ARTICLE OPEN

# A large language model for electronic health records

Xi Yang[1,2], Aokun Chen[1,2], Nima PourNejatian[3], Hoo Chang Shin[3], Kaleb E. Smith[3], Christopher Parisien[3], Colin Compas[3], Cheryl Martin[3], Anthony B. Costa[3], Mona G. Flores[3], Ying Zhang[4], Tanja Magoc[5], Christopher A. Harle[1,5], Gloria Lipori[5,6], Duane A. Mitchell[6], William R. Hogan[1], Elizabeth A. Shenkman[1], Jiang Bian[1,2] and Yonghui Wu[1,2] ✉

There is an increasing interest in developing artificial intelligence (AI) systems to process and interpret electronic health records (EHRs). Natural language processing (NLP) powered by pretrained language models is the key technology for medical AI systems utilizing clinical narratives. However, there are few clinical language models, the largest of which trained in the clinical domain is comparatively small at 110 million parameters (compared with billions of parameters in the general domain). It is not clear how large clinical language models with billions of parameters can help medical AI systems utilize unstructured EHRs. In this study, we develop from scratch a large clinical language model—GatorTron—using >90 billion words of text (including >82 billion words of de-identified clinical text) and systematically evaluate it on five clinical NLP tasks including clinical concept extraction, medical relation extraction, semantic textual similarity, natural language inference (NLI), and medical question answering (MQA). We examine how (1) scaling up the number of parameters and (2) scaling up the size of the training data could benefit these NLP tasks. GatorTron models scale up the clinical language model from 110 million to 8.9 billion parameters and improve five clinical NLP tasks (e.g., 9.6% and 9.5% improvement in accuracy for NLI and MQA), which can be applied to medical AI systems to improve healthcare delivery. The GatorTron models are publicly available at: https://catalog.ngc.nvidia.com/orgs/nvidia/teams/clara/models/gatortron_og.

*npj Digital Medicine* (2022)5:194 ; https://doi.org/10.1038/s41746-022-00742-2

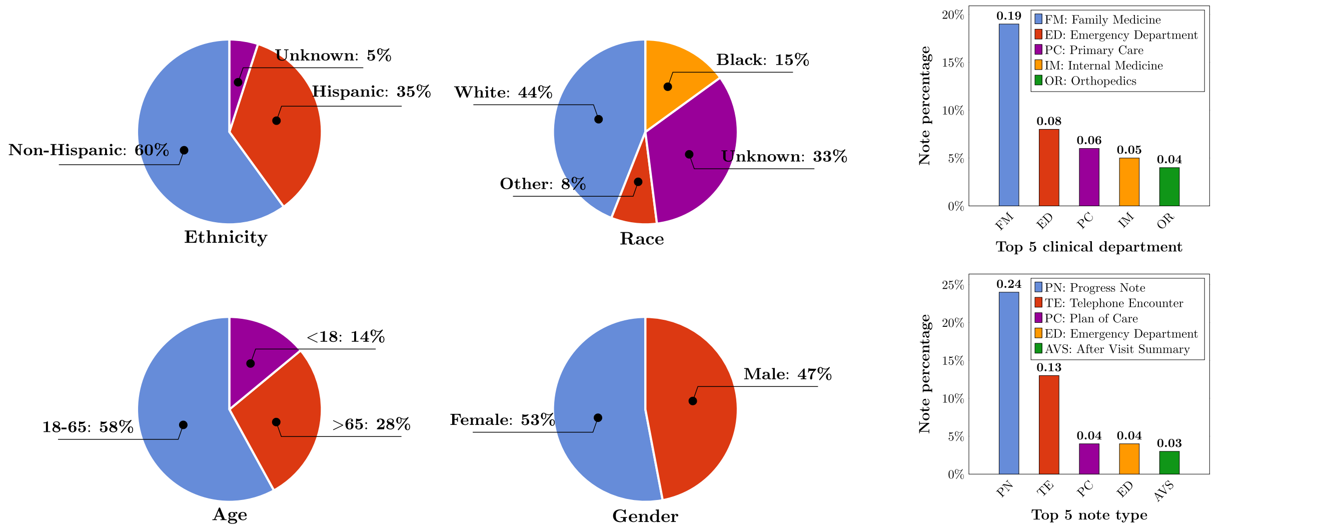
## INTRODUCTION

There is an increasing interest in developing artificial intelligence (AI) systems to improve healthcare delivery and health outcomes using electronic health records (EHRs). A critical step is to extract and capture patients' characteristics from longitudinal EHRs. The more information we have about the patients, the better the medical AI systems that we can develop. In recent decades, hospitals and medical practices in the United States (US) have 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 progress notes). Even though using discrete data fields in clinical documentation has many potential advantages and structured data entry fields are increasingly added into the EHR systems, having clinicians use them remains a barrier, due to the added documentation burden[3]. Physicians and other healthcare providers widely use clinical narratives as a more convenient way to document patient information ranging from family medical histories to social determinants of health[4]. There is an increasing number of medical AI systems exploring the rich, more fine-grained patient information captured in clinical narratives to improve diagnostic and prognostic models[5,6]. 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 medical AI systems to understand clinical language used in healthcare[7].

