medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

GatorTron: A Large Language Model for Clinical Natural Language Processing

Authors: Xi Yang, PhD[1,2], Nima PourNejatian, PhD[3], Hoo Chang Shin, PhD[3], Kaleb E Smith, PhD[3], Christopher Parisien, PhD[3], Colin Compas, PhD[3], Cheryl Martin, BS[3], Mona G Flores, MD[3], Ying Zhang, MS[4], Tanja Magoc, PhD[5], Christopher A Harle, PhD[1,5], Gloria Lipori, MBA[5,6], Duane A Mitchell, MD[7], PhD, William R Hogan, MD, MS[1], Elizabeth A Shenkman, PhD[1], Jiang Bian, PhD[1,2], Yonghui Wu, PhD[1,2] *

Affiliation of the authors:

[1]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, USA.

[2]Cancer Informatics and eHealth core, University of Florida Health Cancer Center, Gainesville, Florida, USA.

[3]NVIDIA, Santa Clara, California, USA.

[4]Research Computing, University of Florida, Gainesville, Florida, USA.

[5]Integrated Data Repository Research Services, University of Florida, Gainesville, Florida, USA.

[6]University of Florida Health and Shands Hospital, Gainesville, FL

[7]Lillian S. Wells Department of Neurosurgery, UF Clinical and Translational Science Institute, University of Florida.

Corresponding author: Yonghui Wu, PhD
Clinical and Translational Research Building
2004 Mowry Road, PO Box 100177
Gainesville, FL, USA, 32610
Phone: 352-294-8436
Email: yonghui.wu@ufl.edu

Keywords: Natural Language Processing
Transformer Model
Deep Learning
Electronic Health Records

Word count: 3,795

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

ABSTRACT

Objective

To develop a large pretrained clinical language model from scratch using transformer architecture; systematically examine how transformer models of different sizes could help 5 clinical natural language processing (NLP) tasks at different linguistic levels.

Methods

We created a large corpus with >90 billion words from clinical narratives (>82 billion words), scientific literature (6 billion words), and general English text (2.5 billion words). We developed GatorTron models from scratch using the BERT architecture of different sizes including 345 million, 3.9 billion, and 8.9 billion parameters, compared GatorTron with three existing transformer models in the clinical and biomedical domain on 5 different clinical NLP tasks including clinical concept extraction, relation extraction, semantic textual similarity, natural language inference, and medical question answering, to examine how large transformer models could help clinical NLP at different linguistic levels.

Results and Conclusion

GatorTron scaled up transformer-based clinical language models to a size of 8.9 billion parameters and achieved state-of-the-art performance on 5 clinical NLP tasks of different linguistic levels targeting various healthcare information documented in unstructured electronic health records (EHRs). The proposed GatorTron models performed remarkably better in much complex clinical NLP tasks such as natural language inference (9.6% and 7.5% improvements) and question answering (9.5% and 7.77% improvements) compared with existing smaller clinical

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

transformer models (i.e., BioBERT and ClinicalBERT), demonstrating the potential of large transformer-based clinical models for advanced medical artificial intelligent (AI) applications such as question answering.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

INTRODUCTION

There has been an increasing interest in developing artificial intelligence (AI) systems by leveraging large electronic health records (EHRs). A critical step in developing medical AI systems is to enable machines to use patients' clinical characteristics captured in longitudinal EHRs. The more information we know about patients, the better medical AI systems that we can develop. In recent decades, hospitals and medical practices in the United States (US) rapidly adopted EHR systems[1,2], resulting in massive stores of electronic patient data, including structured (e.g., disease codes, medication codes) and unstructured (i.e., clinical narratives such as physicians' progress notes and discharge summaries). Physicians and other healthcare workers widely use clinical narratives to document detailed patient information as free text in EHRs. [3] There is an increasing number of studies exploring the rich, more fine-grained information about the patients in clinical narratives that led to improved diagnostic and prognostic models.[4,5] Nevertheless, free-text narratives cannot be easily used in computational models that usually require structured data. Researchers have increasingly turned to natural language processing (NLP) as the key technology to fill the gap of using clinical narratives in clinical studies[6].

Recently, transformer-based deep learning models have become state-of-the-art for many clinical NLP tasks. Compared with traditional machine learning models, transformer-based NLP models usually have very large number of parameters (e.g., 345 million parameters in BERT) to enable automated knowledge learning from a massive amount of text data. There is an increasing interest in examining how scaling up the model size could improve NLP. In the general NLP domain, many large transformer-based NLP models have been developed, such as the Generative

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Pre-trained Transformer 3 (GPT-3) model [7], which has 175 billion parameters and was trained using >400 billion words of text. However, few studies have explored large (i.e., billions of parameters) transformer models in the clinical domain. To date, the largest transformer model using clinical narratives, ClinicalBERT[8], has 110 million parameters and was trained using 0.5 billion words of clinical text. It is unclear how transformer-based models developed using significantly more clinical narrative text and more parameters may improve medical AI systems. In this study, we developed a large clinical transformer model, GatorTron, using >90 billion words of clinical narratives, scientific literature, and general English text. We trained GatorTron from scratch using UF's HiperGator-AI cluster and empirically evaluated three models with different settings including (1) a base model with 345 million parameters, (2) a medium model with 3.9 billion parameters, and (3) a large model with 8.9 billion parameters. We compared GatorTron models with existing large transformer models trained using biomedical literature and clinical narratives on 5 clinical NLP tasks including clinical concept extraction (or clinical named entity recognition [CNER]), medical relation extraction (MRE), semantic textual similarity (STS), natural language inference (NLI), and medical question answering (MQA). GatorTron outperformed previous transformer models on 5 clinical NLP tasks at different linguistic levels targeting various patient information.

