unfortunately, the question infers a need for the latest knowledge on vaccine efficacy. Therefore, despite being published in late 2020, it is considered an old article relative to the pandemic dynamics and graded "*relevant*".

Finally, PMC results are assessed 3rd best overall, impressive given its 3s response time. However, an important distinction between COVID-LEAP and PMC is that our solution responds with paragraph-level snippets while PMC query matches return only the title and other paper-level metadata, with no intra-level content as direct evidence. A search could not determine the explicit retrieval strategy supporting the PMC search engine, but we assume some highly optimised variant of BM25

Strategy (→)		**BM25**		**ANCE**	**BM25 + ANCE RD3**		**BM25 + ANCE RD100**		**ANCE + CE**		**ANCE + CE + MF1**		**ANCE + CE + MF2**		**ANCE + CE + MF3**		**ANCE + CE + MF4**		**PMC**	
Identifier (→)		**A**		**B**	**C**		**D**		**E**		**F**		**G**		**H**		**I**		**J**	
Type (→)		**Lexical**		**Dense**	**Lexical + Dense Rerank**		**Lexical + Dense Rerank**		**Dense + Dense Rerank**		**Dense + Dense Rerank + Citation Count**		**Dense + Dense Rerank + Author Citation Ratio**		**Dense + Dense Rerank + PageRank**		**Dense + Dense Rerank + PageRank + Recency**		**Unknown**	
Avg Search Time (s) (→)		**4.5s**		**52s**	**5s**		**5s**		**63s**		**63s**		**63s**		**63s**		**63s**		**3s**	
Grade / Rank Aware (→)		**G**	**GRA**	**G**	**GRA**	**G**	**GRA**	**G**	**GRA**	**G**	**GRA**	**G**	**GRA**	**G**	**GRA**	**G**	**GRA**	**G**	**GRA**	**G**	**GRA**
Question Hit (↓)																				
Q1	1	1	3	-1	-3	-1	-3	-1	-3	-1	-3	1	3	-1	-3	1	3	1	3	-1	-3
	2	-1	-2	1	2	-1	-2	1	2	3	6	-1	-2	1	2	-1	-2	-1	-2	1	2
	3	-1	-1	2	2	1	1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	1	1	-1	-1
			0		1		-4		-2		**2**		**2**		-2		0		**2**		-2
Q2	1	-1	-3	1	3	1	3	-1	-3	-1	-3	-1	-3	-1	-3	-1	-3	-1	-3	1	3
	2	1	2	1	2	-1	-2	2	4	-1	-2	-1	-2	-1	-2	-1	-2	-1	-2	1	2
	3	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1
			0		**6**		2		0		-6		-6		-6		-6		-6		**6**
Q3	1	-1	-3	-1	-3	1	3	2	6	2	6	2	6	2	6	2	6	2	6	2	6
	2	1	2	1	2	-1	-2	-1	-2	1	2	1	2	2	4	2	4	2	4	1	2
	3	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	2	2	2	2	2	2	1	1
			-2		-2		0		5		9		9		**12**		**12**		**12**		9
Q4	1	1	3	1	3	-1	-3	-1	-3	2	6	-1	-3	-1	-3	-1	-3	2	6	1	3
	2	-1	-2	3	6	1	2	-1	-2	2	4	1	2	1	2	1	2	2	4	1	2
	3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	2	2	-1	-1	1	1
			0		8		-2		-6		**9**		0		0		1		**9**		6
Q5	1	1	3	1	3	1	3	1	3	1	3	1	3	-1	-3	1	3	1	3	-1	-3
	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	-1	-2	-1	-2	-1	-2
	3	1	1	2	2	1	1	1	1	1	1	1	1	1	1	-1	-1	1	1	1	1
			6		**7**		6		6		6		6		0		0		2		-4
Total GRA			(6^{th}) 4		(1^{st}) **20**		(8^{th}) 2		(7^{th}) 3		(1^{st}) **20**		(4^{th}) 11		(6^{th}) 4		(5^{th}) 7		(2^{nd}) 19		(3^{rd}) 15

Table 14: Expert graded evaluation of results from 9 experimental IR strategies in COVID-LEAP used to respond to expert-defined question set. G represents grade point awarded per response hit. GRA shows G adjusted for rank awareness. PMC results to same question set also graded. Note PMC only returns title metadata and no paragraph-level evidence

