	doc_1		doc_2		decision	id
cases	authors	• Xi Yang • Aokun Chen • Nima PourNejatian • Hoo Chang Shin • Kaleb E Smith • Christopher Parisien • Colin Compas • Cheryl Martin • Mona G Flores • Ying Zhang • Tanja Magoc • Christopher A Harle • Gloria Lipori • Duane A Mitchell • William R Hogan • Elizabeth A Shenkman • Jiang Bian • Yonghui Wu	authors	• Xi Yang • Nima Pour Nejatian • Hoo Chang Shin • Kaleb Smith • Christopher Parisien • Colin Compas • cheryl Martin • Mona Flores • Ying Zhang • Tanja Magoc • Christopher Harle • Gloria Lipori • Duane Mitchell • William Hogan • Elizabeth Shenkman • Jiang Bian • Yonghui Wu		
	title	GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records	title	GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records	DUPLICATES	153
	publication_date	2022-12-16 00:00:00	publication_date	2022-02-28 00:00:00		
	source	SupportedSources.INTERNET_ARCHIVE	source	SupportedSources.INTERNET_ARCHIVE		
	journal		journal	Cold Spring Harbor Laboratory		
	volume		volume			
	doi		doi	10.1101/2022.02.27.22271257		
	urls	• https://web.archive.org/web/20221220093354/https://arxiv.org/ftp/arxiv/papers/2203/2203.03540.pdf	urls	• https://web.archive.org/web/20220423112131/https://www.medrxiv.org/content/medrxiv/early/2022/03/18/2022.02.27.22271257.full.pdf		
	id	id-7935969406950761467	id	id5957821622821465066		
	abstract	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 5 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 5 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.	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.		
	versions		versions			