BACKGROUND

Researchers have applied various methods including rule-based, machine learning-based, and hybrid solutions in clinical NLP. [9,10] At present, most state-of-the-art NLP models are based on machine learning models. For a long time, researchers had to train different machine learning models for different NLP tasks. Today, most state-of-the-art clinical NLP solutions are based on deep learning models[11] implemented using neural network architectures – a fast-developing

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

sub-domain of machine learning. Convolutional neural networks[12] (CNN) and recurrent neural networks[13] (RNN) have been successfully applied to NLP in the early stage of deep learning. The RNN model implemented using bidirectional long-short term memory (LSTM) with a CRFs layer (LSTM-CRFs) has been widely used in CNER and relation extraction. [14–16] More recently, the transformer architectures[17] (e.g., BERT) implemented with a self-attention mechanism[18] have become state-of-the-art, achieving best performance on many NLP benchmarks. [19–22] In the general domain, the transformer-based NLP models have achieved state-of-the-art performance for many NLP tasks including name entity recognition[23–25], relation extraction[26–30], sentence similarity[31–33], natural language inference[33–36], and question answering[33,34,37,38]. Notably, transformers perform more effectively by decoupling of language model pretraining (i.e., pretrain language models using large unlabeled text corpora) and fine-tuning (i.e., applying the learned language models solving specific tasks often with labeled training data) into two independent phases. After successful pretraining, the learned language model can be used to solve a variety of NLP subtasks through fine-tuning, which is known as transfer learning – a strategy to learn knowledge from one task and apply it in another task[39]. Human language has a very large sample space – the possible combinations of words and sentences are innumerable. Recent studies show that large transformer models trained using massive text data are remarkably better than traditional NLP models in terms of emergence and homogenization.[39]

In the clinical domain, researchers have identified several fundamental NLP tasks such as CNER, MRE, STS, NLI, and MQA. CNER is to recognize phrases that have important clinical meanings (e.g., medications, treatments, adverse drug events). The NLP system has to determine

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

the boundaries of a concept and classify it into predefined semantic categories. Early systems for clinical concept extract are often rule-based, yet, most recent systems are based on machine learning models such as conditional random fields (CRFs)[40,41], CNN [12,42], and LSTM-CRFs [13,14]. Current state-of-the-art solutions for CNER are mainly based on transformers[43]. MRE is to establish medical-related relations (e.g., drug induce adverse events) among clinical concepts (e.g., drugs, adverse events). MRE is usually approached as a classification problem – identify and classify pairs of concepts with valid relations. Various machine learning-based classifiers such as support vector machines (SVMs), random forests (RF), and gradient boosting trees (GBT)[16] have been applied. With the emergence of deep learning models, researchers have explored the LSTM architecture for RE in both general and clinical domains[44,45]. Most recently, several studies adopted the BERT architecture and demonstrated superior performance for MRE on various datasets[43,46–50]. The STS task is to quantitatively assess the semantic similarity between two text snippets (e.g., sentences), which is usually approached as a regression task where a real-value score was used to quantify the similarity between two text snippets. In the general domain, the STS benchmark (STS-B) dataset curated by the Semantic evaluation (SemEval) challenges between 2012 and 2017[51] is widely used for evaluating STS systems[19]. Various machine learning methods have been examined[52–54] but transformer-based systems such as RoBERTa[31], T5[33], and ALBERT[34] are leading the state-of-the-art models for STS. In the clinical domain, the MedSTS dataset[55] that consists of over 1,000 annotated sentence pairs from clinical notes at Mayo Clinic was widely used as the benchmark. MedSTS was used as the gold standard in two clinical NLP open challenges including the 2018 BioCreative/Open Health NLP (OHNLP) challenge[56] and 2019 n2c2/OHNLP ClinicalSTS shared task[57]. Similar to the general

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

domain, pretrained transformer-based models using clinical text and biomedical literature, including ClinicalBERT and BioBERT[58], are current solutions for STS. NLI is also known as recognizing textual entailment (RTE) - a directional relation between text fragments (e.g., sentences)[59]. The goal of NLI is to determine if a given hypothesis can be inferred from a given premise. In the general domain, two benchmark datasets - the MultiNLI[60] and the Stanford NLI[61] are widely used. On both datasets, pretrained transformer models achieved state-of-the-art performances[33,35]. There are limited resources for NLI in the clinical domain. Until recently, the MedNLI – a dataset annotated by doctors based on the medical history of patients[62] was developed as a benchmark dataset in the clinical domain. A previous study[8] showed that a pretrained clinical BERT model achieved the state-of-the-art performance and outperformed the baseline (InferSent[63]) by ~9% accuracy. The MQA task is to build NLP systems that automatically answer medical questions in a natural language. Unlike other tasks focusing on phrases and sentences, MQA is a document-level task that requires information from the whole document to generate answers according to questions. In the general domain, the Stanford Question Answering Datasets (SQuAD 1.1 and 2.0)[64,65] have been widely used as benchmarks. Transformer-based models are the state-of-the-art for both SQuAD1.1[24] and SQuAD2.0[37]. There are several MQA datasets developed in the past few years such as the MESHQA[66], MedQuAD[67], and emrQA[68].

The promise of transformer-based NLP models has led to further interest in exploring how increases in model and data size may improve large (e.g., >billions of parameters) transformer models processing clinical narratives. In the biomedical domain, researchers developed BioBERT[17] (with 110 million parameters) and PubMedBERT[69] (110 million parameters)

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

transformer models using text from PubMed literature. Previously, we developed BioMegatron models in the biomedical domain with different sizes from 345 million to 1.2 billion parameters[70] using PubMed literature. However, few studies have explored large-size transformer models in the clinical domain due to the sensitive nature of clinical narratives that contain Protected Health Information (PHI) and the requirement of massive computing power. By developing not only larger models, but models that use clinical narratives, NLP may perform better in utilizing patient information in ways that can be applied to medical AI systems. To date, the largest transformer model using clinical narratives is ClinicalBERT[8]. ClinicalBERT has 110 million parameters and was trained using 0.5 billion words from the publicly available Medical Information Mart for Intensive Care III[71] (MIMIC-III) dataset. It is unclear how transformer-based models developed using significantly more clinical narrative text and more parameters may improve medical AI systems in extracting and utilizing patient information.