Today, most NLP solutions are based on deep learning models[8] implemented using neural network architectures—a fast-developing sub-domain of machine learning. Convolutional neural networks[9] (CNN) and recurrent neural networks[10] (RNN) have been applied to NLP in the early stage of deep learning. More recently, the transformer architectures[11] (e.g., Bidirectional Encoder Representations from Transformers [BERT]) implemented with a self-attention mechanism[12] have become state-of-the-art, achieving the best performance on many NLP benchmarks[13–16]. In the general domain, the transformer-based NLP models have achieved state-of-the-art performance for name entity recognition[17–19], relation extraction[20–24], sentence similarity[25–27], natural language inference[27–30], and question answering[27,28,31,32]. Typically, transformers are trained in two stages: language model pretraining (i.e., learning using a self-supervised training objective on a large corpus of unlabeled text) and fine-tuning (i.e., applying the learned language models solving specific tasks with labeled training data). One pretrained language model can be applied to solve many NLP tasks through fine-tuning, which is known as transfer learning—a strategy to learn knowledge from one task and apply it in another task[33]. Human language has a very large sample space—the possible combinations of words, sentences, and their meaning and syntax are innumerable. Recent studies show that large transformer models trained using massive text data are remarkably better than previous NLP models in terms of emergence and homogenization[33].

The promise of transformer models has led to further interest in exploring large-size (e.g., >billions of parameters) transformer models. The Generative Pretrained Transformer 3 (GPT-3) model[34], which has 175 billion parameters and was trained using >400 billion words of text demonstrated superior performance. In the biomedical domain, researchers developed BioBERT[11] (with 110 million parameters) and PubMedBERT[35] (110 million parameters)

[1]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, USA. [2]Cancer Informatics and eHealth core, University of Florida Health Cancer Center, Gainesville, FL, USA. [3]NVIDIA, Santa Clara, CA, USA. [4]Research Computing, University of Florida, Gainesville, FL, USA. [5]Integrated Data Repository Research Services, University of Florida, Gainesville, FL, USA. [6]Lillian S. Wells Department of Neurosurgery, UF Clinical and Translational Science Institute, University of Florida, Gainesville, FL, USA. ✉email: yonghui.wu@ufl.edu



Patient distribution by age, gender, race, ethnicity

Note distribution by type and clinical department

**Fig. 1 Patient distribution by age, gender, race, ethnicity; clinical notes distribution by note type, and clinical department.** Ages were calculated as of September 2022.

transformer models using biomedical literature from PubMed. NVIDIA developed BioMegatron models in the biomedical domain with different sizes from 345 million to 1.2 billion parameters[36] using a more expansive set of PubMed-derived free text. However, few studies have explored scaling transformer models in the clinical domain due to the sensitive nature of clinical narratives that contain Protected Health Information (PHI) and the significant computing power required to increase the size of these models. To date, the largest transformer model using clinical narratives is ClinicalBERT[37]. ClinicalBERT has 110 million parameters and was trained using 0.5 billion words from the publicly available Medical Information Mart for Intensive Care III[38] (MIMIC-III) dataset. By developing not only larger models, but models that use clinical narratives, NLP may perform better to improve healthcare delivery and patient outcomes.

In this study, we develop a large clinical language model, GatorTron, using >90 billion words of text from the de-identified clinical notes of University of Florida (UF) Health, PubMed articles, and Wikipedia. We train GatorTron from scratch and empirically evaluate how scaling up the number of parameters benefit the performance of downstream NLP tasks. More specifically, we examine GatorTron models with varying number of parameters 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 also examine how scaling up data size benefit downstream tasks by comparing the GatorTron-base model trained from the full corpus with another GatorTron-base model trained using a random sample of 1/4 of the corpus. We compare GatorTron with existing transformer models trained using biomedical literature and clinical narratives using five 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 models outperform previous transformer models from the biomedical and clinical domain on five clinical NLP tasks. This study scales up transformer models in the clinical domain from 110 million to 8.9 billion parameters and demonstrates the benefit of large transformer models.

## RESULTS

A total number of 290,482,002 clinical notes from 2,476,628 patients were extracted from the UF Health Integrated Data Repository (IDR), the enterprise data warehouse of the UF Health system. These notes were created from 2011–2021 from over 126 clinical departments and ~50 million encounters covering healthcare settings including but not limited to inpatient, outpatient, and emergency department visits. After preprocessing and de-identification, the corpus included >82 billion medical words. Figure 1 summarizes the distribution of patient by age, gender, race, and ethnicity as well as the distribution of notes by clinical department (top 5) and note type (top 5). The detailed number of patients by each category, a full list of clinical departments and the corresponding proportion of notes, and a full list of note types were provided in Supplementary Table 1, Supplementary Table 2, and Supplementary Table 3.

Training GatorTron-large model required ~6 days on 992 A100 80 G GPUs from 124 NVIDIA DGX notes using the NVIDIA SuperPOD reference cluster architecture. Figure 2 shows the training validation loss for all three sizes of GatorTron models. The GatorTron-base model converged in 10 epochs, whereas the medium and large models converged in 7 epochs, which is consistent with prior observations on the faster per sample convergence of larger transformer models.

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 five clinical NLP tasks.

### Scale up the size of training data and the number of parameters

Compared with GatorTron-base trained using a random sample of 1/4 of the corpus, the GatorTron-base model trained using the full corpus achieved improved performance for four tasks except for a sub-task in MQA (on F1 score of medication-related questions). By scaling up the number of parameters from 345 million to 8.9 billion, GatorTron-large demonstrated remarkable improvements for all five tasks, suggesting that GatorTron models scale for canonical clinical downstream tasks and that we are not yet at the limit.

