Extracting Epilepsy Patient Data with Llama 2

Ben Holgate*, Shichao Fang*, Anthony Shek, Matthew McWilliam*, Pedro Viana*, Joel S. Winston*, James T. Teo*, and Mark P. Richardson***
* Department of Basic & Clinical Neuroscience, King's College London
** Guy's and St Thomas' NHS Foundation Trust
benjamin.holgate@kcl.ac.uk

Abstract

We fill a gap in scholarship by applying a generative Large Language Model (LLM) to extract information from clinical free text about the frequency of seizures experienced by people with epilepsy. Seizure frequency is difficult to determine across time from unstructured doctors' and nurses' reports of outpatients' visits that are stored in Electronic Health Records (EHRs) in the United Kingdom's National Health Service (NHS). We employ Meta's Llama 2 to mine the EHRs of people with epilepsy and determine, where possible, a person's seizure frequency at a given point in time. The results demonstrate that the new, powerful generative LLMs may improve outcomes for clinical NLP research in epilepsy and other areas.

1 Introduction

Advances in Natural Language Processing (NLP), in particular pre-trained Transformers (Vaswani et al., 2017) and Large Language Models (LLMs), create opportunities to develop new methodologies to mine free-text Electronic Health Records (EHRs) for clinical research. One such opportunity is to investigate associations between anti-seizure medications (ASMs) and the frequency of seizures suffered by people with epilepsy, which is typically recorded in free text in the UK's National Health Service (NHS). Mostly, this text consists of doctors' and nurses' reports of outpatients' ambulatory visits; the reports are shared with a patient's primary care physician in the form of a letter. The majority of hospital care episodes for people with epilepsy occur in ambulatory care.

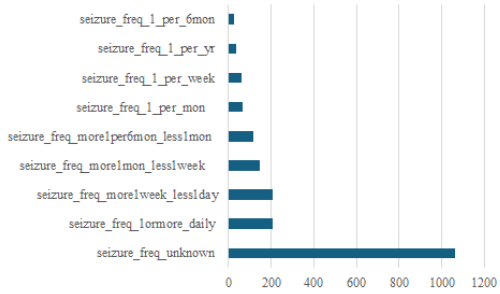

Figure 1: Distribution of 9 seizure frequency categories in annotated dataset.

Yet these reports are unstructured and typically noisy as they include a range of medical and administrative information, such as the patient's medication, other therapies, and details disclosed during previous clinic visits. Moreover, the reports often do not include any information about seizure frequency and, if they do, the language is often imprecise so that the nature of the frequency is vague or unclear. These factors make the application of NLP to EHRs to research seizure frequency challenging.

Epilepsy affects about 1% of the general population (Fiest et al., 2017). Around 30% of people with epilepsy do not respond to ASMs and are therefore regarded as refractory to treatment (Kwan and Brodie, 2000). While there are more than 30 individual ASMs and a much larger number of possible combinations of ASMs taken together, it is not feasible to try them in every refractory patient. This underlines the importance of research in predicting which ASMs would have the greatest impact on epileptic seizures for individual patients.

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 526–535

Although there is some published research on applying pre-trained Transformers to investigate epileptic seizure frequency among EHRs, the more recent opportunity of applying the new, generative LLMs for the same task is under-explored. However, it is expected that the application of generative LLMs to epilepsy research will increase significantly (van Diessen et al., 2024). The paucity of published research in this field may largely be due to the fact that these models are so new.

The most extensive relevant research we found was a long-term study (Xie et al., 2023; Xie et al., 2022a; and Xie et al., 2022b) that used a different methodology from ours to extract seizure frequency information from EHRs. The University of Pennsylvania researchers applied the pre-trained Transformers Bio_ClinicalBERT (for text classification), RoBERTa (for text extraction), and a T-5 model (to summarize sentences with seizure frequency data) to free-text data in EHRs to determine the seizure frequency of a person with epilepsy or whether that person was seizure free. For seizure frequency, they framed the task as an extractive question-answering problem, asking the language model to identify statements that answered the question: "How often does the patient have seizures?" They then simplified each sentence into a standardized format, "X per Y [day/month/year/visit]"; for example, "1 per 1 week". They subsequently manually annotated 1,000 sentences of seizure frequency generated by their models with the formatted summaries, then split them into training (700 sentences) and testing (300 sentences) datasets. Finally, they fine-tuned a T5-large model using Huggingface on the training dataset and made predictions on the test dataset. The researchers declared an "overall accuracy" score of 0.88 for seizure frequency, which comprised scores for each of "sentence accuracy", "summary accuracy", and "quantity accuracy".