MATERIALS AND METHODS

Data Source

The primary data source for this study is the clinical narratives from UF Health Integrated Data Repository (IDR), a research data warehouse of UF Health. We collected a total of 290,482,002 clinical notes from 2011 to early 2021. The data included >82 billion medical words from >290 million notes related to >2 million patients and >50 million patient care encounters. We applied a standard preprocessing pipeline to remove duplicated notes and clean the clinical text – unify character encoding, identify tokens and sentence boundaries. Then, we merged the >82 billion words of clinical corpus with 6 billion words from PubMed (combining PubMed abstracts and full-text commercial-collection)[70], 2.5 billion words from Wikipedia[70], and 0.5 billion

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

words from the MIMIC-III corpus[71] to generate a corpus with > 90 billion words. This study was approved by the UF Institutional Review Board (IRB202100049).

Study design

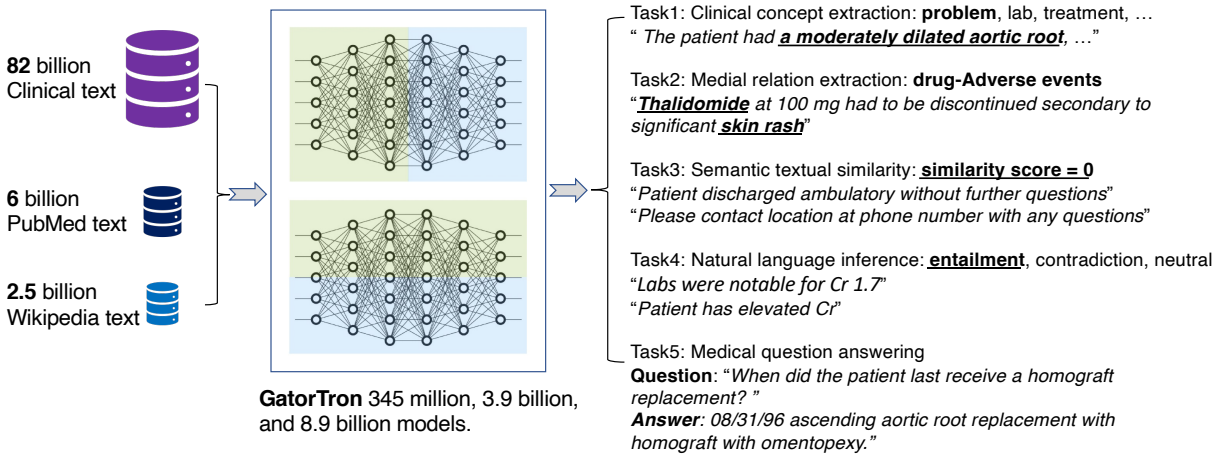

Figure 1. An overview of study design.

Figure 1 shows an overview of the study design. We seek to train a large clinical transformer model, GatorTron, using >90 billion words and examine how and whether scaling up mode size improves clinical NLP tasks. Following standard practice, we first pretrained GatorTron using the >90 billion words as an unsupervised learning procedure and then applied GatorTron to 5 different clinical NLP tasks using a supervised fine-tuning procedure. We adopted the BERT architecture implemented in MagaTron-LM[70] and explored three different settings including a base model of 345 million parameters (i.e., GatorTron-base), a medium model of 3.9 billion parameters (i.e., GatorTron-medium), and a large model of 8.9 billion parameters (i.e., GatorTron-large). Then we compared GatorTron models with an existing transformer model from the clinical domain, ClinicalBERT (trained with 110 million parameters) and two transformer models from the biomedical domain, including, BioBERT (345 million parameters)

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

and BioMegatron (1.2 billion parameters). We examined the models on 5 clinical NLP tasks, including CNER, MRE, STS, NLI, and MQA. We used 6 public clinical benchmark datasets (**Table 2** and **Table 3**) following the default training/test settings and calculated evaluation scores using official evaluation scripts associated with each benchmark dataset.

Training environment

We used a total number of 992 Nvidia DGX A100 GPUs from 124 superPOD nodes at UF's HiperGator-AI cluster to train GatorTron models by leveraging both data-level and model-level parallelisms implemented by the Megatron-LM package[72]. We monitored the training progress by training loss and validation loss and stopped the training when there was no further improvement (i.e., the loss plot became flat).

GatorTron Model Configuration

We developed GatorTron models using three configurations and determined the number of layers, hidden sizes, and number of attention heads according to the guidelines for optimal depth-to-width parameter allocation proposed by Levin et al[73] as well as our previous experience in developing BioMegatron[70]. **Table 1** provides detailed information for the three settings. The GatorTron-base model has 24 layers of transformer blocks, which is similar to the architecture of BERT large model. For each layer, we set the number of hidden units as 1024 and attention heads as 16. The GatorTron-medium model scaled up to 3.9 billion parameters (~10 times of the base setting) and the GatorTron-large model scaled up to 8.9 billion parameters, which is similar to BioMegatron[72] (with 8.3 billion parameters).

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 1. Three configurations of GatorTron model.

Model	# Layers	# Hidden Size	# Attention Heads	# Parameters
GatorTron-base	24	1024	16	345 million
GatorTron-medium	48	2560	40	3.9 billion
GatorTron-large	56	3584	56	8.9 billion

Existing transformer models for comparison

BioBERT.[17] The BioBERT model was developed by further training the original BERT-large model (345 million parameters, 24 layers, 1024 hidden units, and 16 attention heads) using biomedical literature from PubMed Abstracts (4.5 billion words) and PMC Full-text articles (13.5 billion words). In this study, we used version 1.1.