### Recognize clinical concepts and medical relations

Clinical concept extraction is to identify the concepts with important clinical meanings and classify their semantic categories (e.g., diseases, medications). As shown in Table 1, all three

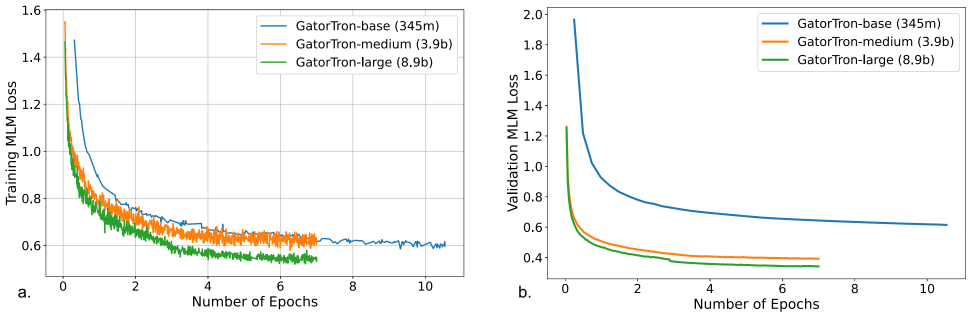


**Fig. 2 Training loss and validation loss for GatorTron-base (345 million), medium (3.9 billion), and large (8.9 billion) models. a** Training loss. **b** Validation loss. MLM masked language modeling.

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

| Transformer | Clinical concept extraction | | | | | | | | | Medical relation extraction | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2010 i2b2[39] | | | 2012 i2b2[40] | | | 2018 n2c2[41] | | | 2018 n2c2[41] | | |
| | 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 | 0.8710 | 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 (1/4 data) | 0.8682 | 0.9046 | 0.8860 | 0.7514 | 0.8013 | 0.7755 | 0.8772 | 0.8992 | 0.8881 | 0.9724 | 0.9457 | 0.9589 |
| 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 presented in bold. 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.

| Transformer | Semantic textual similarity | Natural language inference | Question answering | | | |
|---|---|---|---|---|---|---|
| | 2019 n2c2[66] | MedNLI[71] | emrQA medication[77] | | emrQA relation[77] | |
| | 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 (1/4 data) | 0.8675 | 0.8643 | 0.7281 | 0.2952 | 0.9390 | 0.8579 |
| 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 presented in bold.

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[39] and 2012 i2b2[40]: problem, treatments, lab tests; 2018 n2c2[41]: 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. 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. 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 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 a large 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 number of parameters including 345 million, 3.9 billion, and 8.9 billion 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. The experimental results show that GatorTron models outperformed existing biomedical and clinical transformers for all five clinical NLP tasks evaluated using six different benchmark datasets. We observed monotonic improvements by scaling up the model size of GatorTron for four of the five tasks, excluding the semantic textual similarity task. Our GatorTron model also outperformed the BioMegatron[36], 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 remarkable performance improvements. To the best of our knowledge, GatorTron-large is the largest transformer model in the clinical domain. Among the five tasks, GatorTron achieved remarkable 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 NLP tasks. These results are consistent with observations in the literature on the saturation of simpler benchmarks with large BERT architectures[18,32].

GatorTron was pretrained using self-supervised masked language modeling (MLM) objective. We monitored training loss and calculated validation loss using a subset set of the clinical text (5%) to determine the appropriate stopping time. From the plots of training and validation losses in Fig. 2, we observed that larger GatorTron models converged faster than the smaller model.

GatorTron models perform better in extracting and interpreting patient information documented in clinical narratives, which can be integrated into medical AI systems to improve healthcare delivery and patient outcomes. The rich, fine-grained patient information captured in clinical narratives is a critical resource powering medical AI systems. With better performance in information extraction (e.g., clinical concept extraction and medical relation extraction), GatorTron models can provide more accurate patient 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 for pharmacovigilance. The observed 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.

We conducted error analysis and compared GatorTron with ClinicalBERT to probe the observed performance improvements. We found that the larger, domain-specific pretrained models (e.g., GatorTron) are better at modeling longer phrases and determining semantic categories. For example, GatorTron successfully identified "*a mildly dilated ascending aorta*", where ClinicalBERT identified only "mildly dilated" as a problem; GatorTron successfully categorized "kidney protective effects" as a "TREATMENT", which was mis-classified as "PROBLEM" by ClinicalBERT. For complex NLP tasks such as NLI and MQA, even large language models such as GatorTron still have difficulty in identifying the key pieces of information from longer paragraphs. Our future work will improve GatorTron in handling long pieces of text for complex NLP tasks.

This study demonstrates the advantages of large pretrained transformer models in the medical domain. GatorTron models can be applied to many other NLP 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. This study was approved by the UF Institutional Review Board (IRB202100049). We collected clinical notes from 2011–2021 from over 126 departments, ~2 million patients and 50 million encounters from inpatient, outpatient, and emergency settings. Then, we merged the UF Health clinical corpus with three additional corpora, including the MIMIC-III corpus[38] in the clinical domain with 0.5 billion words, a PubMed (combining PubMed abstracts and full-text commercial-collection) collection[36] in the biomedical domain with 6 billion words, and a Wikipedia articles dump[36] in the general domain with 2.5 billion words, to generate a corpus with >90 billion words.