That study follows a large body of research applying pre-trained Transformers to a wide variety of clinical tasks, such as predicting the risk of seizure recurrence among children with epilepsy (Beaulieu-Jones et al., 2023), inferring cancer disease response from free-text radiology reports (Tan et al., 2023), or detecting dementia with in-hospital clinical notes (Liu et al., 2023). Two other studies used rules-based NLP approaches to identify seizure frequency in unstructured clinic letters (Fonferko-Shadrach et al., 2019; Decker et al., 2022).

Our objective was to apply a new, generative LLM to the task of determining seizure frequency from free-text data. LLMs are built on the architecture of the Transformers but are much larger and more powerful language models. We were encouraged by recent research that demonstrates the benefits of using LLMs with clinical texts (for example, Agrawal et al., 2022; Thirunavukarasu et al., 2023; and Zhou et al., 2023). Our research, however, was restricted to using only an open-source language model because we used confidential NHS medical data that had to remain within the hospital's secure IT network for regulatory reasons. Therefore, we could not experiment with LLMs such as OpenAI's ChatGPT that are only available via an API to an off-site service. We found that Meta's Llama 2 (Touvron et al., 2023) performed best for our purposes within our limitations (see details in section 2.4). The LLM was run on up to eight Nvidia V100 GPUs.

2 Data and Methods

2.1 Data Collection

We selected 41,340 EHRs, the vast majority of which comprised doctors' and nurses' reports of outpatients' ambulatory visits, from King's College Hospital (KCH) in London spanning a decade from 2013-2022. The records related to 6,853 unique adult people with epilepsy being treated at KCH. We defined a person with epilepsy as someone who has at least one record of an epilepsy diagnosis. The selection was done via CogStack, an open-source information retrieval and extraction platform for EHRs developed by researchers at the NIHR Maudsley Biomedical Research Centre in London.[1] CogStack integrates with KCH's EHRs. We defined a set of epilepsy-related keywords and medical codes, and then used CogStack's search functionality to filter out EHRs that matched these definitions. We then used stratified random sampling to select 3,000 EHRs to create an annotated dataset, which ensured proportional distribution across the original dataset in regard to age, gender, and ethnicity to minimize bias. Subsequently, a team of six annotators, comprising four neuroscience clinicians and two data scientists, manually annotated the 3,000 EHRs for

[1] https://cogstack.org

key data categories of the project, in particular seizure frequency, as well as seizure freedom, current anti-epilepsy medication, epilepsy type, seizure type, associated symptoms, and comorbidities. Due to time and resource limitations, as well as tight deadlines, the annotators worked on separate batches of the 3,000 EHRs, rather than having two annotators work on the same batch for moderation. A user guide was written for the annotators with instructions on how to annotate for each key data category, including seizure frequency with eight temporal frequencies and 'unknown' (see section 2.3 and Table 1 for more details).

2.2 Broader Research Project

This research on seizure frequency was part of a broader epilepsy research project run by the Department of Basic & Clinical Neuroscience in the School of Neuroscience at King's College London in the UK. The objective of the broader project is to apply machine learning at scale in an attempt to discover combinations of ASMs that enable refractory people with epilepsy to stop having seizures. Seizure frequency is a critical data point for this broader project.

