Analysis of the development of a LLM in Medicine: A case for Cerebrovascular Disorders using Large Language Models (ALVARO-LLM)

Carlos Utrilla Guerrero
Biomedical Informatics: *Assignment for Prof. Dr. Victor Maojo*
Complementos formativos (*Programa Doctorado DIA*)

December 18, 2024

Abstract

The integration of advanced medical imaging techniques and artificial intelligence (AI) has significantly improved the early detection and diagnosis of cerebrovascular diseases. A particularly compelling question, which this report seeks to address, is how Large Language Models (LLMs) can be employed to enhance the processing, classification, and analysis of cerebrovascular diseases using MRI data. One of the, by now, most interesting published articles and pre-prints I've recently read are [3, 6, 2], in which they explain in a fairly clear and concise methodologies for building foundational models tailored to medical imaging in neuroradiology. Inspired by these contributions, this report briefly outline a system for detecting and classifying brain abnormalities in MRI data using LLMs. A potential approach for automatically detect and classify cerebrovascular diseases is revised and analyzed. As suggested, I followed the protocol and methodology proposed by the slides consisting of the following steps. While many configuration details remain uncertain, the primary goal of this system is to accurately classify two specific types of brain malformations: cerebral cavernous malformations (CCMs) and acute intraparenchymal hemorrhage (AIH) with three steps: brain extraction, candidate malformation detection, and final classification. The system aims to utilize LLMs to enhance the precision and efficiency of medical image interpretation, reflecting the advancements demonstrated in recent literature.

1 Introduction

Cerebrovascular diseases are among the deadliest diseases [1], often requiring timely and accurate diagnosis to improve patient prognosis. Cerebral cavernous malformation (CCM) and small acute intraparenchymal hemorrhage (AIH) represent two distinct entities that require accurate differentiation, particularly when lesions are significantly small. Magnetic resonance imaging (MRI) resources have been recognised as the golden imaging modality for brain abnormalities management, diagnosis and monitoring. Effective utilization of these vast amount of data holds a potential in providing a noninvasive window into the working brain. Yet, analyzing their differentiating between CCM and AIH remains a challenge due to similarities in imaging findings at initial presentation, the complexity of spatiotemporal dynamics, and the inherent difficulty in interpreting indirect indicators

[1] Cerebrovascular diseases, including cerebral cavernous malformations (CCM) and acute intraparenchymal hemorrhage (AIH), show varied prevalence rates in both the general population *(between 0.4% to 0.8%)* [4] and in neonates (0.1 to 1 per 100,000) [5]

of brain functionality. In reported findings in [8], specific AI-assisted software now available for X-ray interpretation can autonomously detect and highlight abnormalities, thus aiding physicians in rapidly locating problem areas and expediting diagnosis [3]. LLMs offer immense potential for enhancing various aspects of medical image processing. Similarly, the application of AI in neuroradiology holds great promise, particularly for tasks like the differentiation of CCM and AIH. Large Language Models (LLMs) based on Transformer architectures, such as GPT-4, have demonstrated potential across numerous radiological tasks, including automating MRI-based neural activity analysis, computer-aided diagnosis of cavernous malformations in brain MR images [12], and genomic data mining for CCM studies [13]. Despite these successes, general-purpose LLMs often lack the specialized architecture necessary for precise segmentation and classification of medical images. Fine-tuning smaller, locally hosted LLMs presents a promising solution to address these limitations. This approach involves training the model on radiology-specific datasets, thereby enhancing its performance on medical tasks while reducing hallucinations and maintaining patient privacy. Moreover, fine-tuned models require fewer computational resources, making them more feasible to deploy in local environments. In this context, this very brief analysis explore the potential the efficacy of fine-tuned LLMs for classifying brain cerebrovascular diseases, specifically focusing on cerebral cavernous malformations (CCM) and acute intraparenchymal hemorrhages (AIH) 1.