ClinicalBERT.[8] The ClinicalBERT model was developed by further training the BioBERT (base version; 110 million parameters with 12 layers, 768 hidden units, and 12 attention heads) using clinical text from the MIMIC-III[71] corpus.

BioMegatron.[70] The BioMegatron models adopted the BERT architecture with a different number of parameters from 345 million to 1.2 billion. Different from BioBERT and ClinicalBERT, the BioMegatron was trained from scratch without leveraging the original BERT model.

Clinical NLP tasks, evaluation matrices, and benchmark datasets

We evaluated GatorTron models using 5 clinical NLP tasks and 6 public benchmark datasets. For CNER, we used three benchmark datasets developed by the 2010 i2b2 challenge, 2012 i2b2 challenge, and 2018 n2c2 challenge to evaluate GatorTron models on identifying various important medical concepts from clinical text. We used standard precision, recall, and F1-score

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

for evaluation. For MRE, we used the dataset developed by the 2018 n2c2 challenge with a focus on relations between medications and adverse drug events. The standard precision, recall, and F1-score were used for evaluation. For STS, we used the dataset developed by the 2019 n2c2/OHNLP challenge on clinical semantic textural similarity[57]. We used the Pearson correlation score for evaluation. For NLI, we evaluated the Gatortron models using the MedNLI dataset and used accuracy for comparison. We used the emrQA dataset, a benchmark dataset widely used for MQA, to evaluate GatorTron. We particularly focused on medications and relations-related questions as Yue et al.[74] found that the two subsets are more consistent. We utilized both F1-score and exact match score for evaluation.

RESULTS

Table 2 and **Table 3** compare GatorTron models with two existing biomedical transformer models (BioBERT and BioMegatron) and one clinical transformer model (Clinical BERT) on 5 clinical NLP tasks.

Table 2. Comparison of GatorTron with existing biomedical and clinical transformer models for CNER and medical MRE.

	CNER									**MRE**		
	2010 i2b2[75]			**2012 i2b2[76]**			**2018 n2c2**[16]			**2018 n2c2**[16]		
Transformer	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
BioBERT	0.8693	0.8653	0.8673	0.7478	0.8037	0.7747	0.8634	0.8921	0.8775	0.9663	0.9451	0.9555
ClinicalBERT	NA	NA	0.8780	NA	NA	0.7890	0.8592	0.8832	0.8710	0.9678	0.9414	0.9544
BioMegatron	0.8614	0.8761	0.8687	0.7591	0.8031	0.7805	0.8707	0.8915	0.8810	0.9711	0.9434	0.9571
GatorTron-base	0.8748	0.9043	0.8893	0.7644	0.8221	0.7922	0.8759	0.9038	0.8896	0.9719	0.9482	0.9599
GatorTron-medium	0.8869	0.9122	0.8994	0.7812	0.8245	0.8022	0.8954	0.9035	0.8994	0.9721	0.9503	0.9611
GatorTron-large	0.8880	0.9116	**0.8996**	0.7862	0.8333	**0.8091**	0.8979	0.9021	**0.9000**	0.9776	0.9482	**0.9627**

CNER: clinical named entity recognition; MRE: medical relation extraction; Pre: precision; Rec: recall; F1: F1-score; Clinical concepts in 2010 i2b2 and 2012 i2b2 challenges: problems, treatments, lab tests; clinical concepts in 2018 n2c2 challenge: drugs, adverse events, and drug-related attributes (e.g., dose). Medical relation in 2018 n2c2 challenge: drug induced adverse events. Best F1 scores are bolded. NA: scores not reported.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 3. Comparison of GatorTron with existing biomedical and clinical transformer models for STS, NLI, and MQA.

	STS	**NLI**	**MQA**			
	2019 n2c2[57]	**MedNLI**[62]	**emrQA Medication**[68]		**emrQA Relation**[68]	
Transformer	Pearson correlation	Accuracy	F1 score	Exact Match	F1 score	Exact Match
BioBERT	0.8744	0.8050	0.6997	0.2475	0.9262	0.8361
ClinicalBERT	0.8787	0.8270	0.6905	0.2406	0.9306	0.8533
BioMegatron	0.8806	0.8390	0.7231	0.2882	0.9405	0.879
GatorTron-base	0.8810	0.8670	0.7181	0.2978	0.9543	0.9029
GatorTron-medium	**0.8903**	0.8720	0.7354	0.3018	0.9677	0.9243
GatorTron-large	0.8896	**0.9020**	**0.7408**	**0.3155**	**0.9719**	**0.9310**

STS: semantic textual similarity; NLI: natural language inference; MQA: medial question answering; The best evaluation scores are bolded.

Recognize clinical concepts and medical relations. As shown in **Table 2**, all three GatorTron models outperformed existing biomedical and clinical transformer models in recognizing various types of clinical concepts on the three benchmark datasets (i.e., 2010 i2b2[75] and 2012 i2b2[76]: problem, treatments, lab tests; 2018 n2c2[16]: drug, adverse events, and drug-related attributes). The GatorTron-large model outperformed the other two smaller GatorTron models and achieved the best F1-scores of 0.8996, 0.8091, and 0.9000, respectively, demonstrating performance gain from scaling up the size of the model. For MRE, the GatorTron-large model also achieved the best F1-score of 0.9627 for identifying drug-cause-adverse event relations outperforming existing biomedical and clinical transformers and the other two smaller GatorTron models. We observed performance improvement when scaling up the size of GatorTron model.