2.3 Seizure Frequency Categories

We chose nine categories of seizure frequency for people with epilepsy, eight of which are for temporal frequencies and the last for unknown, meaning either the Electronic Health Record (EHR) contained no reference to seizures (which is common) or the LLM could not determine the frequency of seizures, due to the ambiguity of the text. We arrived at the nine categories after reviewing other studies (mostly non-NLP research) that investigated the frequency of epilepsy seizures (Wie et al., 2023; Westrhenen et al., 2022; Hsieh et al., 2022; Choi et al., 2014; and van Hout et al., 1997). Our aim was to stress test Llama 2 to gauge to what degree it could identify different seizure frequencies from unstructured text. We created shorthand labels for the nine seizure frequency categories for the annotation dataset (see Figure 1), mainly for ease-of-use when it came to writing Python code to evaluate the performance of Llama 2. Subsequently, we found that Llama 2 could often provide answers on the temporal duration of seizure frequency in an EHR in the format of our shorthand category labels following few-shot prompting instructions. However, after discovering that Llama 2 did not generate accurate enough predictions of seizure frequency over the nine categories, we then aggregated these nine categories into three categories (without performing a new experiment), which in turn resulted in Llama 2 predictions that were more accurate and usable for the broader epilepsy research project (see Table 1).

3 Categories	Aggregation from 9 Categories
Infrequent	once per year once per 6 months > once per 6 months, < once per month once per month
Frequent	> once per month, < once per week once per week > once per week, < once per day once or more per day
Unknown	Unknown

Table 1: 3 seizure frequency categories and aggregation from 9 categories.

2.4 Llama 2: Model Development and Implementation

We used LangChain as our development framework because it provides convenience and flexibility for building applications powered by LLMs.[2] First, we deployed LangChain in our local environment, then we downloaded a 13B parameter version of Llama 2 from Hugging Face and loaded it into LangChain.[3] LangChain offers simple interfaces for loading and initializing LLMs. After the model was loaded and initialized, we loaded various templates into LangChain, allowing us to perform multiple LLM operations in the local environment. While Meta provides 7B, 13B, and 70B different-sized models of Llama 2, our GPU platform did not have the computing power to run the largest 70B model. We used a chat version of Llama 2 13B that had been quantized by GPTQ. Although Meta released Llama 3 in April 2024, this did not provide enough time to run experiments using the latest Llama version in light of the submission deadline for this paper.

[2] https://www.langchain.com

[3] https://huggingface.co/TheBloke/Llama-2-13B-chat-GPTQ

Read the following context then work through these 3 steps.

1. Determine whether the context has any information about the frequency of the epilepsy patient's seizures.

2. If the context does not have any information about the frequency of the epilepsy patient's seizures, then you answer: 'I do not know.'

3. If the context does have information about the frequency of the epilepsy patient's seizures, then you estimate the frequency of the epilepsy seizures and express the frequency in terms of per year, per month, per week, or per day, whichever is most relevant.

Figure 2: Query structure in 3 steps for Llama 2.

Clear example:

"We went through some of his seizures and in March he had two convulsions and three or four petit mal."

Seizure diary example:

"Seizures: Partial seizures: July x 23, Aug x 0, Sept x 1, Oct so far x 7 (x1 daily 7th to 10th, 14th x1, 15th x 2, 18 x1."

Ambiguous example:

"Louise and her mum confirm no seizures with her last seizure was possibly in November but they are not sure."

Figure 3: Examples of context for epilepsy information in Electronic Health Records (excerpts from clinical letters).

As is well known with generative LLMs, the key issues with developing the model for seizure frequency extraction were prompt engineering and minimizing hallucinations. The problem of hallucinations – when LLMs generate plausible yet incorrect information – in clinical settings is explored at length in Pal et al., 2023, a study that found Llama 2's 70B parameter model performed well in one of its tests. The free-text EHRs were passed without modification to the LLM.

We found that the generally accepted default setting for the temperature, 0.7, was too 'creative' for our purposes, encouraging Llama 2 to generate overly colorful answers to our seizure frequency questions and, on occasions, even providing diagnostic advice, including medication prescriptions, for the person with epilepsy. In turn, this increased the false positives. On the other hand, we concluded that a minimal temperature of 0.0001 was sufficient for the model to generate typically fact-based answers without excessive creativity and helped reduce the false positives.