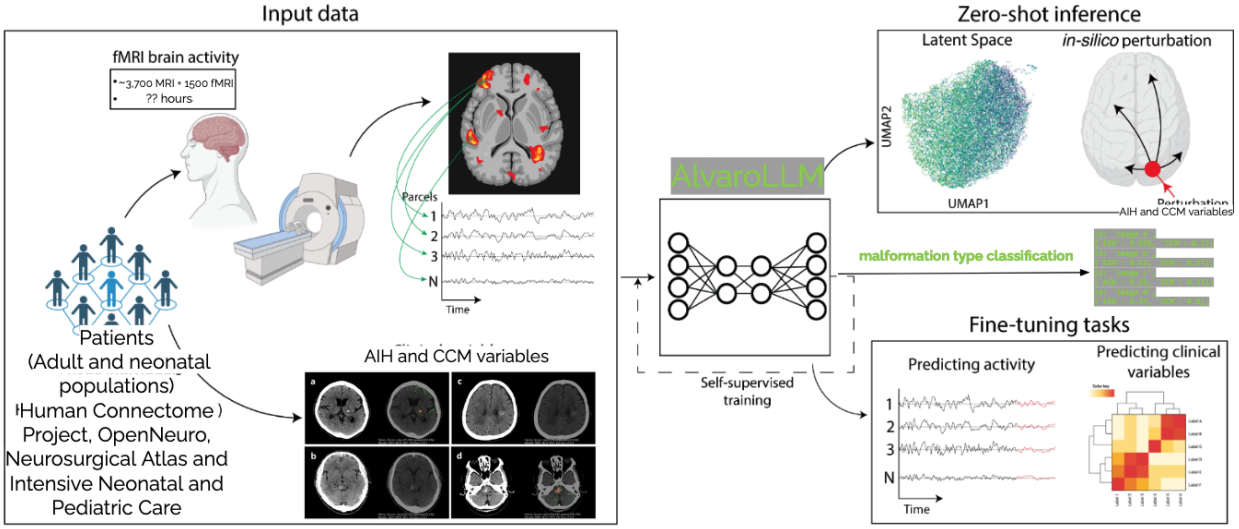

Figure 1: Overview of the AlvaroLLM framework. The model should be training in MRI recordings from datasets. After pretraining, AlvaroLLM should support capabilities through fine-tuning and zero-shot inference. Fine-tuning tasks should demostrate prediction of brain malformations from recordings. Zero-shot applications include inferring functional brain networks from attention weights and using prompting techniques. Adapted figure from BrainLM, (`https://www.biorxiv.org/content/10.1101/2023.09.12.557460v1.full.pdf`).

2 Proposed LLM approach: *AlvaroLLM*

The basic idea of the proposed automated Lesion Visualisation and Analysis for Radiological outcomes and Cerebrovascular disorders (or *AlvaroLLM*) is similar to earlier proposed framework of *BrainLM* [6] as well as Kanzawa Study [3], with significant changes in its selection of generalist LLM and data for fine-tuning. For completeness, I present a brief description of the two frameworks:

1. *BrainLM*: This model aims to build a foundational model (employing Transformer [11] masked autoencoder architecture with BERT [1] and Image Recognition LLM) for MRI functional

analysis (fMRI). We also select BERT variant, but unlike BrainLM, our model variant used for vision-language tasks is OpenFlamingo model, and the processing of imaging is on MRI type in AlvaroLM. A key innovation of BrainLM lies in fine-tuning, as they wee able to decode brain dynamics to predict clinical variables and psychiatric disorders better than baseline models. Likewise, BrainLM identifies intrinsic functional connectivity maps directly from pretraining, clustering parcels into known systems. These innovative techniques could be analysed and integrated in our proposed framework, albeit time-constrain does not allow me to investigate this feature deeply.

2. *Kanzawa Study*: In this study, they investigate the efficacy of utilising fine-tuned LLM to classify MRI into brain tumors type. This study utilised BERT model, version 3 for classifying brain tumor. The model employs subword tokenization and was pretrained using Japanese Wikipedia data. In each session, the model was fine-tuned via the training dataset and its performance was assessed on the validation dataset.

