GatorTron and GatorTronGPT: Large Language Models for Clinical Narratives

Cheng Peng[1], **Xi Yang**[1,2], **Mengxian Lyu**[1], **Kaleb E Smith**[3], **Anthony B. Costa**[3], **Mona G. Flores**[3], **Jiang Bian**[1,2], **Yonghui Wu**[1,2]

[1]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida
[2]Cancer Informatics Shared Resource, University of Florida Health Cancer Center
[3]NVIDIA, Santa Clara, California, USA
{c.peng, alexgre, lvmengxian, bianjiang, yonghui.wu}@ufl.edu, {kasmith, mflores, acosta}@nvidia.com

Abstract

Large language models (LLMs) have become the foundational technology for natural language processing (NLP). We introduce clinical LLMs including GatorTron and GatorTronGPT, summarize their applications, highlight the impact on clinical NLP and artificial intelligence (AI) applications, and provide insights in using LLMs for medical AI applications.

Introduction

Large language models (LLMs) have become the foundational technology used to explore clinical narratives for artificial intelligence (AI) applications in the medical do-main. There is a surge(Singhal et al. 2023a) of studies exploring LLMs for various medical research as the success of ChatGPT. For example, Galactica(Taylor et al. 2022) was trained using a large corpus of papers, reference material, and knowledge bases, which demonstrated good performance on benchmark datasets including PubMedQA(Jin et al. 2019) and MedMCQA(Pal, Umapathi, and Sankarasubbu 2022). Med-PaLM(Singhal et al. 2023a) and later Med-PaLM 2(Singhal et al. 2023b) achieved an "expert" level of performance on the MedQA dataset of USMLE2-style questions. However, most existing studies focus on general-purpose LLMs such as ChatGPT. There are studies developing LLMs using biomedical literature or fine-tuning general-purpose LLMs using a small set of clinical notes from the MIMIC data-base. Nevertheless, there are limited studies developing and examining clinical LLMs developed using clinical text data. We previously have developed clinical LLMs including GatorTron and GatorTronGPT, which have been widely used in the clinical domain. This abstract introduces our clinical LLMs and their applications in the clinical do-main, provides resources to facilitate the use of GatorTron models, and highlights their impact on clinical NLP.

GatorTron and GatorTronGPT Models

GatorTron(Yang et al. 2022) was trained from scratch using 90 billion words of text from the de-identified clinical notes of the University of Florida Health (82 billion words), MIMIC-III corpus (0.5 billion words), PubMed articles (6 billion words), and Wikipedia (2.5 billion words). We adopted the BERT architecture and developed three different configurations including 345 million parameters (i.e., GatorTron-base), 3.9 billion parameters (i.e., GatorTron-medium), and 8.9 billion parameters (i.e., GatorTron-large). GatorTron models are encoder-only LLMs, where only the encoder component of the transformer was used. GatorTron models are available from: https://huggingface.co/UFNLP.

GatorTronGPT(Peng et al. 2023a) was trained from scratch using 277 billion words from the de-identified clinical text of UF Health (82 billion words) and 195 billion words of diverse English text from the Pile(Gao et al. 2020) dataset. GatorTronGPT adopted the GPT-3 architecture with 5 and 20 billion parameters. GatorTronGPT is a decoder-only generative clinical LLM, where only the decoder of the transformer was used.

GatorTron and GatorTronGPT Applications

GatorTron has been applied to many clinical NLP tasks including clinical concept extraction, medical relation extraction (RE), Semantic Textual Similarity (STS), medical Natural Language Inference (NLI), and medical question answering (QA). GatorTron achieved state-of-the-art performance on many benchmark datasets including i2b2 2010(Uzuner et al. 2011), i2b2 2012(Sun, Rumshisky, and Uzuner 2013), and n2c2 2018(Henry et al. 2020). In addition, Table 1 summarizes the clinical applications of GatorTron models in other studies. GatorTron has been applied for predictions of readmission, mortality, length of stay, insurance denial, disease response, and mobility function. Though many studies focus on the generative LLMs based on the decoder-only architecture, the encoder-only LLM, GatorTron, has been widely used for many medical AI applications