Three aspects of prompt engineering proved critical for usable output. First, few-shot prompting significantly improved Llama 2's ability to identify seizure frequency in an EHR, and proved much better than zero-shot prompting. However, we required 11 examples to give the model enough instructions on how to make complex decisions based on our nuanced nine categories of seizure frequency. Second, the characterization instructions in the prefix were a major factor in the model generating acceptable answers. Two key elements were instructing the model to act like a "professional neuroscientist who is responding to fellow neuroscientists" and to provide "succinct answers," the latter helping to eliminate verbosity. Third, we discovered the query was optimally structured by asking the model to logically work through three numbered steps to determine seizure frequency, as distinct from asking a single question (see Figures 2 and 3).

During initial iterations, we experimented with query structures that involved simpler instructions without an explicit logical sequence or numbered steps. The selection process involved group discussion evaluating the model's output from different variations of prompts, which in turn developed the optimal query structure. Of course, in the future improved prompt strategies and new LLMs may enhance the model outputs for extracting seizure frequency from EHRs, and this warrants further investigation.

The few-shot prompting examples provided Llama 2 with enough 'education' to generate answers that typically either matched, or closely resembled, our labels for the nine seizure frequency categories, thereby demonstrating the model's

ability to adapt its answers to idiosyncratic nomenclature. Of the 11 prompting examples, seven covered all but one of the temporal seizure frequency categories, two covered situations in which the patient did suffer seizures but the frequency of them was too difficult to determine from the EHR, and two covered situations in which the patient had not suffered seizures. We found during experimentation that doubling the last two kinds of prompts helped minimize hallucinations, or false positives. However, the model's answers were far from uniformly exact, as it often created its own versions of our category labels, so we devised an algorithm to interpret the model's answers if they either closely matched or were far from matching our labels. (See Appendix A for Llama 2 model architecture diagram.)

2.5 Annotation Dataset

The nine seizure frequency categories in the annotated dataset were dominated by unknowns, which comprised 71% of the 3,000 EHRs. In other words, only 29% of the annotated doctors' and nurses' reports contained any detectable information about seizure frequency. While some references to seizure frequency were clear and precise, especially if based on a patient's seizure diary, unfortunately many others were vague and imprecise. Consequently, the available data is sparse in regard to the core topic, which in turn makes the application of NLP to this task all the more challenging. Moreover, the number of observations in the higher frequency categories of seizure frequency – e.g., 'once or more per day' and 'more than once per week, less than once per day' – were roughly three times more common in the annotated dataset than those in the lower frequency categories (see Figure 1). This meant that Llama 2 found the lower frequency seizure categories more difficult to identify than the higher frequency categories.

2.6 BERT and RoBERTa: Model Development and Implementation

For a comparison to our Llama 2 method, we also fine-tuned BERT Large (Devlin et al., 2019) and RoBERTa Large (Liu et al., 2019) models on the annotated dataset, which was reduced from 3,000 EHRs to 1,720 EHRs to create a balanced dataset that was equally weighted between EHRs in which seizure frequency was known and EHRs in which seizure frequency was unknown. The unknown EHRs were reduced randomly to equal the 860 known EHRs. In turn, this reduced annotated dataset was restricted to the EHR text and the nine seizure frequency categories. Finally, it was split on an 80:10:10 ratio to create training, validation, and test datasets, respectively. We assume independent splits, a normal distribution, and a 95% confidence interval.

Both the BERT Large and RoBERTa models were used with PyTorch, an AdamW optimizer, threshold of 0.5 for the sigmoid, batch size of 4 (due to GPU memory limitations), and a learning rate of $1e^{-5}$. The optimal number of epochs varied for each model: 10 for BERT Large and 6 for RoBERTa Large. While the optimal dropout rate was 0.3 for BERT Large and 0.4 for RoBERTa Large. The maximum number of tokens for each EHR was set at 512, the upper limit for these two models.