### Preprocessing and de-identification of text

We performed minimal preprocessing including (1) removing empty and duplicated clinical notes, unifying all text into UTF-8 encoding, and removing illegal UTF-8 strings; (2) normalizing special characters (e.g., convert '&' to '&;' '\xa0' to 'space'); (3)

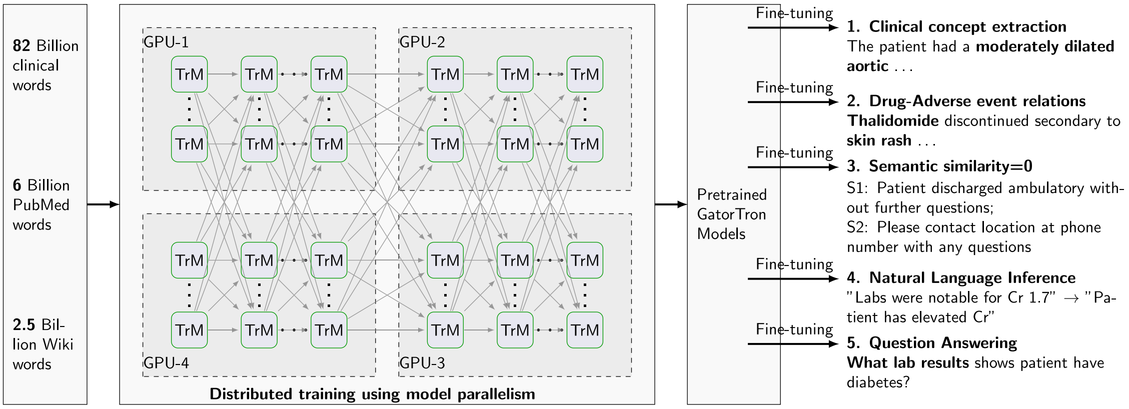

82 Billion clinical words
6 Billion PubMed words
2.5 Billion Wiki words
GPU-1
GPU-2
GPU-3
GPU-4
TrM
Distributed training using model parallelism
Pretrained GatorTron Models
Fine-tuning
1. Clinical concept extraction
The patient had a moderately dilated aortic . . .
2. Drug-Adverse event relations
Thalidomide discontinued secondary to skin rash . . .
3. Semantic similarity=0
S1: Patient discharged ambulatory without further questions;
S2: Please contact location at phone number with any questions
4. Natural Language Inference
"Labs were notable for Cr 1.7" → "Patient has elevated Cr"
5. Question Answering
What lab results shows patient have diabetes?

**Fig. 3 An overview of pretraining and fine-tuning of GatorTron models.** We loaded the base model and the medium model into one GPU for distributed training. We sliced the GatorTron-large model into 4 pieces and loaded model pieces to 4 GPUs for distributed training (i.e., model parallelism). TrM transformer unit.

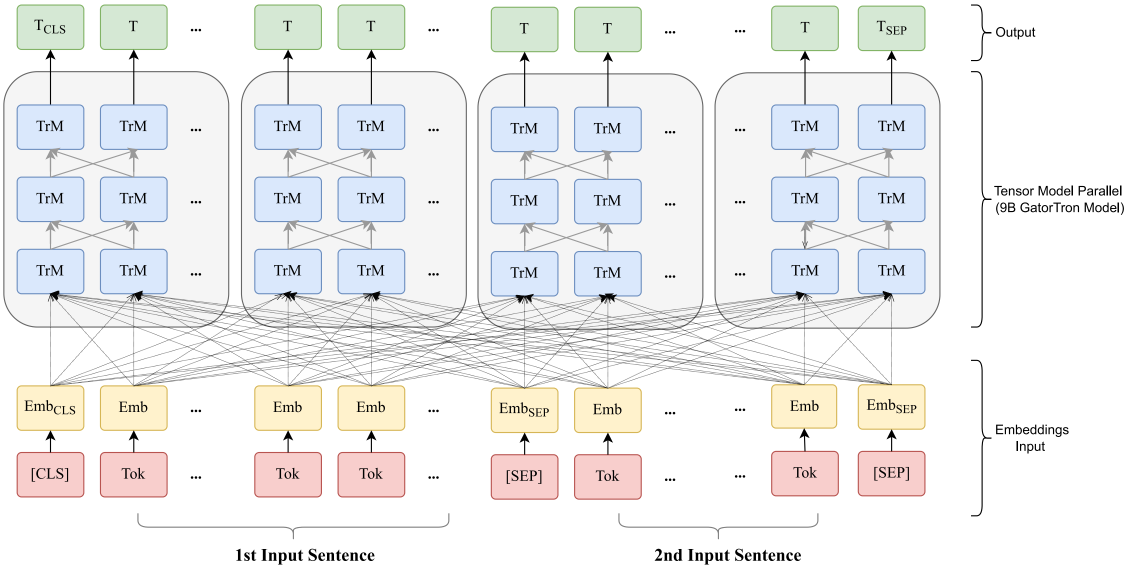


**Fig. 4 Pretraining GatorTron-large model with 9 billion parameters using model parallelism.** Emb embedding, Tok Token from input sentence, Trm Transformer unit. [SEP]: a token defined in BERT to indicate sentence boundaries. [CLS]: a token defined in BERT for sentence-level representation.

tokenization and sentence boundary detection. For clinical text from UF Health, we further applied a de-identification system[42] to remove protected health information (PHI) from clinical text. (Approved under IRB202100049) We adopted the safe-harbor method to identify 18 PHI categories defined in the Health Insurance Portability and Accountability Act (HIPAA) and replaced them with dummy strings (e.g., replace people's names into [**NAME**]).

