GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records

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] *

Affiliations:

[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

Abstract: There is an increasing interest in developing massive-size deep learning models in natural language processing (NLP) - the key technology to extract patient information from unstructured electronic health records (EHRs). However, there are limited studies exploring large language models in the clinical domain; the current largest clinical NLP model was trained with 110 million parameters (compared with 175 billion parameters in the general domain). It is not clear how large-size NLP models can help machines understand patients' clinical information from unstructured EHRs. In this study, we developed a large clinical transformer model – GatorTron – using >90 billion words of text and evaluated it on 5 clinical NLP tasks including clinical concept extraction, relation extraction, semantic textual similarity, natural language inference, and medical question answering. GatorTron is now the largest transformer model in the clinical domain that scaled up from the previous 110 million to 8.9 billion parameters and achieved state-of-the-art performance on the 5 clinical NLP tasks targeting various healthcare information documented in EHRs. GatorTron models perform better in understanding and utilizing patient information from clinical narratives in ways that can be applied to improvements in healthcare delivery and patient outcomes.

Introduction

There has been an increasing interest in developing artificial intelligence (AI) systems to improve healthcare delivery and health outcomes by leveraging medical knowledge in electronic health records (EHRs). A critical step in developing medical AI systems is training machines to understand patients' clinical characteristics captured in longitudinal EHRs. The more information we know about the 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 in free text, leading to >80% of important patient information buried in unstructured text.[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 enable machines to understand clinical language used by humans to support various downstream clinical studies[6].

Historically, researchers designed several critical subtasks in clinical NLP such as clinical concept extraction[7] and relation extraction[8]. For a long time, researchers had to train different models for different NLP subtasks. Today, most NLP solutions are based on deep learning models[9] implemented using neural network architectures – a fast-developing sub-domain of

machine learning. Convolutional neural networks[10] (CNN) and recurrent neural networks[11] (RNN) have been successfully applied to NLP in the early stage of deep learning. More recently, the transformer architectures[12] (e.g., BERT) implemented with a self-attention mechanism[13] have become state-of-the-art, achieving best performance on many NLP benchmarks. [14–17] In the general domain, the transformer-based NLP models have achieved state-of-the-art performance for many NLP tasks including name entity recognition[18–20], relation extraction[21–25], sentence similarity[26–28], natural language inference[28–31], and question answering[28,29,32,33]. Notably, transformers perform more effectively by decoupling of language understanding (i.e., learning of language models using large unlabeled text corpora) and language utilization (i.e., applying the learned language models solving specific tasks often with labeled training data) into two independent phases, i.e., pretraining and fine-tuning. 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[34]. 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.[34]

The promise of transformer models has led to further interest in exploring how increases in model and data size may improve massive-size (e.g., >billions of parameters) transformer models understanding clinical narratives. The Generative Pre-trained Transformer 3 (GPT-3) model[35], which has 175 billion parameters and was trained using >400 billion words of text demonstrated superior performance and capability in many NLP tasks. In the biomedical

domain, researchers developed BioBERT[12] (with 110 million parameters) and PubMedBERT[36] (110 million parameters) transformer models using text from PubMed literature. Nvidia developed BioMegatron models in the biomedical domain with different sizes from 345 million to 1.2 billion parameters[37] 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 understanding and utilizing patient information in ways that can be applied to improvements in healthcare delivery and patient outcomes. To date, the largest transformer model using clinical narratives is ClinicalBERT[38]. ClinicalBERT has 110 million parameters and was trained using 0.5 billion words from the publicly available Medical Information Mart for Intensive Care III[39] (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 understanding, extracting, and utilizing patient information.

In this study, we developed a large clinical transformer model, GatorTron, using >90 billion words of text from the University of Florida (UF) Health, PubMed articles, and Wikipedia. We trained GatorTron from scratch 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 using 5 clinical NLP tasks including clinical concept extraction (or named entity recognition [NER]), medical relation extraction (MRE), semantic textual similarity (STS), natural language

inference (NLI), and medical question answering (MQA). GatorTron outperformed previous transformer models from the biomedical and clinical domain on 5 clinical NLP tasks at different linguistic levels targeting various patient information. This study significantly scaled up transformer models in the clinical domain from 110 million to 8.9 billion parameters and improved our understanding of how large transformer models help AI systems understand and utilize patient information documented in clinical narratives.