3 Results

Our objective was to test an LLM against nine nuanced seizure frequency categories to determine how accurately they could identify seizure frequency from unstructured EHRs. The model F1 score for Llama 2 on the full annotated dataset of 3,000 EHRs was 0.73 and the model accuracy 0.94 (see Table 3), although the accuracy figure is misleading because it is boosted by a high number of true negatives, hence we prefer F1 as a measure of performance. We found that Llama 2 did well in identifying letters that had no or ambiguous information about seizure frequency, recording an F1 score of 0.87, and did moderately well on the most common known categories ('more than once a week', 0.35; and 'one or more daily', 0.41). But Llama 2 struggled with the remaining six temporal categories, ranging from 'once a week' to 'once a year' (see Table 2). Therefore, we aggregated the nine seizure frequency categories into three categories (infrequent, frequent, and unknown) to improve the performance of the model (see Table 4). Under the three categories, Llama 2 posted F1 scores of 0.87 for the unknowns, 0.62 for frequent seizures, and a lower 0.30 for infrequent seizures. Results are the average of three different runs of Llama 2. The LLM's output was highly consistent on each run, reflecting the low temperature of 0.0001 that in turn minimizes 'creativity' in answers.

	Seizure Frequency 9 Categories: F1 Score								
Model	**Once / year**	**Once / 6 months**	**> once / 6 months < once / month**	**Once / month**	**> once / month < once / week**	**Once / week**	**> once / week < once / day**	**1 or more / day**	**Unknown**
LlaMA 2 13B	0.11	0.06	0.17	0.42	0.36	0.06	0.35	0.41	0.87
RoBERTa Large	0.00	0.00	0.61	0.00	0.63	0.00	0.48	0.59	0.74
BERT Large	0.00	0.00	0.47	0.00	0.36	0.00	0.55	0.58	0.76

Table 2: Model performance evaluation on 9 seizure frequency categories.

Model	**Model F1 Score**	**Model Accuracy**
LlaMA 2 13B	0.73	0.94
RoBERTa Large	0.58	0.91
BERT Large	0.55	0.90

Table 3: Model performance for F1 and accuracy

	Seizure Frequency 3 Categories: F1 Score		
Model	**Infrequent**	**Frequent**	**Unknown**
LlaMA 2 13B	0.30	0.62	0.87
RoBERTa Large	0.43	0.76	0.74
BERT Large	0.39	0.77	0.76

Table 4: Model performance evaluation on 3 seizure frequency categories.

By comparison, Llama 2 performed better than BERT Large and RoBERTa Large, although it must be noted that our testing methodology for Llama 2 was different to that for the pre-trained Transformers. Llama 2 was mainly tested against the full annotated dataset of 3,000 EHRs (Llama 2 does not require fine-tuning), whereas BERT Large and RoBERTa Large, which required 80% of the balanced annotated dataset of 1,720 EHRs as a training dataset, were tested against a much smaller test dataset of 172 EHRs (or 10% of 1,720). Under this scenario, Llama 2's model F1 score of 0.73 was higher than RoBERTa Large's 0.58 and BERT Large's 0.55 (the results of the pre-trained Transformers are the average of three different runs with the same random states). Moreover, Llama 2 recorded a positive F1 score in all nine seizure frequency categories, whereas the pre-trained Transformers both posted F1 scores of zero in at least four categories, suggesting Llama 2 is better at identifying seizure frequency in the sparse categories.

However, we also tested Llama 2's performance on the same smaller test dataset of 172 EHRs used for BERT Large and RoBERTa Large. In this case, Llama 2's model F1 score dropped to 0.54, broadly in line with the pre-trained Transformers, and Llama 2 recorded F1 scores of zero in three of the nine seizure frequency categories. There are two possible explanations for this apparent difference in performance. First, the small test dataset represents only 6% of the full annotated dataset of 3,000 EHRs, therefore the latter is a better guide of actual model performance. Second, only 60% of EHRs in the small test dataset contained data on seizure frequency, and of those EHRs there was very little data on four categories ('once per week', 'once per month', 'once per six months', 'once per year'), therefore the paucity of data in the less common categories presented a greater challenge for the few-shot prompting structure for the LLM.