## Study design

Figure 3 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 model size improves performance on five clinical NLP tasks. We first pretrained GatorTron using the >90 billion words by optimizing a masked language model (MLM) and then applied GatorTron to five different clinical NLP tasks using a supervised fine-tuning. We adopted the BERT architecture (Fig. 4) implemented in Megatron-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 compared the models on five clinical NLP tasks, including clinical concept extraction, relation extraction, semantic textual similarity, natural language inference, and medical question answering. We used six 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[43]. 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).

**Table 3.** Technical details of GatorTron models.

| 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 |

## 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.[44] 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[43] (with 8.3 billion parameters).

## Train GatorTron models from scratch

We pretrained a vocabulary from scratch using >90 billion words of corpus following the byte-pair-encoding algorithm[45]. We inherited the BERT-style architecture and trained GatorTron models from scratch using two self-supervised tasks, including masked language modeling (MLM) and sentence-order prediction (SOP). We followed the similar strategy in the BERT model[46] to randomly mask 15% of the input tokens with a special token (i.e., [MASK]) in the MLM. The SOP was formulated as a task to predict the order of two consecutive segments of text[28]. The input for SOP consists of two consecutive sentences from the training corpus in random orders and the training objective is to determine whether the two input sentences are in the correct order. The GatorTron-large model with 8.9 billion parameters is too large to fit one GPU, therefore, we sliced it into four pieces for distributed training using model parallelism. We pretrained the GatorTron-base and medium model without model slicing. The default loss function defined in BERT model[46] was used. Figure 4 shows the distributed training of GatorTron-large model using model parallelism. (See https://github.com/NVIDIA/Megatron-LM for more details)

## Existing transformer models for comparison

BioBERT[11]: 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[37]: 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[38] corpus.

BioMegatron[36]: 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.

## Fine-tune GatorTron for five clinical NLP tasks, evaluation matrices, and benchmark datasets

We fine-tuned pretrained GatorTron models for five different clinical NLP tasks using experts' annotations from six public benchmark datasets. Specifically, we first generated distributed representation from the inputs of a specific task, then added additional output layers (classification or regression) to generate target outputs. We used cross-entropy (CE) loss for classification tasks and mean square error loss for regression tasks. For a classification task with $N$ categories, let $C_i$ be the score generated by a transformer model for category $i$, the probability $_{Pi}$ of a given sample be classified to category $i$ was calculated as:

$$P_i = \frac{e^{C_i}}{\sum_{j=1}^{N} e^{C_j}} \tag{1}$$

Let $t_i$ be the ground truth category, the cross-entropy loss $L_{CE}$ is defined as:

$$L_{CE} = -\sum_{i=1}^{N} t_i \log(P_i) \tag{2}$$

*Fine-tune GatorTron for clinical concept extraction.* This is a task to recognize phrases with important clinical meanings (e.g., medications, treatments, adverse drug events). The task is 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)[47,48], convolutional neural networks (CNN)[9,49], and recurrent neural networks (RNN) implemented with long-short-term memory strategy (LSTM)[10,50]. Current state-of-the-art models are based on transformers such as the ClinicalBERT. We approached clinical concept extraction as a sequence labeling problem and adopted 'BIO' labeling schema, where 'B-' and 'I-' are prefixes indicating words at the beginning and inside of a concept, and 'O' stands for words located outside of any concepts of interest. Using this definition, we approached the task as a classification problem—for each word in a sentence, predict a label in ['B', 'I', 'O']. When there are multiple categories of concepts, a suffix was attached to 'BIO' for discrimination (e.g., 'B-drug', 'I-drug'). Based on the representation generated by pretrained GatorTron models, we added a classification layer (a linear layer with softmax activation) to calculate a probability score for each 'BIO' category. The cross-entropy loss was used for fine-tuning. We trained a unified classifier to extract all concepts for datasets without overlapped concepts. For datasets with overlapped concepts, we trained individual models to recognize each category of concept separately following our previous strategy[51]. We used three benchmark datasets developed by the 2010 i2b2 challenge[39], 2012 i2b2 challenge[40], and 2018 n2c2 challenge[41] to evaluate GatorTron models focusing on identifying important medical concepts (e.g., medications, adverse drug events, treatments) from clinical text. We used precision, recall, and F1 score for evaluation.

*Fine-tune GatorTron for medical relation extraction.* MRE is to establish medical-related relations (e.g., induce relation) among clinical concepts (e.g., drugs, adverse events). MRE is usually approached as a classification problem—identify pairs of concepts with valid relations and classify the relation type. Various machine learning-based classifiers such as support vector machines (SVMs), random forests (RF), and gradient boosting trees (GBT)[41] 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]. We approached MRE as a classification task. First, candidate concept pairs were generated using heuristic rules developed in our previous study[41]. Then, we identified two sentences where the two concepts in a pair were located. We introduced two sets of entity markers (i.e., [S1], [E1] and [S2], [E2]) to indicate the two concepts. If the two concepts were in the same sentence, the two input sentences will be the same but labeled with different markers (e.g., [S1] and [E1] were used in the first sentence; [S2] and [E2] were used in the second sentence). To determine the relation type, we concatenated the representations of the model special [CLS] token and all four entity markers and added a classification layer (a linear layer with softmax activation) for classification. Similarly, the cross-entropy loss was used to fine-tune GatorTron. We used the dataset developed by the 2018 n2c2 challenge[41] with a focus on relations between medications and adverse drug events. The precision, recall, and F1 score were used for evaluation.