Results

We collected all clinical narrative notes from the UF Health Integrated Data Repository (IDR) 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 (articles and abstracts)[37], 2.5 billion words from Wikipedia[37], and 0.5 billion words from the MIMIC-III corpus[39] to generate a corpus with > 90 billion words. Using this corpus, we trained GatorTron with different settings (base - 345 million, medium - 3.9 billion, and large - 8.9 billion parameters) from scratch and fine-tuned GatorTron models for 5 clinical NLP tasks at different linguistic and complexity levels. Fig. 1 shows an overview of the study design. We used 6 public clinical benchmark datasets (Table 1 and Table 2) following the default training/test settings and calculated evaluation scores using official evaluation scripts associated with each benchmark dataset. Table 1 and Table 2 compare GatorTron models with two existing biomedical transformer models (BioBERT and BioMegatron) and one clinical transformer model (Clinical BERT) on the 5 clinical NLP tasks.

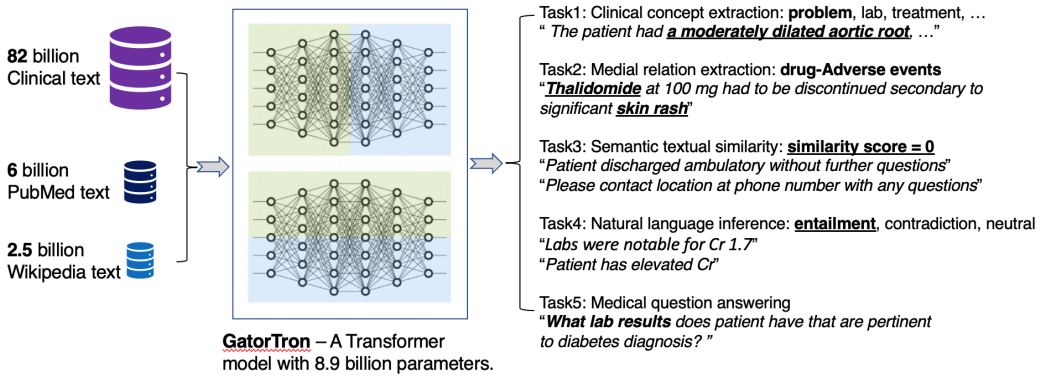

Fig. 1 An overview of study design.

Table 1. Comparison of GatorTron with existing biomedical and clinical transformer models for clinical concept extraction and medical relation extraction.

	Clinical concept extraction									**Medical relation extraction**		
	2010 i2b2[40]			**2012 i2b2**[41]			**2018 n2c2**[42]			**2018 n2c2**[42]		
Transformer	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
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	0.871	0.9678	0.9414	0.9544
BioMegatron	0.8614	0.8761	0.8687	0.7591	0.8031	0.7805	0.8707	0.8915	0.881	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**

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.

Table 2. Comparison of GatorTron with existing biomedical and clinical transformer models for semantic textual similarity, natural language inference, and question answering.

	Semantic textual similarity	**Natural language inference**	**Question answering**			
	2019 n2c2[43]	**MedNLI**[44]	**emrQA Medication**[45]		**emrQA Relation**[45]	
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**

The best evaluation scores are bolded.

Recognize clinical concepts and medical relations. Clinical concept extraction is a phrase-level task to identify the boundary of concepts with clinical meanings and classify their semantic categories (e.g., diseases, medications). As shown in Table 1, 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[40] and 2012 i2b2[41]: problem, treatments, lab tests; 2018 n2c2[42]: 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 medical relation extraction – a task to identify medical relations between two clinical concepts, 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 consistently observed performance improvement when scaling up the size of the GatorTron model.

