Fine-tuning LLMs to Extract Epilepsy Seizure Frequency Data from Health Records

Ben Holgate*, Joe Davies*, Shichao Fang*,
Joel S. Winston*, James T. Teo*, and Mark P. Richardson*
* Department of Basic & Clinical Neuroscience, King's College London
benjamin.holgate@kcl.ac.uk

Abstract

We developed a new methodology of extracting the frequency of a patient's epilepsy seizures from unstructured, free-text outpatient clinic letters by: first, devising a singular unit of measurement for seizure frequency; and second, fine-tuning a generative Large Language Model (LLM) on our bespoke annotated dataset. We measured frequency by the number of seizures per month: one seizure or more requires an integer; and less than one a decimal. This approach enables us to track whether a patient's seizures are improving or not over time. We found fine-tuning improves the F1 score of our best-performing LLM, Ministral-8B-Instruct-2410, by around three times compared to an untrained model. We also found Ministral demonstrated an impressive ability for mathematical reasoning.

1 Introduction

Extracting key patient data from longitudinal Electronic Health Records (EHRs) is critical to developing AI models that help improve patient treatments. Yet unstructured, free-text narratives are typically not suited to computational models that require structured data, and so medical researchers are increasingly utilizing Natural Language Processing (NLP) tools to enable clinical AI models to understand medical terminology and concepts (Yang et al., 2022).

In recent years, much clinical NLP research has focused on generative Large Language Models (LLMs). On the one hand, this has involved the development of LLMs with some degree of clinical expertise, such as ClinicalBERT (Huang et al., 2019), GatorTron (Yang et al., 2022), and ClinicalMamba (Yang et al., 2024). On the other hand, researchers have applied general knowledge LLMs to extract data from clinical texts (for example, Agrawal et al., 2022; Thirunavukarasu et al., 2023; and Zhou et al., 2023). In turn, this field of research has led to the creation of a benchmark, ClinicBench, to evaluate the performance of 22 LLMs in a clinical setting (Liu et al., 2024).

Yet the application of LLMs to epilepsy research is still relatively uncommon, although it is expected that this field will increase significantly in future (van Diessen et al., 2024). Epilepsy affects about 1% of the general population (Fiest et al., 2017) and contributes to an estimated half a percent of the global disease burden (WHO. Epilepsy. 2019). About 30% of people with epilepsy do not respond to anti-seizure medications (ASMs) and are therefore regarded as refractory to treatment (Kwan and Brodie, 2000). In the United Kingdom over the last decade, more than 30 individual ASMs have been available to prescribe and the number of possible combinations of ASMs taken as polytherapy is much larger. Consequently, it is not feasible to try all possible monotherapy and polytherapy options 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.

The most extensive relevant research on LLMs and epilepsy remains a long-term study (Xie et al., 2022a; Xie et al., 2022b; Xie et al., 2023; and Xie at al., 2024) that used a different methodology from ours to extract seizure frequency information from Electronic Health Records (EHRs). In their 2022-23 papers, 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 EHRs to determine the seizure frequency of a person with epilepsy or whether that person was seizure free. They declared an "overall accuracy" score of 0.88 for seizure frequency. In

Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025), pages 44–55

their 2024 paper, the team tested for bias (race, ethnicity, sex, income, and health insurance) in a ClinicalBERT model that they had fine-tuned on 700 manually annotated epileptologist notes and which classified whether a clinic note specified if a patient was seizure free or had recent seizures. They found no evidence of bias in the model.

Our previous, 2024 study was the first published paper to use a generative LLM to determine seizure frequency for people with epilepsy from unstructured, free-text EHRs (Holgate et al., 2024). We utilized Llama 2 13B (Touvron et al., 2023) to classify seizure frequency within eight temporal categories – ranging from once a year at one end of the spectrum to one or more per day at the other end – and in our analysis grouped the temporal categories into a binary split between infrequent and frequent seizures. We achieved an overall F1 score of 0.73 with Llama 2 13B.

An even more recent epilepsy study (Goldenholz et al., 2025) utilizes three different LLMs for different purposes: 1) Meta's Llama 2 13B to generate a randomized clinical trial for the ASM Cenobamate and generate 480 synthetic clinical notes; 2) Mistral's Mistral 7B v0.1 to summarize the clinical notes, specifically in regard to the number of seizures during the observation period and any symptoms associated with the ASM; and 3) Anthropic's Claude 2 to improve on the formatting and results of the data table. They used LLMs from different AI companies to ensure separation of technologies for the discrete tasks. Importantly, none of the LLMs were specially trained in medical language. The researchers concluded that their methodology demonstrated a capacity for inductive reasoning "from large sets of unstructured clinical encounters." Consequently, they recommended "a paradigm shift away from perfectly understanding the individual patient towards generalizable knowledge extracted from groups of patients. This new paradigm capitalizes on the strengths of LLMs … [while] acknowledging their weakness at high precision."