Assess semantic textual similarity. As shown in **Table 3**, all GatorTron models outperformed existing biomedical and clinical transformer models in assessing STS. Among the three GatorTron models, the GatorTron-medium model achieved the best Pearson correlation score of 0.8903, outperforming both GatorTron-base and GatorTron-large. Although we did not observe consistent improvement by scaling up the size of the GatorTron model, the GatorTron-large

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

model significantly outperformed GatorTron-base and its performance is very close to the GatorTron-medium model (0.8896 vs. 0.8903).

Natural language inference. GatorTron models outperformed existing biomedical and clinical transformers, and the GatorTron-large model achieved the best accuracy of 0.9020, outperforming the BioBERT and ClinicalBERT by 9.6% and 7.5%, respectively. We observed a monotonic performance improvement by scaling up the size of GatorTron.

Medical question answering. All GatorTron models outperformed existing biomedical and clinical transformer models in MQA (e.g., "What lab results does patient have that are pertinent to diabetes diagnosis?"). For medication-related questions, the GatorTron-large model achieved the best exact match score of 0.3155, outperforming the BioBERT and ClinicalBERT by 6.8% and 7.5%, respectively. For relation-related questions, GatorTron-large also achieved the best exact match score of 0.9301, outperforming BioBERT and ClinicalBERT by 9.5% and 7.77%, respectively. We also observed a monotonic performance improvement by scaling up the model size of GatorTron.

DISCUSSION

In this study, we developed a large pretrained language model, GatorTron, using a corpus of >90 billion words. We trained GatorTron from scratch with different model sizes and evaluated its performance on 5 clinical NLP tasks at different linguistic levels (phrase level, sentence level, and document level) using 6 publicly-available benchmark datasets from the clinical domain. The experimental results show that GatorTron models outperformed existing biomedical and clinical transformers for all 5 clinical NLP tasks. We observed monotonic improvements by scaling up the model size of GatorTron for 4 of the 5 tasks, excluding the STS task. Our

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

GatorTron model also outperformed the BioMegatron[70], a transformer model with a similar model size developed in our previous study using >8.5 billion words from PubMed and Wikipedia (a small proportion of the >90 billion words of corpus for developing GatorTron). This study scaled up the clinical transformer models to 8.9 billion parameters in the clinical domain and demonstrated performance improvements. To the best of our knowledge, GatorTron-large is the largest transformer model in the clinical domain.

Scaling up model size and performance improvement. There is an increasing interest in examining massive-size deep learning models in NLP as they demonstrated novel abilities such as emergence and homogenization[39]. In the general domain, the Megatron-Turing NLG model has scaled up to 530 billion parameters following the GPT-3[7] model with 175 billion parameters. However, there are limited studies examining large transformer models in the clinical domain due to the sensitive nature of clinical text and massive computing requirements. Prior to our study, the largest transformer in the clinical domain was ClinicalBERT with 110 million parameters trained using 0.5 billion words. Our study scaled the transformer to 8.9 billion parameters and demonstrated performance improvement for 5 clinical NLP tasks on 6 public benchmark datasets. Among the 5 tasks, GatorTron achieved significant improvements for sentence-level and document-level NLP tasks such as NLI and MQA, but moderate improvements for phrase-level tasks such as CNER and MRE, indicating that large transformer models are more helpful to sentence-level and document-level NLP tasks.

Model size and converge speed. GatorTron was pretrained using unsupervised learning to optimize a mask language model (MLM). We monitored training loss and calculated validation loss using a subset set of the clinical text (5%) to determine when to stop the training. **Figure 2** shows the training loss and validation loss for GatorTron models with three different settings.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

We observed that the larger GatorTron models converged faster than the base model. For example, the GatorTron-base model converged in 10 epochs, whereas the medium and large models converged in 7 epochs. This may indicate that larger transformer models learn faster than smaller models. The training of the GatorTron-large model used about 6 days on 992 GPUs from 124 Nvidia SuperPOD nodes.

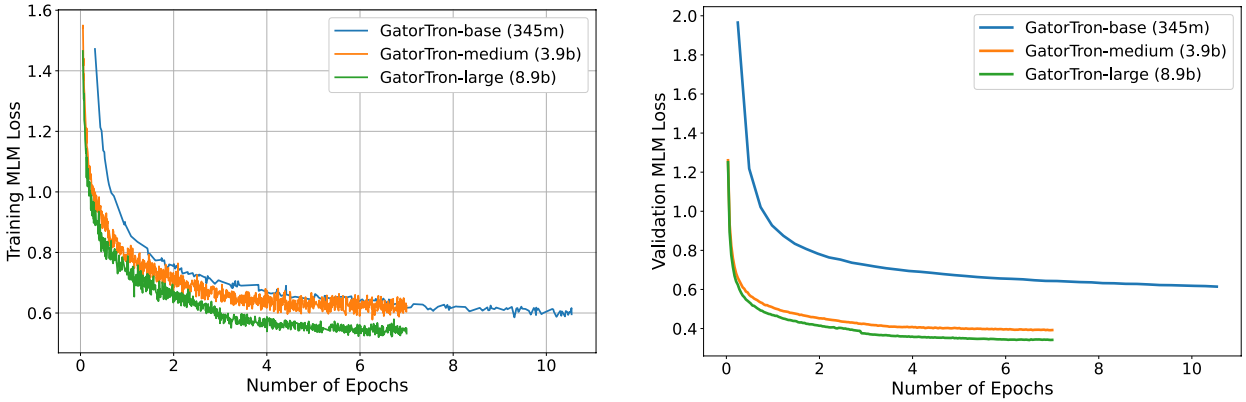

Figure 2. Training loss and validation loss for GatorTron base (345 million), medium (3.9 billion), and large (8.9 billion) models.