Keeping in mind these two unique studies on how to use of fine-tuned LLM for report classification task, I then attempt to propose the Alvaro approach for especifially, classifying cerebrovascular diseases using MRI. I followed the protocol and methodology proposed by the slides consisting of the following steps:

2.1 Data collection from open-source and expected volume.

There is limited open-source data specifically targeting the classification of cerebral cavernous malformations (CCM) and acute (AIH) in adults patients, albeit neonates and pediatric patients is also desired. These repositories provide high-quality MRI and medical records, suitable for training, validation, and testing. Although the sources are not directly covering specifically brain lesions, e leveraged three sources of data to construct a comprehensive dataset for training, validation, and testing:

1. Human Connectome Project (HCP): The HCP provides 1,000 high-quality fMRI recordings from healthy adults scanned plus medical records, including demographic and cognitive information. Provides a benchmark dataset for anatomical and functional modeling.

2. OpenNeuro: The OpenNeuro is an open-source platform containing diverse datasets of MRI and fMRI recordings, along with associated metadata. For this project, we focus on datasets with structural and functional MRI relevant to cerebrovascular diseases.

3. Neurosurgical Atlas: A detailed repository of annotated neurosurgical cases, including cavernous malformations and hemorrhages. Serves as an expert-guided annotation and reference standard source.

   Given the scarcity of dataset publicly available on MRI for brain malformations, patterning with hospitals or neurosurgical university hospital centers in MRI trials (e.g., MAYO trial seem ideal (and future work), enabling accessing (under research agreements) as well as categorization by expert radiologist. In addition, request patient dataset to the corresponding authors of [3] and [4] is another goal of this. For sake of simplicity, these resources are excluded in the expected dataset as shown in Table 1.

Table 1: Summary of Dataset Characteristics

Source	MRI Rough Volume	Purpose
Human Connectome Project (HCP)	1,00	Benchmark for anatomical and functional modeling.
OpenNeuro	∼2,500	Broadens age ranges and includes pathologies.
Neurosurgical Atlas	200	Gold standard for validation and lesion-specific training.
Combined Total	∼3,700 MRI	Training LLMs.

Table 2: Dataset Splits for Training, Validation, and Testing

Dataset Split	Percentage (%)	Number of MRI Recordings	Purpose
Training Set	80	∼2,700	Model training.
Validation Set	10	∼50	Hyperparameter tuning and model selection.
Test Set	10	∼50	Final evaluation on unseen data.

2.2 Data Pre-processing

Table 2 shows a roughly estimation and composition of each dataset. The model should train on 80% of the dataset (∼4,000) and evaluated on the held-out 10% and the full HCP dataset. All recording underwent standard pre-processing including motion correction, normalization, temporal filtering, and de-noising to prepare the data according to [6, 4].

2.3 Model Architecture

I envision a open-source architecture that consists of a Transformer-based masked autoencoder architecture of BERT and Vision Transformer such Flamingo-80B (see Table 2. BERT and Flamingo-80B are used combinatiolly, where BERT and Falmingo-80B processes the MRI image patches, and later generate natural language outputs specifically for images. This architecture allows for fusing both text-based understanding and image-based classification, which is ideal for medical applications involving MRI scans and clinical records. The base-sized model consists of the following key components:

- BERT:
  - 12 layers of multi-headed self-attention.
  - 768 hidden state dimensions.
  - 12 attention heads.
  - Approximately 110 million parameters in total.
- Flamingo-80B:
  - Vision Transformer with 80 billion parameters.
  - Multi-modal integration of text and image processing.

– Specialized in processing both image patches and associated text for improved contextual understanding.

