Utilizing Large Language Models for Disease Phenotyping in Obstructive Sleep Apnea

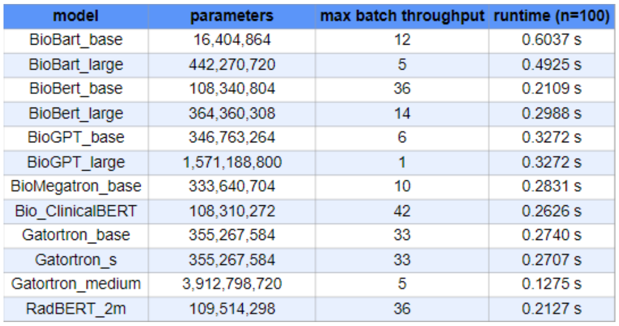
Ifrah Khurram
San Juan Bautista School of Medicine
ifrahk@sanjuanbautista.edu

I Introduction

To understand the interplay between Obstructive Sleep Apnea (OSA) and its associated medical conditions, we harness the capabilities of Large Language Models (LLMs) to characterize patient health from clinical text. OSA, an illness involving airway obstruction during sleep, poses a significant health challenge, particularly among U.S. Veterans. Severe complications, like heart failure (HF), are intricately linked to OSA, urging the VA to optimize medical resource allocation for effective management. This is particularly important with the demographic shift occurring in the United States toward a predominantly elderly population [1]. This research aims to explore the capacity of LLMs to identify disease subtypes linked to OSA. The study seeks to inform ongoing sleep apnea research by contextualizing potential OSA-related symptoms and refining cohort phenotype definitions by employing LLMs to categorize patients based on their clinical text.

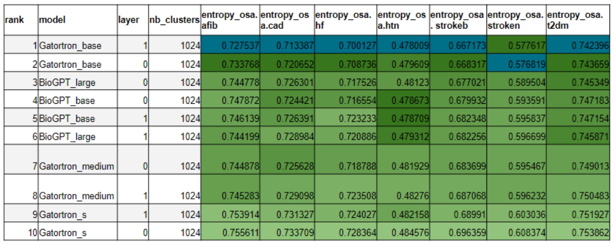
II Methods

We computed document-level embeddings for 331,793 discharge reports from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database using the NERSC's new Perlmutter supercomputer. The Perlmutter system provides sufficient computational resources to scale the embedding of large corpora using contemporary clinical LLMs. Twelve clinical LLMs including BioBART, BioBERT, BioGPT, BioMegatron, Bio_ClinicalBERT, Gatortron, and RadBERT, were benchmarked on Perlmutter's NVidia A100 GPUs, and data parallelism was used to distribute the embedding of clinical text across 40 GPUs. Across 12 different LLM variants, 220 compute hours were used to embed the corpus of discharge reports in the publicly available clinical dataset MIMIC-IV. Embeddings were clustered using K-Means. Then, the purity of clusters was measured by Shannon entropy. Additionally, we assessed the quality of each model by visualizing its latent space with UMAP and manually reading sampled clinical text from clusters of interest.



model	parameters	max batch throughput	runtime (n=100)
BioBart_base	16,404,864	12	0.6037 s
BioBart_large	442,270,720	5	0.4925 s
BioBert_base	108,340,804	36	0.2109 s
BioBert_large	364,360,308	14	0.2988 s
BioGPT_base	346,763,264	6	0.3272 s
BioGPT_large	1,571,188,800	1	0.3272 s
BioMegatron_base	333,640,704	10	0.2831 s
Bio_ClinicalBERT	108,310,272	42	0.2626 s
Gatortron_base	355,267,584	33	0.2740 s
Gatortron_s	355,267,584	33	0.2707 s
Gatortron_medium	3,912,798,720	5	0.1275 s
RadBERT_2m	109,514,298	36	0.2127 s

Figure 1: Parameter count, input throughput, and inference runtime for clinical LLMs.



rank	model	layer	nb_clusters	entropy_osa.afib	entropy_osa.cad	entropy_osa.hf	entropy_osa.htn	entropy_osa. strokeb	entropy_osa.stroken	entropy_osa.t2dm
1	Gatortron_base	1	1024	0.727537	0.713387	0.700127	0.478009	0.667173	0.577617	0.742396
2	Gatortron_base	0	1024	0.733768	0.720652	0.708736	0.479609	0.668317	0.576819	0.743659
3	BioGPT_large	0	1024	0.744778	0.726301	0.717526	0.48123	0.677021	0.589504	0.745349
4	BioGPT_base	0	1024	0.747872	0.724421	0.716554	0.478673	0.679932	0.593591	0.747183
5	BioGPT_base	1	1024	0.746139	0.726391	0.723233	0.478709	0.682348	0.595837	0.747154
6	BioGPT_large	1	1024	0.744199	0.728984	0.720886	0.479312	0.682256	0.596699	0.745871
7	Gatortron_medium	0	1024	0.744878	0.725628	0.718788	0.481929	0.683699	0.595467	0.749013
8	Gatortron_medium	1	1024	0.745283	0.729098	0.723508	0.48276	0.687068	0.596232	0.750483
9	Gatortron_s	1	1024	0.753914	0.731327	0.724027	0.482158	0.68991	0.603036	0.751927
10	Gatortron_s	0	1024	0.755611	0.733709	0.728364	0.484576	0.696359	0.608374	0.753862

Figure 2: Ranking the models by Shannon entropy within each independent OSA + comorbidity combination. Teal is the best model for that comorbidity.