Potentials in improving healthcare delivery and patient outcomes. GatorTron models perform better in utilizing patient information in clinical narratives, which can be applied to various medical AI systems. The rich, fine-grained patients' information captured in clinical narratives is a critical resource powering medical AI system. With better performance in information extraction tasks (e.g., CNER and MRE), GatorTron models have potential to provide more accurate patients' information to identify research-standard patient cohorts using computable phenotypes, support physicians making data-informed decisions by clinical decision support systems, and identify adverse events associated with drug exposures via pharmacovigilance. The significant improvements in STS, NLI, and MQA can be applied for

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

deduplication of clinical text, mining medial knowledge, and developing next-generation medical AI systems that can interact with patients using human language. The emergence and homogenization abilities[39] inherited from a large transformer architecture make it convenient to apply GatorTron to many other AI tasks through fine-tuning. We believe that GatorTron will improve the use of clinical narratives in developing various medical AI systems for better healthcare delivery and health outcomes.

This study has limitations. We mainly focused on medication and relation-related questions when evaluating GatorTron models due to the limitation of the benchmark dataset for MQA. Future studies should examine the benefit of large clinical transformer models to downstream medical applications such as disease phenotyping and patient cohort construction.

CONCLUSION

Large pretrained clinical language models could benefit a number of downstream clinical NLP tasks, especially for complex NLP tasks such as MQA.

ACKNOWLEDGMENTS

None

FUNDING STATEMENT

This study was partially supported by a Patient-Centered Outcomes Research Institute® (PCORI®) Award (ME-2018C3-14754), a grant from the National Cancer Institute, 1R01CA246418 R01, grants from the National Institute on Aging, NIA R56AG069880 and R21AG062884, and the Cancer Informatics and eHealth core jointly supported by the UF Health

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Cancer Center and the UF Clinical and Translational Science Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding institutions.

Support from UF Research Computing: We would like to thank the UF Research Computing team, led by Dr. Erik Deumens, for providing computing power through UF HiperGator-AI cluster.

COMPETING INTERESTS STATEMENT

Authors have no competing financial interests.

CONTRIBUTORSHIP STATEMENT

XY, YW, JB, NP, and MGF were responsible for the overall design, development, and evaluation of this study. XY had full access to all the data in the study, conducted all the experiments, and takes responsibility for the integrity of the data and the accuracy of the data analysis. XY, YW, JB, and WH did the bulk of the writing, EAS, DAM, TM and CAH also contributed to writing and editing of this manuscript. All authors reviewed the manuscript critically for scientific content, and all authors gave final approval of the manuscript for publication.

ETHICS STATEMENT

IRB (202100049) of the University of Florida gave approval for this work as exempt. The approval includes but is not limited to HIPAA waiver to enroll.

SUPPLEMENTARY MATERIAL

None.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

REFERENCES

1 Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015. /evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php (accessed 20 Dec 2019).
2 Adler-Milstein J, Holmgren AJ, Kralovec P, *et al.* Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. *J Am Med Inform Assoc* 2017;**24**:1142–8.
3 Meystre SM, Savova GK, Kipper-Schuler KC, *et al.* Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 2008;:128–44.
4 Liang H, Tsui BY, Ni H, *et al.* Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med* 2019;**25**:433–8.
5 Yang J, Lian JW, Chin Y-P (Harvey), *et al.* Assessing the Prognostic Significance of Tumor-Infiltrating Lymphocytes in Patients With Melanoma Using Pathologic Features Identified by Natural Language Processing. *JAMA Network Open* 2021;**4**:e2126337.
6 Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc* 2011;**18**:544–51.
7 Floridi L, Chiriatti M. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds & Machines* 2020;**30**:681–94.
8 Alsentzer E, Murphy J, Boag W, *et al.* Publicly Available Clinical BERT Embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: : Association for Computational Linguistics 2019. 72–8.https://www.aclweb.org/anthology/W19-1909 (accessed 12 Jun 2020).
9 Wang Y, Wang L, Rastegar-Mojarad M, *et al.* Clinical information extraction applications: A literature review. *J Biomed Inform* 2018;**77**:34–49.
10 Wu S, Roberts K, Datta S, *et al.* Deep learning in clinical natural language processing: a methodical review. *J Am Med Inform Assoc* 2020;**27**:457–70.
11 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**:436–44.
12 Collobert R, Weston J, Bottou L, *et al.* Natural Language Processing (Almost) from Scratch. *J Mach Learn Res* 2011;**12**:2493–537.
13 Lample G, Ballesteros M, Subramanian S, *et al.* Neural Architectures for Named Entity Recognition. *arXiv:160301360 [cs]* Published Online First: 4 March 2016.http://arxiv.org/abs/1603.01360 (accessed 2 Mar 2018).
14 Wu Y, Yang X, Bian J, *et al.* Combine factual medical knowledge and distributed word representation to improve clinical named entity recognition. In: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association 2018. 1110.
15 Yang X, Lyu T, Li Q, *et al.* A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Med Inform Decis Mak* 2019;**19**:232.
16 Yang X, Bian J, Fang R, *et al.* Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. *Journal of the American Medical Informatics Association* 2020;**27**:65–72.
17 Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* Published Online First: 2019.https://doi.org/10.1093/bioinformatics/btz682