As a generative LLM adopted the decoder-only architecture, GatorTronGPT solved many clinical NLP using unified text-to-text learning, including clinical concept extraction, concept normalization, relation extraction, abbreviation disambiguation, NLI, attribute filling, progress notes understanding. GatorTronGPT was applied to generate narrative sections of clinical notes, which physicians can not differentiate them from real-world clinical notes. Our recent study(Peng et al. 2023a) proposed a unified text-to-

	Model	Evaluation Task	Evaluation Dataset	Performance Ranking
Peng et al. (Peng et al. 2023c)	GatorTron-base	Clinical concept extraction End-to-end clinical relation extraction	2018 n2c2 dataset 2022 n2c2 dataset 2018 n2c2 dataset 2022 n2c2 dataset	1/14 1/7 1/7 2/7
Jiang et al. (Jiang et al. 2023)	GatorTron-base	Readmission In-hospital mortality prediction Comorbidity index prediction Length of stay (LOS) prediction Insurance denial prediction	-	1/6 (Average prediction on five tasks)
Tan et al. (Tan et al. 2023)	GatorTron-base	Natural Language Inference (infer cancer disease response from radiology reports)	RECIST dataset	1/14
Chen et al. (Chen et al. 2023b)	GatorTron-base	Medication mention extraction Event classification Context classification	2022 n2c2 dataset	3/6 2/6 1/6
Pathak et al. (Pathak et al. 2023)	GatorTron-base	Clinical concept extraction (thyroid nodule characteristics extraction)	Ultrasound reports from UF Health	1/5
Chen et al. (Chen et al. 2023a)	GatorTron-base	Clinical concept extraction (delirium symptom extraction)	Delirium symptoms corpus from UF Health	1/8
Ong et al. (Ong et al. 2023)	GatorTron-base	Disease response classification	CT and MRI reports	1/4
Alameldin et al.(Alameldin and Williamson 2023)	GatorTron-base	Natural Language Inference Evidence retrieval	SemEval 2023-Task 7	1/6 1/4
Le et al. (Gao et al. 2020)	GatorTron-base	Mobility functioning classification	n2c2 clinical notes	3/9
Ge et al. (Ge et al. 2023)	GatorTron-base	Multi-label text classification (medical diagnosis prediction)	RAA dataset MIMIC-III	1/9 1/9
Peng et al. (Peng et al. 2023b)	GatorTron-base GatorTron-medium GatorTron-large	Clinical concept extraction End-to-end clinical relation extraction	2018 n2c2 dataset 2022 n2c2 dataset 2018 n2c2 dataset 2022 n2c2 dataset	1/5 1/5 1/5 1/3

Table 1: Performance ranking of GatorTron-base models in various evaluation tasks.

text learning architecture, which solved seven major clinical NLP tasks using a unified GatorTronGPT model using a strategy to freeze LLMs, i.e., keep the model parameters unchanged during prompting.

Conclusion and Discussion

GatorTron and GatorTronGPT models have greatly improved clinical NLP and medical AI applications. The encoder-only GatorTron models have been widely used for patient information extraction and various prediction-based tasks. The decoder-only GatorTronGPT can generate synthetic clinical text for the development of synthetic NLP models to fill the gap of sharing clinical corpora. In addition, GatorTronGPT provides a solution to solve many diverse information extraction and classification tasks using unified text-to-text learning.

Most evaluations of LLMs used standard NLP tasks and datasets, there is an absence of holistic evaluation frameworks to examine LLMs in real-world health, which is a significant disconnection between the LLM evaluation regimes and expected clinical benefits(Wornow et al. 2023).

References

Alameldin, A.; and Williamson, A. 2023. Clemson NLP at SemEval-2023 Task 7: Applying GatorTron to Multi-Evidence Clinical NLI. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 1598–1602.

Chen, A.; Paredes, D.; Yu, Z.; Lou, X.; Brunson, R.; Thomas, J. N.; Martinez, K. A.; Lucero, R. J.; Magoc, T.; Solberg, L. M.; et al. 2023a. Identifying Symptoms of Delirium from Clinical Narratives Using Natural Language Processing. *arXiv preprint arXiv:2304.00111*.