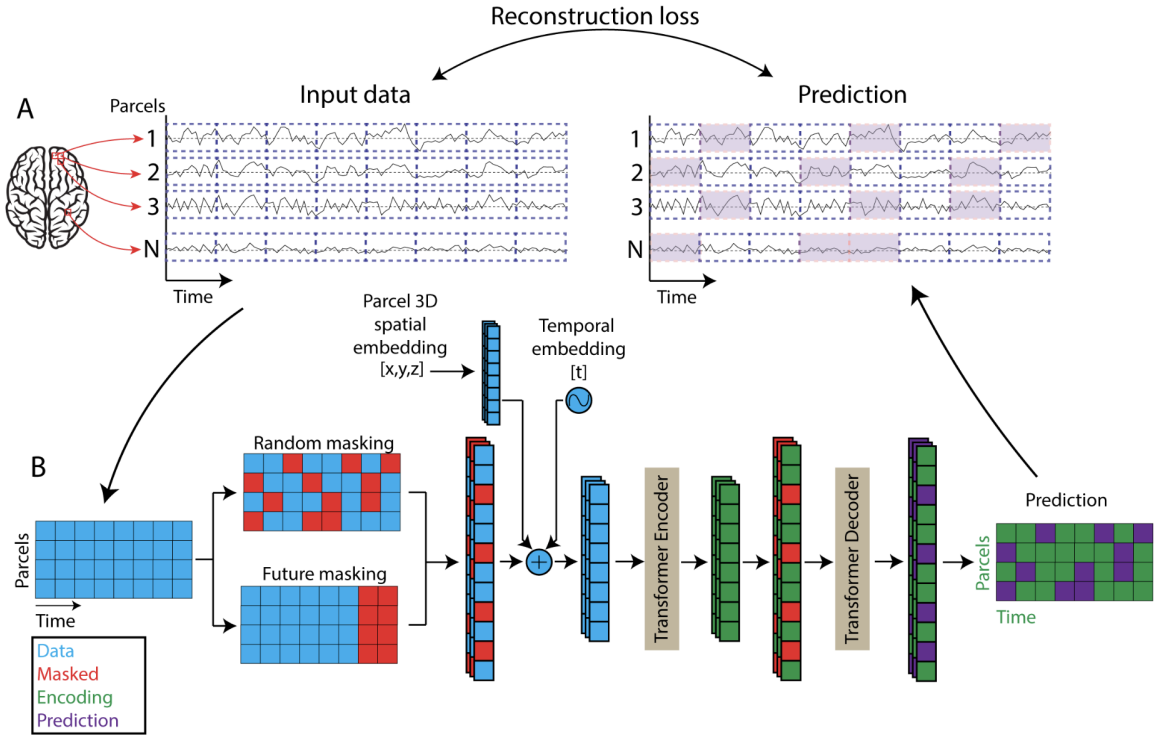

Figure 2: AlvaroLLM architecture and training procedure (see BrainLM for further details). A) The MRI recordings are compressed into dimensions (parcels). The recordings are randomly trimmed to N time points. For each parcel, the temporal signal is split into patches of smaller time points each (blue dashed boxes). The resulting patches are converted into tokens via a learnable linear projection. B) From the total number of tokens (blue), a subset is masked (red), either randomly or at future timepoints. We then add the learnable spatial and temporal embeddings to each token. These visible tokens (blue) are then processed by a series of Transformer blocks (Encoder). The input to the Decoder is the full set of tokens, consisting of encoded visible tokens (green) and masked tokens (red). The Decoder also consists of Transformer blocks and ultimately projects the tokens back to data space. Finally, we compute the reconstruction loss between the prediction (purple) and the original input data (blue)Adapted figure from BrainLM, (`https://www.biorxiv.org/content/10.1101/2023.09.12.557460v1.full.pdf`).

To test the medical reasoning of the system and its ability to stratify medical images for classification task of brain malformations, the integration of prompting interface is foreseen via Flamingo to easily interact with the system using text and images. This approach allows to directly focus to particular anatomical regions, structures, facilitating the classification task.

2.4 Training Procedure

To train the model, I envision a method similar to recently published approaches [6, 4]. In these contributions, the training session was repeated 15 times with the same hyperparameters and training data to account for the inherent randomness in the fine-tuning process. Each session consisted of 10 epochs to optimize the model's performance. The number of epochs for the fine-tuning process was determined empirically based on the model's performance on the training and validation datasets. In each session, the model was fine-tuned via the training dataset and its

performance was assessed on the validation dataset. Other hyperparameters were set to default values of the Transformers library. The performance of the model before fine-tuning was also evaluated on the validation dataset. For each fMRI recording, we sampled random 200-timestep subsequences. The parcel time series were divided into segments of 20 timesteps, yielding 10 segments per subsequence. These were embedded into a 512-dimensional space and masked with a ratio of 20%, 50%, or 75%. The unmasked segments were encoded via a Transformer encoder with 4 self-attention layers and 4 heads. This was decoded by a 2-layer Transformer to reconstruct all segments. We trained with batch size 512 and the Adam optimizer for 100 epochs, minimizing the mean squared error between original and predicted embeddings.