6.4 Retrospective

We have empirically shown that search strategies tasks with complex IR benefit from model passage-level pre-training and domain calibration. Overall, the dense ANCE model with a cross-encoder reranking strategy consistently offers the best retriever performance in-domain and cross-domain. This suggests that ANCE is a good retriever candidate in scenarios where a shared search engine repository may index literature from multiple domains, such as biomedical, neuroscience specialisations, finance, and patents. However, such performance is expensive, with a full semantic comparison between embedded question and 4.7M paragraph vectors followed by re-ranking top results costing one minute. While BM25 proved hard to beat in intrinsic tests, the human evaluation was its undoing, with accuracy assessed poor.

It is unfortunate that the strategies re-ranking BM25 results with dense models evaluate so poorly in extrinsic tests given their very low additional overhead. These approaches require no full semantic search or additional vector encoding and therefore are easier to scale as document repositories grow. From a practical perspective, we would like to explore our winning ANCE strategy further to determine if its performance indicates readiness for production deployment and its inherent latency trade-off is acceptable. We have yet to appraise the potential latency improvements through vertical scaling GPU and ElasticSearch resources. With the current configuration, the paper quality metrics detrimentally impact models, and further refinement is necessary.

7. Conclusion

This Master's project presented COVID-LEAP for highly visual, interactive insights into the latest view of the COVID-19 vaccine knowledge landscape and efficient search of the CORD-19 corpus. We believe this project makes a contribution by integrating these typically disparate silos of knowledge.

The proposed solution experimented with and benchmarked nine IR strategies, including lexical models (BM25) and dense models based on BERT transformer architecture. Fine-tuning of models to the CORD-19 corpus allowed experimentation with methods to optimise for in-domain IR. This proved fruitful, as our results show, with base models responding positively to domain calibration. With the ANCE variant, we demonstrated the ability of state-of-the-art dense architectures refined in several iterations with MS MARCO passage training, CORD-19 tuning, and an enhanced learning mechanism to deliver a consistently winning performance. Empirical evaluation showed dense ANCE outperforming the lexical BM25 method and "*go-to*" PubMedCentral search site in most tasks. There is a consensus between automated synthetic benchmarking and assessment by our medical expert that ANCE was the strongest model.

Manual evaluation is highly time-consuming, so we are incredibly grateful to our medical expert for their time, diligence, and feedback in grading such a large set of IR strategy results. As each set of results came back, it only encouraged us to strive for better performance yet grounding us with realistic expectations. While our models, particularly those based on ANCE, retrieved relevant documents that went some way towards addressing the information need, it was a clear reminder that biomedical IR with NLP is not yet solved. The fundamental problem is that complex questions cannot be answered with knowledge from one document. Expert biomedical researchers draw on many years of experience in the domain to compile answers from many sources.

Through the project's experimentation with transformer models, we learned how models can be fine-tuned and deployed, and better understand how they behave in specialised and cross-domain IR tasks. A key learning was how much can be achieved through simple heuristics, including linking papers to trials and extracting useful technical drug names from unstructured trial text.

7.1 Limitations

We experimented with bibliometrics; however, the extrinsic grading of search results from models weighted by our paper quality metrics detrimentally affected relevancy. More work is required to dampen their overly aggressive influence. Topic modelling to classify documents into themes of field study needs re-evaluation. In discussing the label assignments, the opinion of our medical expert is that associating a single label to a document is an inherently weak and rudimentary approach leading to distrust in the label.

There is certainly room for improvement with the disambiguation of authors. Our current strategy uses a combination of author name and paper journal to constrain attribution of papers to author. We acknowledge drawbacks with this approach, and there is interest in further research to refine citation assembly.

7.2 Future Work

As discussed in the extrinsic evaluation results, models struggled to address expert question 2, beginning "*What T-cell epitopes…*", explicitly, to locate articles discussing T-cells. Having recognised the technical problem, we would first like to understand how common such named entity terms are in the biomedical literature and then explore custom tokenizers to prevent the splitting of hyphenated entities.