Assess semantic textual similarity. The task of measuring semantic similarity is to determine the extent to which two sentences are similar in terms of semantic meaning. As shown in Table 2, all GatorTron models outperformed existing biomedical and clinical transformer models in assessing the semantic similarity between two sentences. 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 model significantly outperformed GatorTron-base and its performance is very close to the GatorTron-medium model (0.8896 vs. 0.8903).

Natural language inference. The task of NLI is to determine whether a conclusion can be inferred from a given sentence – a sentence-level NLP task. As shown in Table 2, all 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 the GatorTron model.

Medical question answering. MQA is a complex clinical NLP task that requires understand information from the entire document. As shown in Table 2, all GatorTron models outperformed existing biomedical and clinical transformer models in answering medication and relation-related questions (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 size of the GatorTron model.

Discussion

In this study, we developed the largest clinical transformer model, GatorTron, using a corpus of >90 billion words from UF Health (>82 billion), Pubmed (6 billion), Wikipedia (2.5 billion), and

MIMIC III (0.5 billion). We trained GatorTron with different model sizes including 345 million, 3.9 billion, and 8.9 billion parameters and evaluated its performance on 5 clinical NLP tasks with different complexity (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 evaluated using 6 different benchmark datasets. We observed monotonic improvements by scaling up the model size of GatorTron for 4 of the 5 tasks, excluding the semantic textual similarity task. Our GatorTron model also outperformed the BioMegatron[37], 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 from 345 million (ClinicalBERT) to 8.9 billion parameters in the clinical domain and demonstrated significant 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[34], which are more close to human intelligence than previous models. In the general domain, the Megatron-Turing NLG model has scaled up to 530 billion parameters following the GPT-3[35] 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 successfully scaled the transformer from 110 million 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 complex NLP tasks such as natural language inference and medical question answering, but moderate improvements for easier tasks such as clinical concept extraction and medical relation extraction, indicating that large transformer models are more helpful to complex clinical 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. Fig. 2 shows the training loss and validation loss for GatorTron models with three different settings. 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.

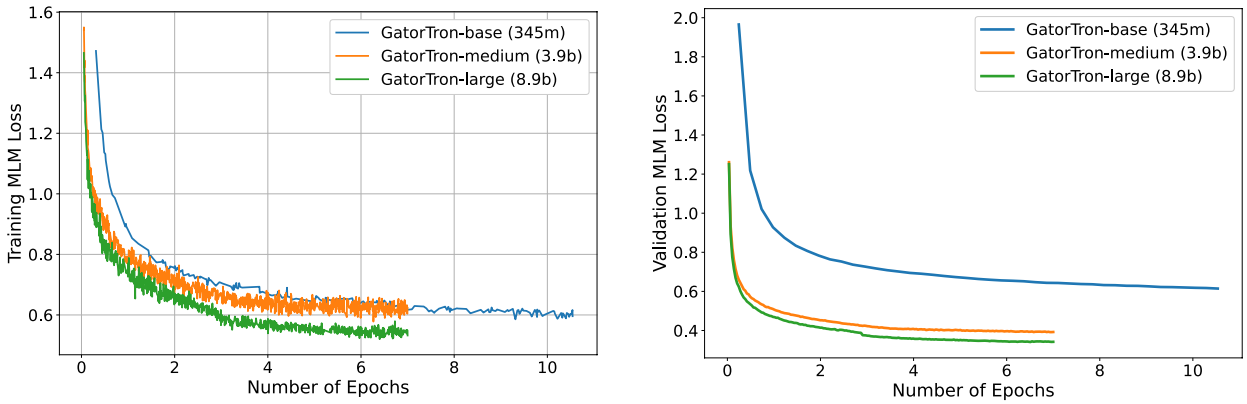

Fig. 2 Training loss and validation loss for GatorTron base (345 million), medium (3.9 billion), and large (8.9 billion) models.