2.5 Computational Resources

Implementation of fine-tuning will be developed using Python 3.10.13 (https://www.python.org/) and possibly with Transformers library 4.35.2 (https://huggingface.co/) on a workstation with a Core™ i9-12900 F central processing unit, an NVIDIA GeForce RTX 3090 graphic processing unit, and 128 GB of random access memory. The Transformers library's *AutoModelForSequenceClassification* class method might be also employed to configure the model for categorizing reports into three groups according to logits. Specifically, for *Flamingo-80B*, which contains 80 billion parameters, training or fine-tuning requires access to high-performance GPUs or TPUs. The later might be challenging to implement, given the limited resources. A pipeline of the prototyped system should be publicly available in Github at later stage.

2.6 Evaluation and Statistical Analysis

Regarding the target malformation, fine-tuned Alvaro should be proceeded to predict clinical variables from MRI recordings. Based on the analysis [5], the pretrained encoder tend to be appended with 3 Multilayer perception (MLP) head and trained to predict targets including sizes of CMM and AIH groups, distribution of lesion locations in lobar, basal gangline and thalamus, probability scores of hemorrhage in a slicewise and patientwise manner. For normalization of scores, a standard log transformation to make values less exponentially distributed and mix-max scaling range is divised. By analyzing the optimized perturbations, we can identify which MRI features are most influential in altering the model's predicted brain state. This reveals insights into relationships learned by AlvaroLM. The performance of the classifiers will be evaluated using standard metrics such as area under the reciever-operator curve (AUC), F1-score according to Hugh et al, as well as P- values and estimated probability distributions using bootstrapping with 1,000 replicates and parired 2-tailed t-test.

2.7 Ethical requirements and limitations

This approach should ensures privacy of patients by keeping sensitive data within a secure local environment. For that, this system should be conducted in accordance with the tenets of the Declaration of Helsinki and was approved by the local institutional review board. Similarly, a systematically examination of the imaging dataset across the sources is mandatory, in order to comply with and GDPR for AI.

3 Discussion

Our analysis is focus on the promised approach on is unique on the use of fine-tuned LLM for report classification task in Neurobiology. Given the optimal performance of fine-tuned LLM in classifying brain MRI previously, I analyses and proposed a very first approach that can be extended to a variety of other radiology tasks such AIH and CNN differentiation. This analysis aims to explore the idea that even small, locally available LLMs can achieve sufficient performance through fine-tuning.

Adult patients instead of Neonatalogy and pediatric ages The initial desire goal of this study was to study these vascular malformations on neonatal intensive care [7], particularly in the context of diagnosing CMMs and AIH in newborns and pediatric ages. These malformations are extremely rare, with limited understanding of their presentation, condition and management [5]. The scarcity of open-source MRI dataset made this objective a daunting task: the vast majority of data isn't publicly available, and most existing open datasets for neonatal brain abnormalities are found in utilizing electroencephalography (EEG) data, which is largely the most cost-effective data collection method for neonatal brain malformation [9] and ML, but are insufficient for building accurate LLMs for CCM and AIH classification. While the usefulness of signal processing of EGG data is very important for real-time monitoring and seizure detection, they lack the spatial resolution and anatomical detail MRI provides, making less suitable for detecting structural abnormalities like CCMs or AIH. The rarity of specific datasets, particularly for neonatal CCM and AIH, means that current models and studies predominantly rely on available adult datasets, with research in the pediatric realm lagging behind. Further efforts are needed to create accessible, high-quality MRI datasets focusing on neonates and children with these conditions to enable better model training and more precise clinical applications.