Further experimentation with transformers would build upon a solid foundation of evaluation results gained from this project. First, assessing the capabilities of Big Bird by Google, a recent transformer designed to capture representations of longer text sequences. Fine-tuning with larger synthetic query sets would look to verify possible performance improvements. With our hope that COVID-LEAP as an application can prove useful beyond the COVID-19 domain, we would like to evaluate its IR capabilities in other literature domains, including neuroscience. In closing, we share our medical expert's view of the COVID-LEAP dashboard:

"*The COVID-19 vaccine dashboard provides an overview of the global COVID-19 vaccine development landscape. It presents an impressive amount of vaccine development and related knowledge via an intuitive user-interface. Automatically generated visualizations of the preclinical and clinical development pipelines empower end-users to quickly and efficiently navigate and understand the landscape and how it is changing over time. Linked tabulations of knowledge allow users to quickly drill-down to more deeply understand the development and related details of any given vaccine or class of vaccines. The Dashboard is an invaluable resource for anyone working in the field of COVID-19 vaccine development.*"

BIBLIOGRAPHY

A, R.J., Beigel, J.H., Paolino, K.M., Jocelyn, V., Castellano, A.R., Zonghui, H., Muñoz, P., …2017. A Recombinant Vesicular Stomatitis Virus Ebola Vaccine. N. Engl. J. Med. 376, 330–341.

Colavizza, G., Costas, R., Traag, VA., van Eck, N., van Leeuwen, T., Waltman, L., 2021. A scientometric overview of CORD-19.

Ahamed, S., Samad, M., 2020. Information Mining for COVID-19 Research From a Large Volume of Scientific Literature. ArXiv200402085 Cs Q-Bio.

Wadden, D., Shanchuan, L., Lo, K., Wang, L., van Zuylen, M., Cohan, A., Hajishirzi, H., 2020. Verifying scientific claims.

Allot, A., Chen, Q., Kim, S., Vera Alvarez, R., Comeau, D.C., Wilbur, W.J., Lu, Z., 2019. LitSense: making sense of biomedical literature at sentence level. Nucleic Acids Res. 47, W594–W599.

Lever, J., Altman, R., 2020. Analyzing the vast coronavirus literature with CoronaCentral | bioRxiv

Artaud, C., Kara, L., Launay, O., 2019. Vaccine Development: From Preclinical Studies to Phase 1/2 Clinical Trials. Methods Mol. Biol. Clifton NJ 2013, 165–176.

Basu, S., Chakraborty, S., Hassan, A., Siddique, S., Anand, A., 2020. ERLKG: Entity Representation Learning and Knowledge Graph based association analysis of COVID-19 through mining of unstructured biomedical corpora, in: Proceedings of the First Workshop on Scholarly Document Processing. Presented at the EMNLP-sdp 2020, Association for Computational Linguistics, Online, pp. 127–137.

Beltagy, I., Lo, K., Cohan, A., 2019. SciBERT: A Pretrained Language Model for Scientific Text. ArXiv190310676 Cs.

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., 2003. A Neural Probabilistic Language Model

Brainard, J., 2020. Scientists are drowning in COVID-19 papers. Can new tools keep them afloat?

Bras, P.L., Gharavi, A., Robb, D.A., Vidal, A.F., Padilla, S., Chantler, M.J., 2020. Visualising COVID-19 Research. ArXiv200506380 Cs.

CDC, 2020. COVID-19 and Your Health. Cent. Dis. Control Prev.

Chen, Q., Allot, A., Lu, Z., 2021. LitCovid: an open database of COVID-19 literature. Nucleic Acids Res. 49, D1534–D1540.

Chen, Q., Peng, Y., Lu, Z., 2019. BioSentVec: creating sentence embeddings for biomedical texts. 2019 IEEE Int. Conf. Healthc. Inform. ICHI 1–5.

Chorba, T., 2020. The Concept of the Crown and Its Potential Role in the Downfall of Coronavirus. Emerg. Infect. Dis. 26, 2302–2305.

CORD-19: COVID-19 Open Research Dataset — Allen Institute for AI

UN News, 2021. COVID's led to 'massive' income and productivity losses, UN labour estimates show.

Cyranoski, D., 2020. Profile of a killer: the complex biology powering the coronavirus pandemic. Nature 581, 22–26.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv181004805 Cs.

Dong, M., Cao, X., Liang, M., Li, L., Liu, G., Liang, H., 2020. Understand research hotspots surrounding COVID-19 and other coronavirus infections using topic modeling. Infectious Diseases.

WHO, 2021. Draft landscape of COVID-19 candidate vaccines.