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

18 Vaswani A, Shazeer N, Parmar N, *et al.* Attention is All you Need. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc. 2017. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (accessed 28 Oct 2021).
19 Wang A, Singh A, Michael J, *et al.* GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv:180407461 [cs]* Published Online First: 22 February 2019.http://arxiv.org/abs/1804.07461 (accessed 21 Mar 2020).
20 Wang A, Pruksachatkun Y, Nangia N, *et al.* SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *arXiv:190500537 [cs]* Published Online First: 12 February 2020.http://arxiv.org/abs/1905.00537 (accessed 8 May 2021).
21 Qiu X, Sun T, Xu Y, *et al.* Pre-trained Models for Natural Language Processing: A Survey. *arXiv:200308271 [cs]* Published Online First: 24 April 2020.http://arxiv.org/abs/2003.08271 (accessed 8 May 2021).
22 Tay Y, Dehghani M, Bahri D, *et al.* Efficient Transformers: A Survey. *arXiv:200906732 [cs]* Published Online First: 16 September 2020.http://arxiv.org/abs/2009.06732 (accessed 8 May 2021).
23 Yu J, Bohnet B, Poesio M. Named Entity Recognition as Dependency Parsing. *arXiv:200507150 [cs]* Published Online First: 13 June 2020.http://arxiv.org/abs/2005.07150 (accessed 29 Oct 2021).
24 Yamada I, Asai A, Shindo H, *et al.* LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. *arXiv:201001057 [cs]* Published Online First: 2 October 2020.http://arxiv.org/abs/2010.01057 (accessed 29 Oct 2021).
25 Li X, Sun X, Meng Y, *et al.* Dice Loss for Data-imbalanced NLP Tasks. *arXiv:191102855 [cs]* Published Online First: 29 August 2020.http://arxiv.org/abs/1911.02855 (accessed 29 Oct 2021).
26 Xu B, Wang Q, Lyu Y, *et al.* Entity Structure Within and Throughout: Modeling Mention Dependencies for Document-Level Relation Extraction. *arXiv:210210249 [cs]* Published Online First: 19 February 2021.http://arxiv.org/abs/2102.10249 (accessed 29 Oct 2021).
27 Ye D, Lin Y, Sun M. Pack Together: Entity and Relation Extraction with Levitated Marker. *arXiv:210906067 [cs]* Published Online First: 10 October 2021.http://arxiv.org/abs/2109.06067 (accessed 29 Oct 2021).
28 Cohen AD, Rosenman S, Goldberg Y. Relation Classification as Two-way Span-Prediction. *arXiv:201004829 [cs]* Published Online First: 17 April 2021.http://arxiv.org/abs/2010.04829 (accessed 29 Oct 2021).
29 Lyu S, Chen H. Relation Classification with Entity Type Restriction. *arXiv:210508393 [cs]* Published Online First: 18 May 2021.http://arxiv.org/abs/2105.08393 (accessed 29 Oct 2021).
30 Wang J, Lu W. Two are Better than One: Joint Entity and Relation Extraction with Table-Sequence Encoders. *arXiv:201003851 [cs]* Published Online First: 8 October 2020.http://arxiv.org/abs/2010.03851 (accessed 29 Oct 2021).
31 Jiang H, He P, Chen W, *et al.* SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 2020;:2177–90.
32 Yang Z, Dai Z, Yang Y, *et al.* XLNet: Generalized Autoregressive Pretraining for Language Understanding. In: *NeurIPS*. 2019.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

33 Raffel C, Shazeer N, Roberts A, *et al.* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv:191010683 [cs, stat]* Published Online First: 24 October 2019.http://arxiv.org/abs/1910.10683 (accessed 27 Mar 2020).

34 Lan Z-Z, Chen M, Goodman S, *et al.* ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *ArXiv* 2019;**abs/1909.11942**.

35 Wang S, Fang H, Khabsa M, *et al.* Entailment as Few-Shot Learner. *arXiv:210414690 [cs]* Published Online First: 29 April 2021.http://arxiv.org/abs/2104.14690 (accessed 29 Oct 2021).

36 Zhang Z, Wu Y, Zhao H, *et al.* Semantics-aware BERT for Language Understanding. *arXiv:190902209 [cs]* Published Online First: 4 February 2020.http://arxiv.org/abs/1909.02209 (accessed 29 Oct 2021).

37 Zhang Z, Yang J, Zhao H. Retrospective Reader for Machine Reading Comprehension. *arXiv:200109694 [cs]* Published Online First: 11 December 2020.http://arxiv.org/abs/2001.09694 (accessed 29 Oct 2021).

38 Garg S, Vu T, Moschitti A. TANDA: Transfer and Adapt Pre-Trained Transformer Models for Answer Sentence Selection. *arXiv:191104118 [cs]* Published Online First: 20 November 2019.http://arxiv.org/abs/1911.04118 (accessed 29 Oct 2021).

39 Bommasani R, Hudson DA, Adeli E, *et al.* On the Opportunities and Risks of Foundation Models. *arXiv:210807258 [cs]* Published Online First: 18 August 2021.http://arxiv.org/abs/2108.07258 (accessed 16 Oct 2021).

40 Wu Y, Xu J, Jiang M, *et al.* A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text. *AMIA Annu Symp Proc* 2015;**2015**:1326–33.

41 Soysal E, Wang J, Jiang M, *et al.* CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association* 2018;**25**:331–6.

42 Wu Y, Jiang M, Lei J, *et al.* Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network. *Stud Health Technol Inform* 2015;**216**:624–8.

43 Yang X, Yu Z, Guo Y, *et al.* Clinical Relation Extraction Using Transformer-based Models. *arXiv preprint arXiv:210708957* 2021.

44 Kumar S. A Survey of Deep Learning Methods for Relation Extraction. *arXiv:170503645 [cs]* Published Online First: 10 May 2017.http://arxiv.org/abs/1705.03645 (accessed 8 May 2021).

45 Lv X, Guan Y, Yang J, *et al.* Clinical relation extraction with deep learning. *International Journal of Hybrid Information Technology* 2016;**9**:237–48.

46 Wei Q, Ji Z, Si Y, *et al.* Relation Extraction from Clinical Narratives Using Pre-trained Language Models. *AMIA Annu Symp Proc* 2020;**2019**:1236–45.

47 Guan H, Devarakonda M. Leveraging Contextual Information in Extracting Long Distance Relations from Clinical Notes. *AMIA Annu Symp Proc* 2020;**2019**:1051–60.