While we agree that LLMs hallucinate at individual patient level for seizure frequency, based on our experience, we disagree that they are not useful for micro analysis. On the contrary, our study demonstrates that some of the latest generative LLMs are, in fact, very good at estimating seizure frequency in unstructured, free-text EHRs based on our new methodology that incorporates a singular unit of measurement and fine-tuning.

2 Data and Methods

2.1 Data Collection

We selected 51,760 EHRs from King's College Hospital NHS Foundation Trust (KCH) that relate to 5,767 unique adult people with epilepsy being treated at KCH. The data spans more than a decade, from 1 January 2013 to 30 September 2023. The vast majority of the records comprise doctors' and nurses' reports of outpatients' ambulatory visits. 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 (see below for further annotation details).

2.2 Seizure Frequency Measurement

We followed the logic of the U Penn team to create a standardized format to denote seizure frequency in a given EHR. However, our methodology differed in two ways. First, the U Penn researchers used three language model pipelines with three different language models – for text classification, text extraction, and summarization of sentences with seizure frequency data – whereas we used only one generative LLM for all classification, extraction, and calculation tasks, largely because the newest LLMs are much more powerful than the ones they used. Second, the U Penn researchers initially used different time periods – day, month, year, or visit –depending on the period specified in the text, and then converted that by a rules-based quantifier into a standardized format of the number of seizures per month, whereas we required only

[1] https://cogstack.org/

one step by fine-tuning an LLM on our annotated dataset that denoted the text's data as the number of seizures per month.

Our project's lead data scientist annotated 1,480 EHRs in accordance with our singular unit of measurement for seizure frequency – that is, the number of seizures per month. The EHRs had previously undergone an initial annotation process. In our previous study (Holgate et al., 2024), we used stratified random sampling to select 3,000 EHRs to create an annotated dataset, which ensured proportional distribution across the original dataset for age, gender, and ethnicity to minimize bias. Subsequently, a team of six annotators, comprising four neuroscience clinicians (including two epileptologists) and two data scientists, manually annotated the 3,000 EHRs for 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. The annotators categorized seizure frequency into eight temporal frequencies – ranging from one seizure per year to one or more per day – plus 'unknown.' Due to time and resource limitations, the annotators worked on separate batches of the 3,000 EHRs, rather than having two annotators work on the same batch for moderation. However, the two epileptologists reconvened to create a 'gold standard' annotated dataset of 300 EHRs; their inter-annotator agreement was a Cohen's kappa score of 0.84, which signified near perfect agreement.

In turn, the lead data scientist used the 300 EHRs from this 'gold standard' annotated dataset plus a further 1,180 annotated EHRs to create a training and testing dataset to fine-tune LLMs on seizure frequency. The reason why the training / testing dataset was about half the size of the original annotated dataset was that about the same proportion of the KCH EHRs extracted contained information about a patient's seizure frequency. The lead data scientist converted the annotator's original annotation for seizure frequency to our new measurement system, in which one seizure or more per month required an integer, and less than one seizure per month a decimal (see Table 1). Two other categories were required for notation. If an EHR contained reference to seizures but the duration was unspecified or unclear, the number '1000' was used (essentially a proxy figure to denote incomplete information). Or if an EHR contained no reference to seizures, a '0' was used.

This methodology provided three key advantages: first, a single numerical metric makes it easy to track a patient's seizure trajectory over time (a declining number means the frequency of their seizures is reducing, while an increasing number means the frequency of their seizures is rising); second, a single numerical metric is easier to understand than eight, discrete temporal categories to record seizure frequency; and third, a single numerical metric is a more accurate and reliable input to feed into a seizure prediction model that we are developing as part of our wider epilepsy research project.

2.3 Model Development and Implementation

Environments and Models: 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 the four LLMs we experimented with in this study from Hugging Face and loaded the models into LangChain, which allowed us to perform multiple LLM operations in the local environment.[3] LangChain offers simple interfaces for loading and initializing LLMs.

We also employed parameter-efficient fine-tuning techniques, or PEFT, in particular parameter updates by low-rank adaptation, or LoRA. The latter hacks the regular backpropagation updates by splitting the update matrix into two smaller matrices which, when multiplied together, can give back the original update matrix. LoRA can accelerate training while reducing the computational demands.

We experimented with four LLMs that were released in 2024 or 2025 and developed by three different AI companies: US-based Meta's Llama 3.1 8B Instruct (Grattafiori et al., 2024); France-based Mistral's Mistral Nemo Instruct 2417 (Mistral AI Team, 2024a) and Ministral 8B Instruct 2410 (Mistral AI Team, 2024b); and China-based Alibaba's Qwen 2.5 7B Instruct (Yang et al., 2025). We were restricted to only using open-source language models because we used confidential

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

[3] https://huggingface.co

You are a professional neuroscientist.

Analyze the text and work through these 4 steps:

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

2. If the text does have information about the frequency of the patient's epilepsy seizures, then estimate the frequency of the seizures, and return the answer as the number of seizures per month.

3. If the text does refer to seizures but you cannot estimate the frequency of the seizures, then return the answer '1000'.