Furthermore, other metrics demonstrate that even when evaluated on the small test dataset, Llama 2 was more reliable than the other two models. Llama 2 predicted that 59% of the EHRs in the test dataset contained either no, or vague, information about seizure frequency, which was higher than the annotators' 40% but lower than RoBERTa Large's 65% and BERT Large's 71%. Also, while Llama 2 always made a prediction on every EHR, RoBERTa Large failed to make a prediction on average on 17% of the test EHRs and BERT Large failed on 30%.

It is difficult to compare the results of our study to those of Xie et al. (2023, 2022a, and 2022b) because they provided an "overall accuracy" score of 0.88 for seizure frequency and did not break down accuracy for individual seizure frequency categories. However, in broad terms in appears our Llama 2 methodology produced at least similar performance given its model accuracy was 0.94

and its accuracy rate for the infrequent category was 0.92 and for the frequent category 0.85.

4 Discussion

While our initial aim to determine whether the LLM could identify the frequency of seizures in unstructured outpatient reports for eight temporal categories proved too ambitious, when the temporal categories were reduced to frequent and infrequent the output of Llama 2 was much improved. The key objective of our broader epilepsy project is to track the effects of different combinations of anti-seizure medications on seizure frequency in individual patients and consequent changes. In this respect, Llama 2's F1 scores of 0.87 for the unknowns and 0.62 for frequent seizures is useful. Although the model's F1 score was a lower 0.30 for infrequent seizures, we are mindful that the number of observations of frequent seizures is roughly three times that of infrequent seizures, as previously stated, and so while more work is required to optimize the model's output for infrequent seizures, its overall performance aids our broader objective.

During experimentation it was clear that Llama 2's pre-training on vast general corpora had imbued it with a noticeable degree of expert knowledge about epilepsy. This may be one reason why Llama 2 proved superior to the pre-trained Transformers in identifying seizure frequency in unstructured, free-text EHRs. Another reason is that Llama 2 is a much bigger language model – we used the 13B parameter version – than BERT Large with 336M parameters and RoBERTa Large with 356M parameters.

Llama 2, like other generative LLMs, has three advantages over pre-trained Transformer language models. First, Llama 2 does not have to be fine-tuned on an annotated dataset, which saves substantial time and resources by obviating annotations for a training dataset. Second, Llama 2 does not have a built-in maximum token length for processing long texts. Third, Llama 2 is 'guided' on a particular language task by prompt engineering, which typically takes less time than adjusting multiple hyperparameters to optimize the performance of a pre-trained Transformer model.

On the other hand, Llama 2 has a distinct disadvantage: because of its large size, the LLM requires a longer running time. In this case, Llama 2 took on average 3.6 seconds to process one EHR, or about one hour for 1,000 EHRs.

A drawback of this particular study, however, is that the results are not reproduceable by other researchers because the patient EHRs are confidential and can only be accessed via the hospital's secure IT network.

5 Conclusion

Llama 2, as a popular LLM widely regarded as producing impressive performance on a variety of NLP tasks, performed well on the specific clinical NLP task of identifying seizure frequency from unstructured, free-text EHRs. This demonstrates that the new, generative LLMs are useful for epilepsy research in particular and clinical NLP research in general. The key question for our broader epilepsy research project was whether a new, generative LLM could identify seizure frequency among the EHRs to a sufficient degree to use the model's predictions as a basis for further research into different anti-epilepsy medications and their effects on seizure frequency. Our conclusion is that Llama 2 can.

Limitations

The confidential nature of the EHRs creates two limitations of this study. First, the model outputs are not reproduceable by research teams outside the hospital where the authors worked because the data has to remain within the hospital's secure IT network for regulatory reasons. Second, we could not experiment with LLMs such as OpenAI's ChatGPT that are only available via an API to an off-site service due to privacy reasons. However, with more time we could have experimented with other open-source LLMs. Another limitation is that, because of time and resource constraints, our annotation methodology of having six expert annotators working on separate batches of the 3,000, rather than having two annotators work on the same batch for moderation, did not allow for a measurement of inter-annotator agreement. Also, our research was also limited by the computing power generated by our GPU platform (eight Nvidia V100 GPUs). For example, this did not have the capacity to work with Llama 2's 70B parameter version on our dataset. Finally, the dataset of epilepsy patients from King's College Hospital may differ from datasets of epilepsy patients from other hospitals.