48 Alimova I, Tutubalina E. Multiple features for clinical relation extraction: A machine learning approach. *Journal of Biomedical Informatics* 2020;**103**:103382.

49 Mahendran D, McInnes BT. Extracting Adverse Drug Events from Clinical Notes. *arXiv:210410791 [cs]* Published Online First: 21 April 2021.http://arxiv.org/abs/2104.10791 (accessed 26 May 2021).

50 Yang X, Zhang H, He X, *et al.* Extracting Family History of Patients From Clinical Narratives: Exploring an End-to-End Solution With Deep Learning Models. *JMIR Medical Informatics* 2020;**8**:e22982.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

51 Cer D, Diab M, Agirre E, *et al.* Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:170800055* 2017.
52 Farouk M. Measuring Sentences Similarity: A Survey. *ArXiv* Published Online First: 2019.
53 Ramaprabha J, Das S, Mukerjee P. Survey on Sentence Similarity Evaluation using Deep Learning. *J Phys: Conf Ser* 2018;**1000**:012070.
54 Gomaa WH, Fahmy A. A Survey of Text Similarity Approaches. Published Online First: 2013.
55 Wang Y, Afzal N, Fu S, *et al.* MedSTS: a resource for clinical semantic textual similarity. *Lang Resources & Evaluation* 2020;**54**:57–72.
56 Rastegar-Mojarad M, Liu S, Wang Y, *et al.* BioCreative/OHNLP Challenge 2018. In: *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. New York, NY, USA: : ACM 2018. 575–575.http://doi.acm.org/10.1145/3233547.3233672
57 Wang Y, Fu S, Shen F, *et al.* Overview of the 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity. *JMIR Medical Informatics* 2020.
58 Mahajan D, Poddar A, Liang JJ, *et al.* Identification of Semantically Similar Sentences in Clinical Notes: Iterative Intermediate Training Using Multi-Task Learning. *JMIR Medical Informatics* 2020;**8**:e22508.
59 Dagan I, Glickman O, Magnini B. The PASCAL Recognising Textual Entailment Challenge. In: Quiñonero-Candela J, Dagan I, Magnini B, *et al.*, eds. *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*. Berlin, Heidelberg: : Springer Berlin Heidelberg 2006. 177–90.
60 Williams A, Nangia N, Bowman SR. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *arXiv:170405426 [cs]* Published Online First: 19 February 2018.http://arxiv.org/abs/1704.05426 (accessed 16 Aug 2020).
61 Bowman SR, Angeli G, Potts C, *et al.* A large annotated corpus for learning natural language inference. *arXiv:150805326 [cs]* Published Online First: 21 August 2015.http://arxiv.org/abs/1508.05326 (accessed 8 Nov 2021).
62 Shivade C. MedNLI — A Natural Language Inference Dataset For The Clinical Domain. 2017.https://physionet.org/content/mednli/ (accessed 23 Apr 2021).
63 Conneau A, Kiela D, Schwenk H, *et al.* Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *arXiv:170502364 [cs]* Published Online First: 8 July 2018.http://arxiv.org/abs/1705.02364 (accessed 6 Sep 2021).
64 Rajpurkar P, Zhang J, Lopyrev K, *et al.* SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv:160605250 [cs]* Published Online First: 10 October 2016.http://arxiv.org/abs/1606.05250 (accessed 16 Aug 2020).
65 Rajpurkar P, Jia R, Liang P. Know What You Don't Know: Unanswerable Questions for SQuAD. *arXiv:180603822 [cs]* Published Online First: 11 June 2018.http://arxiv.org/abs/1806.03822 (accessed 8 Nov 2021).
66 Zhu M, Ahuja A, Juan D-C, *et al.* Question Answering with Long Multiple-Span Answers. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: : Association for Computational Linguistics 2020. 3840–9.https://aclanthology.org/2020.findings-emnlp.342 (accessed 8 Nov 2021).
67 Ben Abacha A, Demner-Fushman D. A question-entailment approach to question answering. *BMC Bioinformatics* 2019;**20**:511.

medRxiv preprint doi: https://doi.org/10.1101/2022.02.27.22271257; this version posted March 18, 2022. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

68 Pampari A, Raghavan P, Liang J, *et al.* emrQA: A Large Corpus for Question Answering on Electronic Medical Records. *arXiv:180900732 [cs]* Published Online First: 3 September 2018.http://arxiv.org/abs/1809.00732 (accessed 24 Oct 2021).
69 Gu Y, Tinn R, Cheng H, *et al.* Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans Comput Healthcare* 2022;**3**:1–23.
70 Shin H-C, Zhang Y, Bakhturina E, *et al.* BioMegatron: Larger Biomedical Domain Language Model. *arXiv:201006060 [cs]* Published Online First: 13 October 2020.http://arxiv.org/abs/2010.06060 (accessed 5 Apr 2021).
71 Johnson AEW, Pollard TJ, Shen L, *et al.* MIMIC-III, a freely accessible critical care database. *Scientific Data* 2016;**3**:160035.
72 Shoeybi M, Patwary M, Puri R, *et al.* Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. *arXiv:190908053 [cs]* Published Online First: 13 March 2020.http://arxiv.org/abs/1909.08053 (accessed 20 Oct 2021).
73 Levine Y, Wies N, Sharir O, *et al.* The Depth-to-Width Interplay in Self-Attention. *arXiv:200612467 [cs, stat]* Published Online First: 17 January 2021.http://arxiv.org/abs/2006.12467 (accessed 23 Oct 2021).
74 Yue X, Gutierrez BJ, Sun H. Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset. *arXiv:200500574 [cs]* Published Online First: 1 May 2020.http://arxiv.org/abs/2005.00574 (accessed 6 Sep 2021).
75 Uzuner Ö, South BR, Shen S, *et al.* 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011;**18**:552–6.
76 Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J Am Med Inform Assoc* 2013;**20**:806–13.