LLM selection and architecture In this report, I presented some potential approach of using generalist models in favor of specialised ones that have been employed to achieve *state-of-the-art* performance. Specialised LLMs such as PubMedBERT, BiooGPT, Med-PaLM and many others have been the predominant technique for solving biomedical tasks. However, the research project conducted in [2] provides an evidence of losing their advantage over generalist ones such GPT-4, BERT or Flamingo-80B. This may explain why this generalist model, with appropriate prompting techniques, excels in several specialized domains such as medicine. The decision to use BERT as the foundational LLM was driven by its strength in understanding context and performing classification tasks. While there are several general-purpose models available, such as GPT, OpenCLIP, OpenFlamingo, and even U-Net, each comes with unique strengths and limitations. GPT excels in generative tasks but may require extensive fine-tuning for classification problems. OpenCLIP and OpenFlamingo are multimodal models designed for image-text interaction, making them promising candidates for medical imaging tasks. U-Net is specialized for segmentation and could complement an LLM for localization tasks. Although LLMs can not outperform classic deep learning models like CNN or ResNet on benchmarked datasets, it is worth noting that they can serve as chat assistants to provide pre-diagnosis before making decisions [10]. A comprehensive evaluation to compare the performance of these models should be performed, a desired goal of me.

Low Computational Resource The low computational requirements of smaller LLMs make them feasible to implement in clinical settings. However, the lack of resources (*even medical knowledge!*) in notorious and the limited computational resources available for this study posed significant

challenges. The primary development environment relied on an Apple M2 chip with 16 GB of RAM and macOS Sequoia, which lacks sufficient GPU support for training large models like BERT or OpenFlamingo. To mitigate these limitations, we utilized Google Colab's GPU capabilities for model training and fine-tuning. While this provided some relief, the dependency on cloud-based resources restricts scalability and may not accommodate larger datasets or more complex architectures. Future implementations would benefit from access to high-performance computing clusters or dedicated GPUs, which are essential for training large-scale multimodal models (in case is needed in the future).

Ethical Implications This report underscores the importance of integrating pediatric-specific datasets, such as Accept-AI, into future work. Datasets like Accept-AI, which provide data on pediatric medical conditions, could significantly enhance the development of AI models tailored to neonates and children. However, using such sensitive data requires compliance with ethical and legal frameworks. The EU AI Act, which regulates AI systems, particularly in high-risk domains like healthcare, imposes strict guidelines on transparency, data quality, and accountability. Projects using pediatric data must align with these regulations to ensure ethical AI deployment and maintain public trust.

The implications of the EU AI Act extend to this project, as models for medical diagnosis are classified as high-risk applications. The act mandates rigorous documentation of data sources, validation of model accuracy, and clear accountability mechanisms. Incorporating these requirements into the project's workflow will not only enhance its reliability but also prepare it for real-world clinical deployment.

4 Limitations

This report has limitations and leaves proposal for future work and improvements. It represents a proof-of-concept focused solely on imaging data. Integrating imaging with more complex information, such as clinical reports and patient histories, is expected to significantly improve the system's diagnostic quality. However, the current system exhibits hallucinations when addressing questions outside its training context. Another limitation is that the images used for evaluation are not specific to CCM and AIH, but instead focus on general tasks to assess the model's classification skills. Consequently, the system's ability to accurately identify these specific lesions requires further validation. Follow-up work is necessary to refine the approach and establish its clinical utility.

A fourth limitation is the lack of discussion on interpretability and explainability. The system does not include interpretability methods, such as attention analysis, to reveal how it processes input data. Future efforts should prioritize extracting insights from the model's representations by visualizing self-attention weights and identifying the regions prioritized during encoding. Attention analysis could provide valuable interpretability, potentially yielding neuroscientific insights.

Finally, this report points BERT as the LLM "backbone". While proprietary models like GPT-4V (OpenAI) and Gemini Ultra (Google) are likely more powerful, BERT variations currently represent the state-of-the-art among open-source models. Due to their time-limitation nature,I was unable to test these other models, but we anticipate that they—and future advancements in open-source LLMs—will enable the development of even more high-performing vision-language systems.

5 Conclusion

This report underscores the pivotal role of LLMs in advancing medical image processing for neuroradiology. AlvaroLLM approach represents a promising step forward in leveraging LLM for applications at the intersection of cerebrovascular diseases and artificial intelligence. By integrating state-of-the-art vision-language models with advanced medical imaging data such MRI, our analysis demonstrate the potential of AlvaroLLM to enhance diagnostic workflows and facilitate cerebral malformation classification. Its capabilities could assist clinicians in identifying and prioritizing brain abnormalities, such as cavernous malformations, and acute hemorrhages, thereby reducing the risk of missed diagnoses in radiological reports. Furthermore, AlvaroLLM desires is to become a valuable tool for researchers, enabling the extraction of relevant cases from extensive datasets, which can accelerate progress in understanding and managing cerebrovascular diseases.