Ebadi, A., Xi, P., Tremblay, S., Spencer, B., Pall, R., Wong, A., 2021. Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing. Scientometrics 126, 725–739.

Elasticsearch: The Official Distributed Search & Analytics Engine.

Esteva, A., Kale, A., Paulus, R., Hashimoto, K., Yin, W., Radev, D., Socher, R., 2020. CO-Search: COVID-19 Information Retrieval with Semantic Search, Question Answering, and Abstractive Summarization. ArXiv200609595 Cs.

Farokhnejad, M., Pranesh, R.R., Vargas-Solar, G., Mehr, D.A., 2020. S_Covid: An Engine to Explore COVID-19 Scientific Literature

FiQA - 2018

Guo, M., Yang, Y., Cer, D., Shen, Q., Constant, N., 2020. MultiReQA: A Cross-Domain Evaluation for Retrieval Question Answering Models. ArXiv200502507 Cs.

Gupta, A., Aeron, S., Agrawal, A., Gupta, H., 2020. Trends in COVID-19 Publications: Streamlining Research Using NLP and LDA.

He, S., Bakhtiari, Z., 2020. Developing Answers to Scientific Questions with BERT

WHO, 2020. Impact of COVID-19 on people's livelihoods, their health and our food systems

Mikolov, T., Karafiat, M., Burget, L., Cernocky, J., Khudanpur, S., Recurrent Neural Network Based Language Model, 2010.

Sundermeyer, M., Schluter, R., Ney, H., 2012. LSTM Neural Networks for Language Modeling

Johns Hopkins University, 2020. COVID-19 Map

Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G., 2016. MIMIC-III, a freely accessible critical care database. Sci. Data 3, 160035.

Kim, Y.C., Dema, B., Reyes-Sandoval, A., 2020. COVID-19 vaccines: breaking record times to first-in-human trials. Npj Vaccines 5, 1–3.

Lanying Du, Yuxian He, Yusen Zhou, Shuwen Liu, Bo-Jian Zheng, Shibo Jiang, 2009. The spike protein of SARS-CoV — a target for vaccine and therapeutic development. Nat. Rev. Microbiol. 7, 226–236.

Lee, J., Yi, S.S., Jeong, M., Sung, M., Yoon, W., Choi, Y., Ko, M., Kang, J., 2020. Answering Questions on COVID-19 in Real-Time. ArXiv200615830 Cs.

Lee, J., Yoon, W., Kim, Sungdong, Kim, D., Kim, Sunkyu, So, C.H., Kang, J., 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics btz682.

WHO, 2021. Listings of WHO's response to COVID-19.

MacAvaney, S., Cohan, A., Goharian, N., 2020. SLEDGE-Z: A Zero-Shot Baseline for COVID-19 Literature Search, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Presented at the EMNLP 2020, Association for Computational Linguistics, Online, pp. 4171–4179.

University of Heidelberg, 2021. NFCorpus: A Full-Text Learning to Rank Dataset for Medical Information Retrieval.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L., 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset.

Nguyen, V., Rybinski, M., Karimi, S., Xing, Z., 2020. Pandemic Literature Search: Finding Information on COVID-19

Oniani, D., Wang, Y., 2020. A Qualitative Evaluation of Language Models on Automatic Question-Answering for COVID-19. ArXiv200610964 Cs.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidle, S., 2015, BookCorpus Dataset

Roberts, A., Raffel, C., Lee, K., Matena, M., Shazeer, N., Liu, P.J., Narang, S., Li, W., Zhou, Y., 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Robertson, S., Zaragoza, H., 2009. The Probabilistic Relevance Framework: BM25 and Beyond.

Rosa, G.M., Rodrigues, R.C., Lotufo, R., Nogueira, R., 2021. Yes, BM25 is a Strong Baseline for Legal Case Retrieval.

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv191001108 Cs.

Sastre, J., Vahid, A.H., McDonagh, C., Walsh, P., 2020. A Text Mining Approach to Discovering COVID-19 Relevant Factors.

Sohrab, M.G., Duong, K., Miwa, M., Topić, G., Masami, I., Hiroya, T., 2020. BENNERD: A Neural Named Entity Linking System for COVID-19

Soni, S., Roberts, K., 2020. An Evaluation of Two Commercial Deep Learning-Based Information Retrieval Systems for COVID-19 Literature. ArXiv200703106 Cs.