*Fine-tune GatorTron for 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[13]. Various machine learning methods have been examined[61–63] but transformer-based systems such as RoBERTa[25], T5[27], and ALBERT[28] 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[66]. Similar to the general domain, pretrained transformer-based models using clinical text and biomedical literature, including ClinicalBERT and BioBERT[67], achieved state-of-the-art performance. In this study, we formulated STS as a regression problem. We applied pretrained GatorTron models to learn the sentence-level representations of the two pieces of text and adopted a linear regression layer to calculate the similarity score. Different from classification models, we used MSE as the loss function. We used the dataset developed by the 2019 n2c2/OHNLP[66] challenge on clinical semantic textual similarity[66]. The Pearson correlation score was used for evaluation.

*Fine-tune GatorTron for natural language inference*. NLI is also known as recognizing textual entailment (RTE)—a directional relation between text fragments (e.g., sentences)[68]. 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[69] and the Stanford NLI[70] are widely used. On both datasets, pretrained transformer models achieved state-of-the-art performances[27,29]. 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[71] was developed as a benchmark dataset in the clinical domain. A previous study[37] showed that a pretrained clinical BERT model achieved the state-of-the-art performance and outperformed the baseline (InferSent[72]) by ~9% accuracy. In this study, we approached NLI as a classification problem. We concatenated the hypothesis and premise as the input separated using a special token [SEP] and applied pretrained GatorTron models to generate distributed representations, which were fed into a classification layer (a linear layer with softmax activation) to calculate a probability for each of the three categories of entailment, contradiction, and neutral. The cross-entropy loss was used for fine-tuning. We evaluated the GatorTron models on NLI using the MedNLI dataset[71] and used accuracy for comparison.

*Fine-Tune GatorTron for 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 five 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)[73,74] have been widely used as benchmarks. Transformer-based models are state-of-the-art for both SQuAD1.1[18] and SQuAD2.0[31]. There are several MQA datasets developed in the past few years such as the MESHQA[75], MedQuAD[76], and emrQA[77]. In this study, we approached MQA using a machine reading comprehension (MRC) technique where the goal is to extract the most relevant responses (i.e., short text snippets or entities) from the given context according to questions. We applied a span classification algorithm to identify the start and end offsets of the answer from the context. More specifically, we packed the question and the context into a single sequence as input for GatorTron and applied two linear layers to predict the start and end position of the answer, respectively. As GatorTron models were developed using a maximum token length of 512, we limited the maximum length of questions to 64 tokens and the rest of the 446 tokens (including special tokens such as [CLS] and [SEP]) were used for the context. We truncated questions with more than 64 tokens. For contexts the had more than 446 tokens, we adopted a sliding window strategy to scan the whole document using a window size of 446 tokens and a stride size of 396 tokens, so that two consecutive windows had the same 50 tokens overlapped. We also limited the answers to a maximum length of 32 tokens. We used the emrQA dataset[77], which is widely used as a benchmark dataset for MQA. We particularly focused on medications and relations-related questions as Yue et al.[78] found that the two subsets are more consistent. We utilized both F1 score and exact match score for evaluation.

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## DATA AVAILABILITY

The benchmark datasets that support the findings of this study are available from the official websites of natural language processing challenges with Data Use Agreements. More specifically: (1) i2b2 2010, 2012 datasets and n2c2 2018, 2019 datasets: https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/. (2) MedNLI dataset: https://physionet.org/content/mednli/1.0.0/. (3) emrQA dataset: https://github.com/panushri25/emrQA#download-dataset. (4) MIMIC III dataset: https://physionet.org/content/mimiciii/1.4/. (5) PubMed dataset: https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/. (6) Wikipedia dataset: https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2. (7) UF Health IDR clinical notes are not open to the public due to patient privacy information. The GatorTron models pretrained using >90 billion words of text is publicly available at: https://catalog.ngc.nvidia.com/orgs/nvidia/teams/clara/models/gatortron_og.

## CODE AVAILABILITY

The computer codes to train GatorTron models are available from: https://github.com/NVIDIA/Megatron-LM and https://github.com/NVIDIA/NeMo. The computer codes for preprocessing of text data are available from: https://github.com/uf-hobi-informatics-lab/NLPreprocessing https://github.com/uf-hobi-informatics-lab/GatorTron.

Received: 21 June 2022; Accepted: 13 December 2022;
Published online: 26 December 2022