III Results

Among the evaluated LLMs, Gatortron_base's second to last layer consistently achieved low entropy scores along K-means clusters (k=1024). This performance was observed when discriminating OSA from 6 out of 7 comorbid phenotypes including atrial fibrillation, coronary artery disease, heart failure, hypertension, stroke (broad definition), and type 2 diabetes mellitus. Adding on, as the number of clusters increased, model performance at organizing the clusters increased. When ranking clusters by class purity, clinical notes were sampled (n=100) from clusters of high rates of OSA patients with heart failure. These notes describe admissions of patients with a history of OSA, dyspnea as a chief complaint, and no prescription/adherence to Continuous Positive Airway Pressure (CPAP) treatment. For the analysis of the UMAP plots, each point is a patient document. The comorbidity of focus for the plots below is heart failure. There is clustering of the documents by color, separating the comorbidities: HF + OSA. HF is blue, OSA is red, and both comorbidities in a patient is green. Our method shows potential for clustering comorbidities as the color clusters are prominent.

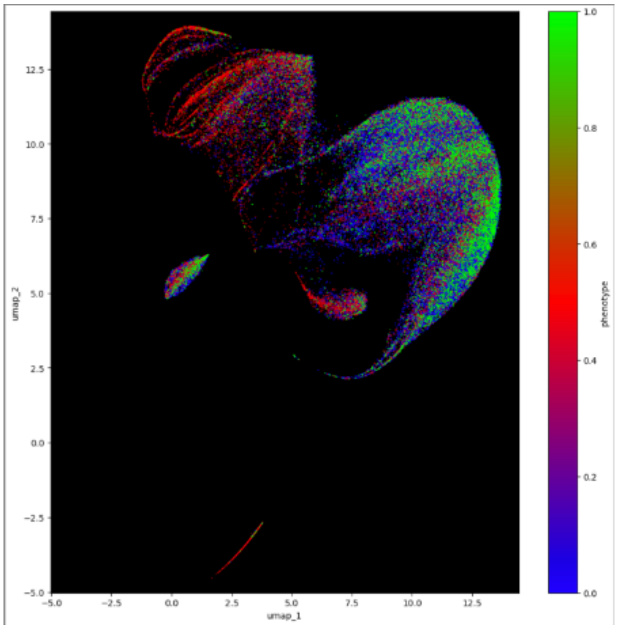


Figure 3: Gatortron_base, layer = 1, K = 1024 on HF + OSA

IV Conclusions

Leveraging LLMs for feature discovery, this work serves as a stepping stone toward understanding how statistical language models organize clinical information and if relations between latent representations capture phenotypes. Identifying sub-cohorts of OSA patients at risk of severe complications is crucial for optimizing the allocation of costly CPAP treatment. Even though LLMs have yet to scale in the narrow domain of clinical medicine, our results show models trained on larger sets of data perform better at organizing corpora by clinically relevant measures.

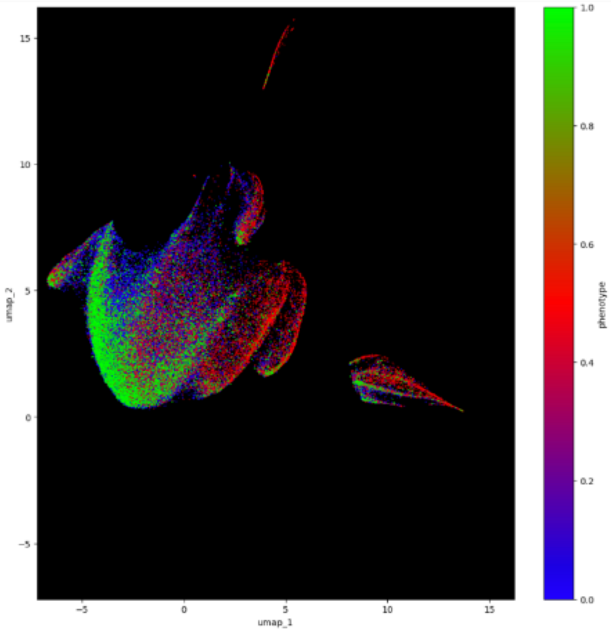


Figure 4: BioGPT_large, layer = 0, K = 1024 on HF + OSA

V Acknowledgements

This work used the resources of the National Energy Research Scientific Computing Center (NERSC). This work was also supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) Program.

References

[1] Leatherby, Lauren. *How a vast demographic shift will reshape the world.* The New York Times. Jul 2023.