Potentials in improving healthcare delivery and patient outcomes. GatorTron models perform better in understanding and utilizing patient information in clinical narratives, which can be applied to various medical AI systems to improve healthcare delivery and patient outcomes. The rich, fine-grained patients' information captured in clinical narratives is a critical resource powering medical AI systems. With better performance in information extraction tasks (e.g., clinical concept extraction and medical relation extraction), GatorTron models can 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 semantic textual similarity, natural language inference, and medical question answering can be applied for 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[34] 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.

Methods

Data Source

The primary data source for this study is the clinical narratives from UF Health IDR, a research data warehouse of UF Health. We collected a total of 290,482,002 clinical notes from 2011-2021 from over 2 million patients and 50 million encounters. After removing empty notes and duplicated notes, we generated a UFHealth clinical corpus with over 82 billion tokens. Then, we merged the UFHealth clinical corpus with three additional corpora, including the MIMIC-III

corpus[39] in the clinical domain with 0.5 billion words, a PubMed (combining PubMed abstracts and full-text commercial-collection) collection[37] in the biomedical domain with 6 billion words, and a Wikipedia articles dump[37] in the general domain with 2.5 billion words, to generate a corpus with >90 billion words. We performed minimal preprocessing including tokenization and sentence boundary detection.

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 model performance on 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 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 the three GatorTron models to 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) and BioMegatron (1.2 billion parameters). We examined the models on 5 clinical NLP tasks, including clinical concept extraction, relation extraction, semantic textual similarity, natural language inference, and medical question answering. We used 6 public benchmark datasets in the clinical domain.

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[46]. 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 with 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[47] as well as our previous experience in developing BioMegatron. Table 3 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[46] (with 8.3 billion parameters).

Table 3.

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.[12] 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.[38] 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[39] corpus.

BioMegatron.[37] 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 five clinical NLP tasks at different linguistic levels.

Clinical concept extraction. This is a task to recognize phrases with important clinical meanings (e.g., medications, treatments, adverse drug events). The NLP system has to determine 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)[48,49], convolutional neural networks (CNN)[10,50], and recurrent neural networks (RNN) implemented with long short-term memory strategy (LSTM)[11,51]. Current state-of-the-art models are based on transformers such as the ClinicalBERT. We used three benchmark datasets developed by the 2010 i2b2 challenge, 2012 i2b2 challenge, and 2018 n2c2 challenge to evaluate GatorTron models focusing on identifying important medical concepts (e.g., medications, adverse drug events, treatments) from clinical text. We used standard precision, recall, and F1-score for evaluation.

Medical Relation extraction. MRE is to establish medical-related relations (e.g., induce relation) among clinical concepts (e.g., drugs, adverse events), which is a crucial task for high-

level clinical NLP applications. 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)[42] have been applied. With the emergence of deep learning models, researchers have explored the long-short term memory (LSTM) architecture for RE in both general and clinical domains[52,53]. Most recently, several studies adopted the BERT architecture and demonstrated superior performance for MRE on various datasets[54–59]. In this study, 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.

Semantic textual similarity. 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[60] is widely used for evaluating STS systems[14]. Various machine learning methods have been examined[61–63] but transformer-based systems such as RoBERTa[26], T5[28], and ALBERT[29] are leading the state-of-the-art models for STS. In the clinical domain, the MedSTS dataset[64] that consists of over 1000 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[65] and 2019 n2c2/OHNLP ClinicalSTS shared task[43]. Similar to the general domain, pretrained transformer-based models using clinical text and biomedical literature, including ClinicalBERT and BioBERT[66], are state-of-the-art solutions. In this study, we used

the dataset developed by the 2019 n2c2/OHNLP challenge on clinical semantic textural similarity[43]. We used the Pearson correlation score for evaluation.

Natural language inference. NLI is also known as recognizing textual entailment (RTE) - a directional relation between text fragments (e.g., sentences)[67]. 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[68] and the Stanford NLI[69] are widely used. On both datasets, pretrained transformer models achieved state-of-the-art performances[28,30]. 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[44] was developed as a benchmark dataset in the clinical domain. A previous study[38] showed that a pretrained clinical BERT model achieved the state-of-the-art performance and outperformed the baseline (InferSent[70]) by ~9% accuracy. In this study, we evaluated the Gatortron models on NLI using the MedNLI dataset and used accuracy for comparison.