Ethical Considerations

The main ethical consideration was that the confidential EHRs of patients had to remain within the hospital's secure IT network. Therefore, researchers could only access the data and ingest the data into models via the hospital's IT network. Researchers and clinicians required clearance from the hospital. The project operated under London 593 South-East Research Ethics Committee (reference 18/LO/2048), approval granted to the King's 595 Electronic Records Research Interface (KERRI).

Acknowledgements

This work was supported by Epilepsy Research Institute UK (project reference 2209 Richardson Angelini) in conjunction with Angelini Pharma. The authors also wish to thank Angus Roberts and Joe Davies for their advice on early drafts, as well as the reviewers for the BioNLP Workshop for their constructive and helpful feedback.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1998-2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Brett K. Beaulieu-Jones, Mauricio F Villamar, Phil Scordis, Ana Paula Bartmann, Waqar Ali, Benjamin D Wissel, Emily Alsentzer, Johann de Jong, Arijit Patra, Prof Isaac Kohane. 2023. Predicting seizure recurrence after an initial seizure-like episode from routine clinical notes using large language models: a retrospective cohort study. *The Lancet Digital Health*, vol. 5, issue 12: pages e882-94.

Iz Beltagy, Matthew E. Peters, Arman Cohan. 2020. Longformer: The Long-Document Transformer. arXiv:2004.05150v2. Version 2.

Hyunmi Choi, Marla J. Hamberger, Heidi Munger Clary, Rebecca Loeb, Frankline M. Onchiri, GusBaker, W. Allen Hauser, and John B. Wong. 2014. Seizure frequency and patient-centered outcome assessment in epilepsy. *Epilepsia* 55(8): pages 1205-1212.

Barbara M. Decker, Alexandra Turco, Jian Xu, Samuel W. Terman, Nikitha Kosaraju, Alisha Jamil, Kathryn A. Davis, Brian Litt, Colin A. Ellis, Pouya Khankhanian, Chloe E. Hill. Development of a natural language processing algorithm to extract seizure types and frequencies from the electronic health record. *Seizure: European Journal of Epilepsy* 101 (2022): 48-51. doi.org/10.1016/j.seizure.2022.07.010

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (Google AI Language). 2019. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805v2. Version 2.

Eric van Diessen, Ramon A. van Amerongen, Maeike Zijlmans, Willem M. Otte. 2024. Potential merits and flaws of large language models in epilepsy care: A critical review. *Epilepsia* 00: pages 1-14. https://doi.org/10.1111/epi.17907.

Kirsten M. Fiest, Khara M. Sauro, Samuel Wiebe, Scott B. Patten, Churl-Su Kwon, Jonathan Dykeman, Tamara Pringsheim, Diane L. Lorenzetti, Nathalie Jetté. 2017. Prevalence and incidence of epilepsy: A systematic review and meta-analysis of international studies. *Neurology* 88(3): pages 296-303.

Beata Fonferko-Shadrach, Arron S Lacey, Angus Roberts, Ashley Akbari, Simon Thompson, David V Ford, Ronan A Lyons, Mark I Rees, William Owen Pickrell. Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the ExECT (extraction of epilepsy clinical text) system. *BMJ* Open. 2019 Apr 1; 9(4): e023232. doi: 10.1136/bmjopen-2018-023232.

Ben van Hout, Dennis Gagnon, Eric Souetre, Sibylle Ried, Claude Remy, Gus Baker, Pierre Genton, Herve Vespignani, and Pauline McNulty. 1997. Relationship Between Seizure Frequency and Costs and Quality of Life of Outpatients with Partial Epilepsy in France, Germany, and the United Kingdom. *Epilepsia* 38(11): pages 1221-1226.