References

[1] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. en. May 2019. URL: http://arxiv.org/abs/1810.04805.

[2] Tianyu Han et al. "Multimodal Large Language Models are Generalist Medical Image Interpreters". In: *medRxiv* (Dec. 2023), p. 2023.12.21.23300146. DOI: 10.1101/2023.12.21.23300146. URL: https://www.medrxiv.org/content/10.1101/2023.12.21.23300146v3https://www.medrxiv.org/content/10.1101/2023.12.21.23300146v3.abstract.

[3] Jun Kanzawa et al. "Automated classification of brain MRI reports using fine-tuned large language models". In: *Neuroradiology* (July 2024), pp. 1–7. ISSN: 14321920. DOI: 10.1007/S00234-024-03427-7/FIGURES/2. URL: https://link.springer.com/article/10.1007/s00234-024-03427-7.

[4] Jung Youn Kim et al. "Improved differentiation of cavernous malformation and acute intraparenchymal hemorrhage on CT using an AI algorithm". In: *Scientific Reports 2024 14:1* 14.1 (May 2024), pp. 1–7. ISSN: 2045-2322. DOI: 10.1038/s41598-024-61960-0. URL: https://www.nature.com/articles/s41598-024-61960-0.

[5] Ismael Moreno et al. "Giant cerebral cavernous malformation in a newborn: a rare case report and review of literature". In: *Child's Nervous System* 40.7 (July 2024), pp. 2215–2221. ISSN: 14330350. DOI: 10.1007/S00381-024-06401-Z/TABLES/1. URL: https://link.springer.com/article/10.1007/s00381-024-06401-z.

[6] Josue Ortega Caro et al. "BrainLM: A foundation model for brain activity recordings". In: *bioRxiv* (2023), pp. 2023–09.

[7] Janno S. Schouten et al. "From bytes to bedside: a systematic review on the use and readiness of artificial intelligence in the neonatal and pediatric intensive care unit". In: *Intensive Care Medicine* 50.11 (Nov. 2024), pp. 1767–1777. ISSN: 14321238. DOI: 10.1007/S00134-024-07629-8/FIGURES/5. URL: https://link.springer.com/article/10.1007/s00134-024-07629-8.

[8] Dianzhe Tian et al. "The role of large language models in medical image processing: a narrative review". In: *Quantitative Imaging in Medicine and Surgery* 14.1 (Jan. 2024), pp. 1108–1121. ISSN: 22234306. DOI: 10.21037/QIMS-23-892/COIF. URL: https://qims.amegroups.org/article/view/119330/htmlhttps://qims.amegroups.org/article/view/119330.

[9] Turker Tuncer et al. "TATPat based explainable EEG model for neonatal seizure detection". In: *Scientific Reports* 14.1 (Dec. 2024), p. 26688. ISSN: 20452322. DOI: 10.1038/S41598-024-77609-X. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11535284/.

[10] Minh-Hao Van, Prateek Verma, and Xintao Wu. "On Large Visual Language Models for Medical Imaging Analysis: An Empirical Study". In: (Feb. 2024). URL: http://arxiv.org/abs/2402.14162.

[11] Ashish Vaswani et al. "Attention is All you Need". en. In: ().

[12] Huiquan Wang, S. Nizam Ahmed, and Mrinal Mandal. "Computer-aided diagnosis of cavernous malformations in brain MR images". In: *Computerized Medical Imaging and Graphics* 66 (June 2018), pp. 115–123. ISSN: 0895-6111. DOI: 10.1016/J.COMPMEDIMAG.2018.03.004.

[13] Yiqi Wang et al. "Large language models assisted multi-effect variants mining on cerebral cavernous malformation familial whole genome sequencing". In: *Computational and Structural Biotechnology Journal* 23 (Dec. 2024), pp. 843–858. ISSN: 2001-0370. DOI: 10.1016/J.CSBJ.2024.01.014.