Medical question answering. The MQA task is to build NLP systems that automatically answer medical questions in a natural language, which is the most complex challenge among the 5 tasks. 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)[71,72] have been widely used as benchmarks. Transformer-based models are the state-of-the-art for both SQuAD1.1[19] and SQuAD2.0[32]. There are several MQA datasets developed in the past few years such as the MESHQA[73], MedQuAD[74], and emrQA[45]. In this study, we used the emrQA dataset, which is widely used as a benchmark dataset for MQA. We particularly focused on medications

and relations-related questions as Yue et al.[75] found that the two subsets are more consistent. We utilized both F1-score and exact match score for evaluation.

References and Notes:

1. Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015. Accessed December 20, 2019. /evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php
2. Adler-Milstein J, Holmgren AJ, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. *J Am Med Inform Assoc JAMIA*. 2017;24(6):1142-1148. doi:10.1093/jamia/ocx080
3. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. Published online 2008:128-144.
4. Liang H, Tsui BY, Ni H, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med*. 2019;25(3):433-438. doi:10.1038/s41591-018-0335-9
5. Yang J, Lian JW, Chin YP (Harvey), et al. Assessing the Prognostic Significance of Tumor-Infiltrating Lymphocytes in Patients With Melanoma Using Pathologic Features Identified by Natural Language Processing. *JAMA Netw Open*. 2021;4(9):e2126337. doi:10.1001/jamanetworkopen.2021.26337
6. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc JAMIA*. 2011;18(5):544-551. doi:10.1136/amiajnl-2011-000464
7. Fu S, Chen D, He H, et al. Clinical concept extraction: A methodology review. *J Biomed Inform*. 2020;109:103526. doi:10.1016/j.jbi.2020.103526
8. Bach N, Badaskar S. A Review of Relation Extraction. :16.
9. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539
10. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural Language Processing (Almost) from Scratch. *J Mach Learn Res*. 2011;12:2493-2537.
11. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural Architectures for Named Entity Recognition. *ArXiv160301360 Cs*. Published online March 4, 2016. Accessed March 2, 2018. http://arxiv.org/abs/1603.01360
12. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. Published online 2019. doi:10.1093/bioinformatics/btz682
13. Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. In: *Advances in Neural Information Processing Systems*. Vol 30. Curran Associates, Inc.; 2017. Accessed October 28, 2021. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
14. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *ArXiv180407461 Cs*. Published online February 22, 2019. Accessed March 21, 2020. http://arxiv.org/abs/1804.07461
15. Wang A, Pruksachatkun Y, Nangia N, et al. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *ArXiv190500537 Cs*. Published online February 12, 2020. Accessed May 8, 2021. http://arxiv.org/abs/1905.00537
16. Qiu X, Sun T, Xu Y, Shao Y, Dai N, Huang X. Pre-trained Models for Natural Language Processing: A Survey. *ArXiv200308271 Cs*. Published online April 24, 2020. Accessed May 8, 2021. http://arxiv.org/abs/2003.08271
17. Tay Y, Dehghani M, Bahri D, Metzler D. Efficient Transformers: A Survey. *ArXiv200906732 Cs*. Published online September 16, 2020. Accessed May 8, 2021. http://arxiv.org/abs/2009.06732
18. Yu J, Bohnet B, Poesio M. Named Entity Recognition as Dependency Parsing. *ArXiv200507150 Cs*. Published online June 13, 2020. Accessed October 29, 2021. http://arxiv.org/abs/2005.07150
19. Yamada I, Asai A, Shindo H, Takeda H, Matsumoto Y. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. *ArXiv201001057 Cs*. Published online October 2, 2020. Accessed October 29, 2021. http://arxiv.org/abs/2010.01057
20. Li X, Sun X, Meng Y, Liang J, Wu F, Li J. Dice Loss for Data-imbalanced NLP Tasks. *ArXiv191102855 Cs*. Published online August 29, 2020. Accessed October 29, 2021. http://arxiv.org/abs/1911.02855