## REFERENCES

1. Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008–2015. *ONC Data Brief*. https://www.healthit.gov/sites/default/files/briefs/2015_hospital_adoption_db_v17.pdf (2016).
2. Adler-Milstein, J. et al. Electronic health record adoption in US hospitals: the emergence of a digital 'advanced use' divide. *J. Am. Med. Inform. Assoc.* **24**, 1142–1148 (2017).
3. Bush, R. A., Kuelbs, C. L., Ryu, J., Jian, W. & Chiang, G. J. Structured data entry in the electronic medical record: perspectives of pediatric specialty physicians and surgeons. *J. Med. Syst.* **41**, 1–8 (2017).
4. Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C. & Hurdle, J. F. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb. Med. Inform.* **17**, 128–144 (2008).
5. Liang, H. et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat. Med.* **25**, 433–438 (2019).
6. Yang, J. 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* **4**, e2126337 (2021).
7. Nadkarni, P. M., Ohno-Machado, L. & Chapman, W. W. Natural language processing: an introduction. *J. Am. Med. Inform. Assoc.* **18**, 544–551 (2011).
8. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
9. Collobert, R. et al. Natural language processing (almost) from scratch. *J. Mach. Learn Res.* **12**, 2493–2537 (2011).
10. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. & Dyer, C. Neural architectures for named entity recognition. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 260–270 (2016).
11. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics.* **36**, 1234–1240 (2020).
12. Vaswani, A. et al. Attention is All you Need. *Advances in Neural Information Processing Systems*. **30** (2017).
13. Wang, A. et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. 353–355* (2018).
14. Wang, A. et al. SuperGLUE: a stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems. **32*** (2019).
15. Qiu, X. et al. Pre-trained models for natural language processing: a survey. *Science China Technological Sciences.* **63**, 1872–1897 (2020).
16. Tay, Y., Dehghani, M., Bahri, D. & Metzler, D. Efficient transformers: a survey. *ACM Computing Surveys*. **55**, 1–28 (2020).
17. Yu, J., Bohnet, B. & Poesio, M. Named entity recognition as dependency parsing. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6470–6476 (2020).
18. Yamada, I., Asai, A., Shindo, H., Takeda, H. & Matsumoto, Y. LUKE: deep contextualized entity representations with entity-aware self-attention. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6442–6454 (2020).
19. Li, X. et al. Dice loss for data-imbalanced NLP tasks. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 465–476 (2020).
20. Xu, B., Wang, Q., Lyu, Y., Zhu, Y. & Mao, Z. Entity structure within and throughout: modeling mention dependencies for document-level relation extraction. *Proceedings of the AAAI Conference on Artificial Intelligence* **35**, 14149–14157 (2021).
21. Ye, D., Lin, Y. & Sun, M. Pack together: entity and relation extraction with levitated marker. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. **1**, 4904–4917 (2021).
22. Cohen, A. D., Rosenman, S. & Goldberg, Y. Relation classification as two-way span-prediction. *ArXiv* arXiv:2010.04829 (2021).
23. Lyu, S. & Chen, H. Relation classification with entity type restriction. *Findings of the Association for Computational Linguistics: ACL-IJCNLP*. 390–395 (2021).
24. Wang, J. & Lu, W. Two are better than one: joint entity and relation extraction with table-sequence encoders. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1706–1721 (2020).
25. Jiang, H. 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. 2177–2190 (2020).
26. Yang, Z. et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 5753–5763 (2019).
27. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 1–67 (2019).
28. Lan, Z.-Z. et al. ALBERT: a lite BERT for self-supervised learning of language representations. *ArXiv* arXiv:*1909.11942* (2019).
29. Wang, S., Fang, H., Khabsa, M., Mao, H. & Ma, H. Entailment as Few-Shot Learner. *ArXiv* arXiv:*2104.14690* (2021).
30. Zhang, Z. et al. Semantics-aware BERT for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*. **34**, 9628-9635 (2020).
31. Zhang, Z., Yang, J. & Zhao, H. Retrospective reader for machine reading comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*. **35**, 14506-14514 (2021).
32. Garg, S., Vu, T. & Moschitti, A. TANDA: transfer and adapt pre-trained transformer models for answer sentence selection. *Proceedings of the AAAI Conference on Artificial Intelligence. 34, 7780-7788* (2020).
33. Bommasani, R. et al. On the opportunities and risks of foundation models. *ArXiv* arXiv:*2108.07258* (2021).
34. Floridi, L. & Chiriatti, M. GPT-3: its nature, scope, limits, and consequences. *Minds Mach* **30**, 681–694 (2020).
35. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* **3**, 1–23 (2022).
36. Shin, H.-C. et al. BioMegatron: larger biomedical domain language model. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4700–4706 (2020).
37. Alsentzer, E. et al. Publicly Available Clinical BERT Embeddings. in *Proc. 2nd Clinical Natural Language Processing Workshop* 72–78 (2019).
38. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
39. Uzuner, Ö., South, B. R., Shen, S. & DuVall, S. L. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J. Am. Med. Inform. Assoc.* **18**, 552–556 (2011).
40. Sun, W., Rumshisky, A. & Uzuner, O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J. Am. Med. Inform. Assoc.* **20**, 806–813 (2013).
41. Yang, X. et al. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. *J. Am. Med. Inform. Assoc.* **27**, 65–72 (2020).
42. Yang, X. et al. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Med. Inform. Decis. Mak.* **19**, 232 (2019).
43. Shoeybi, M. et al. Megatron-LM: training multi-billion parameter language models using model parallelism. *ArXiv arXiv:1909.08053* (2020).
44. Levine, Y., Wies, N., Sharir, O., Bata, H. & Shashua, A. Limits to depth efficiencies of self-attention. *Advances in Neural Information Processing Systems* **33**, 22640–22651 (2020).
45. Sennrich, R., Haddow, B. & Birch, A. Neural Machine Translation of Rare Words with Subword Units. in *Proc. 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 1715–1725 (Association for Computational Linguistics, 2016).
46. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186 (2019).
47. 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**, 1326–1333 (2015).
48. Soysal, E. et al. CLAMP—a toolkit for efficiently building customized clinical natural language processing pipelines. *J. Am. Med. Inform. Assoc.* **25**, 331–336 (2018).
49. Wu, Y., Jiang, M., Lei, J. & Xu, H. Named entity recognition in chinese clinical text using deep neural network. *Stud. Health Technol. Inform.* **216**, 624–628 (2015).
50. Wu, Y. et al. Combine factual medical knowledge and distributed word representation to improve clinical named entity recognition. in *AMIA Annual Symposium Proceedings* vol. 2018, 1110 (American Medical Informatics Association, 2018).
51. Yang, X. et al. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. *J. Am. Med. Inform. Assoc.* **27**, 65–72 (2020).
52. Kumar, S. A survey of deep learning methods for relation extraction. *ArXiv arXiv:1705.03645* (2017).
53. Lv, X., Guan, Y., Yang, J. & Wu, J. Clinical relation extraction with deep learning. *Int. J. Hybrid. Inf. Technol.* **9**, 237–248 (2016).
54. Wei, Q. et al. Relation extraction from clinical narratives using pre-trained language models. *Amia. Annu. Symp. Proc.* **2019**, 1236–1245 (2020).
55. Guan, H. & Devarakonda, M. Leveraging contextual information in extracting long distance relations from clinical notes. *Amia. Annu. Symp. Proc.* **2019**, 1051–1060 (2020).
56. Alimova, I. & Tutubalina, E. Multiple features for clinical relation extraction: a machine learning approach. *J. Biomed. Inform.* **103**, 103382 (2020).