Chen, A.; Yu, Z.; Yang, X.; Guo, Y.; Bian, J.; and Wu, Y. 2023b. Contextualized medication information extraction using Transformer-based deep learning architectures. *Journal of Biomedical Informatics*, 142: 104370.

Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Ge, X.; Williams, R. D.; Stankovic, J. A.; and Alemzadeh, H. 2023. DKEC: Domain Knowledge Enhanced Multi-Label Classification for Electronic Health Records. *arXiv preprint arXiv:2310.07059*.

Henry, S.; Buchan, K.; Filannino, M.; Stubbs, A.; and Uzuner, O. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1): 3–12.

Jiang, L. Y.; Liu, X. C.; Nejatian, N. P.; Nasir-Moin, M.; Wang, D.; Abidin, A.; Eaton, K.; Riina, H. A.; Laufer, I.; Punjabi, P.; et al. 2023. Health system-scale language models are all-purpose prediction engines. *Nature*, 1–6.

Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W. W.; and Lu, X. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Ong, H. X.; Cai, Y.; Yi, Z. M.; Tee, W. H.; Tan Ying Cong, R. S.; Tan, W. C.; and Bakr Azam, A. 2023. Automated classification of disease response in radiology reports using natural language processing.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, 248–260. PMLR.

Pathak, A.; Yu, Z.; Paredes, D.; Monsour, E. P.; Rocha, A. O.; Brito, J. P.; Ospina, N. S.; and Wu, Y. 2023. Extracting Thyroid Nodules Characteristics from Ultrasound Reports Using Transformer-based Natural Language Processing Methods. *arXiv preprint arXiv:2304.00115*.

Peng, C.; Yang, X.; Chen, A.; Smith, K. E.; PourNejatian, N.; Costa, A. B.; Martin, C.; Flores, M. G.; Zhang, Y.; Magoc, T.; et al. 2023a. A Study of Generative Large Language Model for Medical Research and Healthcare. *NPJ Digital Medicine*, 6(1): 210.

Peng, C.; Yang, X.; Smith, K. E.; Yu, Z.; Chen, A.; Bian, J.; and Wu, Y. 2023b. Model Tuning or Prompt Tuning? A Study of Large Language Models for Clinical Concept and Relation Extraction. *arXiv preprint arXiv:2310.06239*.

Peng, C.; Yang, X.; Yu, Z.; Bian, J.; Hogan, W. R.; and Wu, Y. 2023c. Clinical concept and relation extraction using prompt-based machine reading comprehension. *Journal of the American Medical Informatics Association*, 30(9): 1486–1493.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Sun, W.; Rumshisky, A.; and Uzuner, O. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5): 806–813.

Tan, R. S. Y. C.; Lin, Q.; Low, G. H.; Lin, R.; Goh, T. C.; Chang, C. C. E.; Lee, F. F.; Chan, W. Y.; Tan, W. C.; Tey, H. J.; et al. 2023. Inferring cancer disease response from radiology reports using large language models with data augmentation and prompting. *Journal of the American Medical Informatics Association*, 30(10): 1657–1664.

Taylor, R.; Kardas, M.; Cucurull, G.; Scialom, T.; Hartshorn, A.; Saravia, E.; Poulton, A.; Kerkez, V.; and Stojnic, R. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Uzuner, Ö.; South, B. R.; Shen, S.; and DuVall, S. L. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5): 552–556.

Wornow, M.; Xu, Y.; Thapa, R.; Patel, B.; Steinberg, E.; Fleming, S.; Pfeffer, M. A.; Fries, J.; and Shah, N. H. 2023. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1): 135.

Yang, X.; Chen, A.; PourNejatian, N.; Shin, H. C.; Smith, K. E.; Parisien, C.; Compas, C.; Martin, C.; Costa, A. B.; Flores, M. G.; et al. 2022. A large language model for electronic health records. *NPJ Digital Medicine*, 5(1): 194.