21. Xu B, Wang Q, Lyu Y, Zhu Y, Mao Z. Entity Structure Within and Throughout: Modeling Mention Dependencies for Document-Level Relation Extraction. *ArXiv210210249 Cs*. Published online February 19, 2021. Accessed October 29, 2021. http://arxiv.org/abs/2102.10249
22. Ye D, Lin Y, Sun M. Pack Together: Entity and Relation Extraction with Levitated Marker. *ArXiv210906067 Cs*. Published online October 10, 2021. Accessed October 29, 2021. http://arxiv.org/abs/2109.06067
23. Cohen AD, Rosenman S, Goldberg Y. Relation Classification as Two-way Span-Prediction. *ArXiv201004829 Cs*. Published online April 17, 2021. Accessed October 29, 2021. http://arxiv.org/abs/2010.04829
24. Lyu S, Chen H. Relation Classification with Entity Type Restriction. *ArXiv210508393 Cs*. Published online May 18, 2021. Accessed October 29, 2021. http://arxiv.org/abs/2105.08393
25. Wang J, Lu W. Two are Better than One: Joint Entity and Relation Extraction with Table-Sequence Encoders. *ArXiv201003851 Cs*. Published online October 8, 2020. Accessed October 29, 2021. http://arxiv.org/abs/2010.03851
26. Jiang H, He P, Chen W, Liu X, Gao J, Zhao T. SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization. *Proc 58th Annu Meet Assoc Comput Linguist*. Published online 2020:2177-2190. doi:10.18653/v1/2020.acl-main.197
27. Yang Z, Dai Z, Yang Y, Carbonell JG, Salakhutdinov R, Le QV. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In: *NeurIPS*. ; 2019.
28. Raffel C, Shazeer N, Roberts A, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *ArXiv191010683 Cs Stat*. Published online October 24, 2019. Accessed March 27, 2020. http://arxiv.org/abs/1910.10683
29. Lan ZZ, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *ArXiv*. 2019;abs/1909.11942.
30. Wang S, Fang H, Khabsa M, Mao H, Ma H. Entailment as Few-Shot Learner. *ArXiv210414690 Cs*. Published online April 29, 2021. Accessed October 29, 2021. http://arxiv.org/abs/2104.14690
31. Zhang Z, Wu Y, Zhao H, et al. Semantics-aware BERT for Language Understanding. *ArXiv190902209 Cs*. Published online February 4, 2020. Accessed October 29, 2021. http://arxiv.org/abs/1909.02209
32. Zhang Z, Yang J, Zhao H. Retrospective Reader for Machine Reading Comprehension. *ArXiv200109694 Cs*. Published online December 11, 2020. Accessed October 29, 2021. http://arxiv.org/abs/2001.09694
33. Garg S, Vu T, Moschitti A. TANDA: Transfer and Adapt Pre-Trained Transformer Models for Answer Sentence Selection. *ArXiv191104118 Cs*. Published online November 20, 2019. Accessed October 29, 2021. http://arxiv.org/abs/1911.04118
34. Bommasani R, Hudson DA, Adeli E, et al. On the Opportunities and Risks of Foundation Models. *ArXiv210807258 Cs*. Published online August 18, 2021. Accessed October 16, 2021. http://arxiv.org/abs/2108.07258
35. Floridi L, Chiriatti M. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds Mach*. 2020;30(4):681-694. doi:10.1007/s11023-020-09548-1
36. Gu Y, Tinn R, Cheng H, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans Comput Healthc*. 2022;3(1):1-23. doi:10.1145/3458754
37. Shin HC, Zhang Y, Bakhturina E, et al. BioMegatron: Larger Biomedical Domain Language Model. *ArXiv201006060 Cs*. Published online October 13, 2020. Accessed April 5, 2021. http://arxiv.org/abs/2010.06060
38. Alsentzer E, Murphy J, Boag W, et al. Publicly Available Clinical BERT Embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2019:72-78. doi:10.18653/v1/W19-1909
39. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035. doi:10.1038/sdata.2016.35
40. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc JAMIA*. 2011;18(5):552-556. doi:10.1136/amiajnl-2011-000203
41. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J Am Med Inform Assoc JAMIA*. 2013;20(5):806-813. doi:10.1136/amiajnl-2013-001628
42. Yang X, Bian J, Fang R, Bjarnadottir RI, Hogan WR, Wu Y. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. *J Am Med Inform Assoc*. 2020;27(1):65-72.
43. Wang Y, Fu S, Shen F, Henry S, Uzuner O, Liu H. Overview of the 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity. *JMIR Med Inform*. Published online 2020.