57. Mahendran, D. & McInnes, B. T. Extracting adverse drug events from clinical notes. *AMIA Summits on Translational Science Proceedings*. 420–429 (2021).
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.* **8**, e22982 (2020).
59. Yang, X., Yu, Z., Guo, Y., Bian, J. & Wu, Y. Clinical Relation Extraction Using Transformer-based Models. *ArXiv. arXiv:2107.08957* (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. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 1–14 (2017).
61. Farouk, M. Measuring sentences similarity: a survey. *ArXiv arXiv:1910.03940* (2019).
62. Ramaprabha, J., Das, S. & Mukerjee, P. Survey on sentence similarity evaluation using deep learning. *J. Phys. Conf. Ser.* **1000**, 012070 (2018).
63. Gomaa, W. H. & Fahmy, A. A survey of text similarity approaches. *International journal of Computer Applications* **68**, 13–18 (2013).
64. Wang, Y. et al. MedSTS: a resource for clinical semantic textual similarity. *Lang. Resour. Eval.* **54**, 57–72 (2020).
65. Rastegar-Mojarad, M. et al. BioCreative/OHNLP Challenge 2018. in *Proc. 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* 575–575 (ACM, 2018).
66. Wang, Y. et al. Overview of the 2019 n2c2/OHNLP track on clinical semantic textual similarity. *JMIR Med. Inform.* **8**, e23375 (2020).
67. Mahajan, D. et al. Identification of semantically similar sentences in clinical notes: iterative intermediate training using multi-task learning. *JMIR Med. Inform.* **8**, e22508 (2020).
68. Dagan, I., Glickman, O. & Magnini, B. in *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment* (eds. Quiñonero-Candela, J., Dagan, I., Magnini, B. & d'Alché-Buc, F.) 177–190 (Springer Berlin Heidelberg, 2006).
69. Williams, A., Nangia, N. & Bowman, S. R. A broad-coverage challenge corpus for sentence understanding through inference. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. **1**, 1112–1122 (2018).
70. Bowman, S. R., Angeli, G., Potts, C. & Manning, C. D. A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 632–642 (2015).
71. Shivade, C. MedNLI—a natural language inference dataset for the clinical domain. *PhysioNet* https://doi.org/10.13026/C2RS98 (2017).
72. Conneau, A., Kiela, D., Schwenk, H., Barrault, L. & Bordes, A. Supervised learning of universal sentence representations from natural language inference data. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 670–680 (2017).
73. Rajpurkar, P., Zhang, J., Lopyrev, K. & Liang, P. SQuAD: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2383–2392 (2016).
74. Rajpurkar, P., Jia, R. & Liang, P. Know what you don't know: unanswerable questions for SQuAD. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* **2**, 784–789 (2018).
75. Zhu, M., Ahuja, A., Juan, D.-C., Wei, W. & Reddy, C. K. Question Answering with Long Multiple-Span Answers. in *Findings of the Association for Computational Linguistics: EMNLP 2020* 3840–3849 (Association for Computational Linguistics, 2020).
76. Ben Abacha, A. & Demner-Fushman, D. A question-entailment approach to question answering. *BMC Bioinforma* **20**, 511 (2019).
77. Pampari, A., Raghavan, P., Liang, J. & Peng, J. emrQA: a large corpus for question answering on electronic medical records. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2357–2368 (2018).
78. Yue, X., Gutierrez, B. J. & Sun, H. Clinical reading comprehension: a thorough analysis of the emrQA dataset. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4474–4486 (2020).

## ACKNOWLEDGEMENTS

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. 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

Y.W., J.B., M.G.F., N.P., and X.Y. were responsible for the overall design, development, and evaluation of this study. X.Y. and A.C. 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. Y.W., X.Y., J.B., and W.H. did the bulk of the writing, E.A.S., D.A.M., T.M., C.A.H., A.B.C., and G.L. 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

The authors declare no competing interests.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41746-022-00742-2.

**Correspondence** and requests for materials should be addressed to Yonghui Wu.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.