Jason K. Hsieh, Francesco G. Pucci1, Swetha J. Sundar, Efstathios Kondylis, Akshay Sharma, Shehryar R. Sheikh, Deborah Vegh, Ahsan N. Moosa, Ajay Gupta, Imad Najm, Richard Rammo, William Bingaman, Lara Jehi. 2022. Beyond seizure freedom: Dissecting long- term seizure control after surgical resection for drug-resistant epilepsy. *Epilepsia*, vol. 64: pages 103-113.

P. Kwan and M.J. Brodie. 2000. Early identification of refractory epilepsy. *The New England Journal of Medicine* 342(5): pages 314-9.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Ming Liu, Richard Beare, Taya Collyer, Nadine Andrew, and Velandai Srikanth. 2023. Leveraging Natural Language Processing and Clinical Notes for Dementia Detection. In Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 150-155, Toronto, Canada. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical Domain Hallucination Test for Large Language Models. In Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), pages 314–334, Singapore. Association for Computational Linguistics.

Ryan Shea Ying Cong Tan, Qian Lin, Guat Hwa Low, Ruixi Lin, Tzer Chew Goh, Christopher Chu En Chang, Fung Fung Lee, Wei Yin Chan, Wei Chong Tan, Han Jieh Tey, Fun LoonLeong, Hong Qi Tan, WenLong Nei, WenYeeChay, David WaiMeng Tai, Gillianne Geet Yi Lai, Lionel Tim-Ee Cheng, Fuh Yong Wong, Matthew Chin Heng Chua, Melvin Lee Kiang Chua, Daniel Shao Weng Tan, Choon Hua Thng, Iain Bee Huat Tan, and HweeTouNg. 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): pages 1657-1664.

Arun Thirunavukarasu, Kabilan Elangovan, Darren Shu Jeng Ting, Laura Gutierrez, Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, vol. 29: pages 1930-1940.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288v2. Version 2.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:1706.03762v5. Version 5.

Anouk van Westrhenen1, Ben F. M. Wijnen, Roland D. Thijs. 2022. Parental preferences for seizure detection devices: a discrete choice experiment. *Epilepsia*, vol. 63: pages 1152-1163.

Kevin Xie, Brian Litt, Dan Roth, and Colin A. Ellis. 2022a. Quantifying Clinical Outcome Measures in Patients with Epilepsy Using the Electronic Health Record. In Proceedings of the 21st Workshop on Biomedical Language Processing, pages 369-375, Dublin, Ireland. Association for Computational Linguistics.

Kevin Xie, Ryan S. Gallagher, Erin C. Conrad, Chadric O. Garrick, Steven N. Baldassano, John M. Bernabei, Peter D. Galer, Nina J. Ghosn, Adam S. Greenblatt, Tara Jennings, Alana Kornspun, Catherine V. Kulick-Soper, Jal M. Panchal, Akash R. Pattnaik, Brittany Scheid, Danmeng Wei, Micah Weitzman, Ramya Muthukrishnan, Joongwon Kim, Brian Litt, Colin Ellis, Dan Roth. 2022b. Extracting seizure frequency from epilepsy clinic notes: a machine reading approach to natural language processing. *Journal of the American Medical Informatics Association*, 29(5): pages 873-881.

Kevin Xie, Ryan S. Gallagher, Russell T. Shinohara, Sharon X. Xie, Chloe E. Hill, Erin C. Conrad, Kathryn A. Davis, Dan Roth, Brian Litt, Colin A. Ellis. 2023. Long-term epilepsy outcome dynamics revealed by natural language processing of clinic notes. *Epilepsia* 64(7): pages 1900-1909.

Weipeng Zhou, Majid Afshar, Dmitriy Dligach, Yanjun Gao, and Timothy Miller. 2023. Improving the Transferability of Clinical Note Section Classification Models with BERT and Large Language Model Ensembles. In Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 125–130, Toronto, Canada. Association for Computational Linguistics.

Appendix A. Model Architecture Diagram for Llama 2