Su, D., Xu, Y., Yu, T., Siddique, F.B., Barezi, E.J., Fung, P., 2020. CAiRE-COVID: A Question Answering and Query-focused Multi-Document Summarization System for COVID-19 Scholarly Information Management. ArXiv200503975 Cs.

Tang, R., Nogueira, R., Zhang, E., Gupta, N., Cam, P., Cho, K., Lin, J., 2020. Rapidly Bootstrapping a Question Answering Dataset for COVID-19. ArXiv200411339 Cs.

Tasneem, A., Aberle, L., Ananth, H., Chakraborty, S., Chiswell, K., McCourt, B.J., Pietrobon, R., 2012. The Database for Aggregate Analysis of ClinicalTrials.gov (AACT) and Subsequent Regrouping by Clinical Specialty.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I., 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. ArXiv210408663 Cs.

Brin. S., Page. L., 1998. The anatomy of a large-scale hypertextual Web search engine. Comput. Netw. ISDN Syst. 30, 107–117.

Trewartha, A., Dagdelen, J., Huo, H., Cruse, K., Wang, Z., He, T., Subramanian, A., Fei, Y., Justus, B., Persson, K., Ceder, G., 2020. COVIDScholar: An automated COVID-19 research aggregation and analysis platform. ArXiv201203891 Cs.

Tyagin, I., Kulshrestha, A., Sybrandt, J., Matta, K., Shtutman, M., Safro, I., 2021. Accelerating COVID-19 research with graph mining and transformer-based learning. ArXiv210207631 Cs.

Ubiquitous Knowledge Processing Lab, 2021. UKPLab/sentence-transformers.

National Library of Medicine, 2004. Unified Medical Language System (UMLS).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention Is All You Need. ArXiv170603762 Cs.

Voorhees, E., Alam, T., Bedrick, S., Demner-Fushman, D., Hersh, W.R., Lo, K., Roberts, K., Soboroff, I., Wang, L.L., 2020. TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection. ArXiv200504474 Cs.

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., Kanakia, A., 2020. Microsoft Academic Graph: When experts are not enough. Quant. Sci. Stud. 1, 396–413.

Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., …, 2020. CORD-19: The COVID-19 Open Research Dataset. ArXiv200410706 Cs.

Wang, X., Song, X., Li, B., Guan, Y., Han, J., 2020. Comprehensive Named Entity Recognition on CORD-19 with Distant or Weak Supervision. ArXiv200312218 Cs.

Wang, Y., Wang, L., Li, Y., He, D., 2013. A Theoretical Analysis of NDCG Ranking Measures

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P., Liu, H., 2018. A comparison of word embeddings for the biomedical natural language processing. J. Biomed. Inform. 87, 12–20.

WHO, 2020. Coronavirus disease (COVID-19).

WHO, 2020. WHO issues its first emergency use validation for a COVID-19 vaccine and emphasizes need for equitable global access

NHS UK, 2020. Who's at higher risk from coronavirus (COVID-19)

Wise, C., Ioannidis, V.N., Calvo, M.R., Song, X., Price, G., Kulkarni, N., Brand, R., Bhatia, P., Karypis, G., 2020. COVID-19 Knowledge Graph: Accelerating Information Retrieval and Discovery for Scientific Literature. ArXiv200712731 Cs.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., …, 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. ArXiv160908144 Cs.

Xiong, L., Xiong, C., Li, Y., Tang, K.-F., Liu, J., Bennett, P., Ahmed, J., Overwijk, A., 2020. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. ArXiv200700808 Cs.

Yang, W., Lu, K., Yang, P., Lin, J., 2019. Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models. Proc. 42nd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. 1129–1132.

Zhang, Y., Chen, Q., Yang, Z., Lin, H., Lu, Z., 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH.

Zhu, Z., Lian, X., Su, X., Wu, W., Marraro, G.A., Zeng, Y., 2020. From SARS and MERS to COVID-19: a brief summary and comparison of severe acute respiratory infections caused by three highly pathogenic human coronaviruses.

FDA, 2019. First FDA-approved vaccine for the prevention of Ebola virus disease, marking a critical milestone in public health preparedness and response.

Stanford University, 2016. SQuAD The Stanford Question Answering Dataset

National Library of Medicine, 2021. Medical Subject Headings

The Internet Archive, 2021. Wayback Machine

ClinicalTrials.gov https://clinicaltrials.gov/