44. Shivade C. MedNLI — A Natural Language Inference Dataset For The Clinical Domain. Published online 2017. doi:10.13026/C2RS98
45. Pampari A, Raghavan P, Liang J, Peng J. emrQA: A Large Corpus for Question Answering on Electronic Medical Records. *ArXiv180900732 Cs*. Published online September 3, 2018. Accessed October 24, 2021. http://arxiv.org/abs/1809.00732
46. Shoeybi M, Patwary M, Puri R, LeGresley P, Casper J, Catanzaro B. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. *ArXiv190908053 Cs*. Published online March 13, 2020. Accessed October 20, 2021. http://arxiv.org/abs/1909.08053
47. Levine Y, Wies N, Sharir O, Bata H, Shashua A. The Depth-to-Width Interplay in Self-Attention. *ArXiv200612467 Cs Stat*. Published online January 17, 2021. Accessed October 23, 2021. http://arxiv.org/abs/2006.12467
48. Wu Y, Xu J, Jiang M, Zhang Y, Xu H. A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text. *AMIA Annu Symp Proc*. 2015;2015:1326-1333.
49. Soysal E, Wang J, Jiang M, et al. CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc*. 2018;25(3):331-336. doi:10.1093/jamia/ocx132
50. Wu Y, Jiang M, Lei J, Xu H. Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network. *Stud Health Technol Inform*. 2015;216:624-628.
51. Wu Y, Yang X, Bian J, Guo Y, Xu H, Hogan W. Combine factual medical knowledge and distributed word representation to improve clinical named entity recognition. In: *AMIA Annual Symposium Proceedings*. Vol 2018. American Medical Informatics Association; 2018:1110.
52. Kumar S. A Survey of Deep Learning Methods for Relation Extraction. *ArXiv170503645 Cs*. Published online May 10, 2017. Accessed May 8, 2021. http://arxiv.org/abs/1705.03645
53. Lv X, Guan Y, Yang J, Wu J. Clinical relation extraction with deep learning. *Int J Hybrid Inf Technol*. 2016;9(7):237-248.
54. Wei Q, Ji Z, Si Y, et al. Relation Extraction from Clinical Narratives Using Pre-trained Language Models. *AMIA Annu Symp Proc*. 2020;2019:1236-1245.
55. Guan H, Devarakonda M. Leveraging Contextual Information in Extracting Long Distance Relations from Clinical Notes. *AMIA Annu Symp Proc*. 2020;2019:1051-1060.
56. Alimova I, Tutubalina E. Multiple features for clinical relation extraction: A machine learning approach. *J Biomed Inform*. 2020;103:103382. doi:10.1016/j.jbi.2020.103382
57. Mahendran D, McInnes BT. Extracting Adverse Drug Events from Clinical Notes. *ArXiv210410791 Cs*. Published online April 21, 2021. Accessed May 26, 2021. http://arxiv.org/abs/2104.10791
58. Yang X, Zhang H, He X, Bian J, Wu Y. Extracting Family History of Patients From Clinical Narratives: Exploring an End-to-End Solution With Deep Learning Models. *JMIR Med Inform*. 2020;8(12):e22982.
59. Yang X, Yu Z, Guo Y, Bian J, Wu Y. Clinical Relation Extraction Using Transformer-based Models. *ArXiv Prepr ArXiv210708957*. Published online 2021.
60. Cer D, Diab M, Agirre E, Lopez-Gazpio I, Specia L. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *ArXiv Prepr ArXiv170800055*. Published online 2017.
61. Farouk M. Measuring Sentences Similarity: A Survey. *ArXiv*. Published online 2019. doi:10.17485/ijst/2019/v12i25/143977
62. Ramaprabha J, Das S, Mukerjee P. Survey on Sentence Similarity Evaluation using Deep Learning. *J Phys Conf Ser*. 2018;1000:012070. doi:10.1088/1742-6596/1000/1/012070
63. Gomaa WH, Fahmy A. A Survey of Text Similarity Approaches. Published online 2013. doi:10.5120/11638-7118
64. Wang Y, Afzal N, Fu S, et al. MedSTS: a resource for clinical semantic textual similarity. *Lang Resour Eval*. 2020;54(1):57-72. doi:10.1007/s10579-018-9431-1
65. 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*. BCB '18. ACM; 2018:575-575. doi:10.1145/3233547.3233672
66. Mahajan D, Poddar A, Liang JJ, et al. Identification of Semantically Similar Sentences in Clinical Notes: Iterative Intermediate Training Using Multi-Task Learning. *JMIR Med Inform*. 2020;8(11):e22508. doi:10.2196/22508
67. Dagan I, Glickman O, Magnini B. The PASCAL Recognising Textual Entailment Challenge. In: Quiñonero-Candela J, Dagan I, Magnini B, d'Alché-Buc F, eds. *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*. Springer Berlin Heidelberg; 2006:177-190.

68. Williams A, Nangia N, Bowman SR. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *ArXiv170405426 Cs*. Published online February 19, 2018. Accessed August 16, 2020. http://arxiv.org/abs/1704.05426
69. Bowman SR, Angeli G, Potts C, Manning CD. A large annotated corpus for learning natural language inference. *ArXiv150805326 Cs*. Published online August 21, 2015. Accessed November 8, 2021. http://arxiv.org/abs/1508.05326
70. Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *ArXiv170502364 Cs*. Published online July 8, 2018. Accessed September 6, 2021. http://arxiv.org/abs/1705.02364
71. Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *ArXiv160605250 Cs*. Published online October 10, 2016. Accessed August 16, 2020. http://arxiv.org/abs/1606.05250
72. Rajpurkar P, Jia R, Liang P. Know What You Don't Know: Unanswerable Questions for SQuAD. *ArXiv180603822 Cs*. Published online June 11, 2018. Accessed November 8, 2021. http://arxiv.org/abs/1806.03822
73. Zhu M, Ahuja A, Juan DC, Wei W, Reddy CK. Question Answering with Long Multiple-Span Answers. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics; 2020:3840-3849. doi:10.18653/v1/2020.findings-emnlp.342
74. Ben Abacha A, Demner-Fushman D. A question-entailment approach to question answering. *BMC Bioinformatics*. 2019;20(1):511. doi:10.1186/s12859-019-3119-4
75. Yue X, Gutierrez BJ, Sun H. Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset. *ArXiv200500574 Cs*. Published online May 1, 2020. Accessed September 6, 2021. http://arxiv.org/abs/2005.00574

Acknowledgments:

Funding: 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 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.

Author contributions: YW, JB, MGF, NP, and XY were responsible for the overall design, development, and evaluation of this study. XY had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. YW, XY 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.

Competing interests: Authors have no competing financial interests.

Materials & Correspondence: Yonghui Wu, PhD

Statistical information: N/A

Data availability: The benchmark datasets that support the findings of this study are available from the official websites of natural language processing challenges.

Computer code: The computer codes to train GatroTron models are available from: https://github.com/NVIDIA/Megatron-LM and https://github.com/NVIDIA/NeMo.