4. If the text does not have any information about the patient's epilepsy seizures, then return the answer '0'.

Figure 1: Prompt query structure.

medical data from the UK's National Health System (NHS) that had to remain within the hospital's secure IT network for regulatory reasons. We ran the LLMs on up to eight Nvidia V100 GPUs.

Pre-processing: We implemented two pre-processing elements. First, we found that an LLM's performance was slightly improved by reducing the length of each EHR, deleting non-relevant administrative information at the top and bottom of each clinic letter. As a result, this minimized noise from the unstructured text. We deleted all text before the clinic date at the top of the letter, and removed all text after the letter writer (typically a doctor or nurse) signed off "yours sincerely" (a UK letter writing convention) towards the end. In the event there was no specified date or sign-off, we set a default deletion of the first 40 characters and final 500 characters of each letter.

Second, we created a balanced dataset from the 1,480 annotated EHRs to train, test, and validate the LLMs. In each of the dataset's 1,480 observations, the input consisted of the EHR text, and the required output was the annotated decimal or integer for the corresponding seizure frequency, if stated in the document. A label for seizure frequency was assigned to the entire clinical note, based on the frequency for the patient at the time of the clinic visit. In other words. we fine-tuned the LLM on the annotated output. The balanced dataset was of various sizes, ranging from 375 to 813 EHRs in order to create training datasets ranging from 300 to 650 EHRs in increments of 50. The balanced dataset was structured by: taking a specified number of EHRs annotated with seizure frequency measurements of 0.1 to 999 (meaning these letters contained a reference to seizures with a specified frequency) and selected at random from the 1,480 annotated EHRs; then taking 25% of the number of the 0.1-999 category letters from the '1000' category letters, selected at random; and finally taking the same 25% portion from the '0' category letters, again selected at random. For example, 500 of the 0.1-999 letters were combined with 125 of the '1000' letters and 125 of the '0' letters to make a balanced dataset of 750 EHRs in total. The train/test/validation split was 80%/10%/10%. So in this example the training dataset consisted of 600 letters, the testing dataset 75 letters, and the validation dataset 75 letters. We use the term 'balanced' to mean that the dataset used to fine-tune the LLM was not weighted too far towards any of the three annotated categories. During experiments we found that this ratio of 25% of the total 0.1-999 letters for each of the '1000' and '0' letters worked best for adequately fine-tuning the LLMs on our seizure frequency task.

A fundamental challenge for this project was that the NHS EHRs used, mostly doctors' and nurses' reports of outpatients' ambulatory visits, were unstructured and typically noisy. The reports included a range of medical and administrative information, such as the patient's medication, other therapies, and details disclosed during previous clinic visits. Furthermore, the reports often did not include any information about seizure frequency and, if they did, the language was often imprecise, so that the nature of the frequency was vague or unclear. These factors make the application of LLMs to EHRs to research seizure frequency challenging.

Prompt Engineering: Although fine-tuning the LLM on hundreds of examples was the primary methodology in meeting this challenge, a secondary methodology was prompt engineering. We found that the structure of the prompt query made a difference to the quality of an LLM's answers. After experimentation, we concluded the optimal approach was Chain of Thought reasoning,

Seizure Frequency		Performance Evaluation	
Categories	**Measurement / Month**	**Purist Method**	**Pragmatic Method**
1 per year	0.08	$0 < x \leq 0.16$	
1 per 6 months	0.17	$0.16 < x \leq 0.18$	
> 1 per 6 months, < 1 per month	> 0.17, < 1	$0.18 < x \leq 0.99$	
1 per month	1	$0.99 < x \leq 1.1$	$0 < x \leq 1.1$
> 1 per month, < 1 per week	> 1, < 4	$1.1 < x \leq 3.9$	
1 per week	4	$3.9 < x \leq 4.1$	
> 1 per week, < 1 per day	> 4, < 30	$4.1 < x \leq 29$	
1 or more per day	30 - 999	$29 < x \leq 999$	$1.1 < x \leq 999$
Unknown frequency	1000	1000	1000
No seizure information	0	0	0

Table 1: Seizure frequency categories and measurements per month, performance evaluation methods.

asking the LLM to work through four logical steps, each of which was numbered (see Figure 1). The first step was to determine whether the EHR contained any information about the frequency of a patient's seizures (because often the letters did not). The second step asked the LLM to estimate the frequency as the number of seizures per month. The third step asked to return an answer of '1000' if the frequency of seizures of was too difficult to answer. The fourth and final step asked to return '0' if there was no information about seizures. At the start of the prompt, we asked the LLM to take on the role of a professional neuroscientist, as we found this slightly improved the quality of answers. We hypothesize that contextualizing the reasoning task for the LLM assists it in logically connecting the prompt (question) and text (EHR) with the relevant medical parts of the vast corpora that the LLM was originally trained on.

Hyperparameters: We kept the temperature at a very low 0.0001 (0 does not work for some LLMs) because we wanted the LLMs to generate typically fact-based answers and be consistent in their answers across multiple runs. In addition, our aim was to minimize both the LLMs' 'creativity' and hallucinations.

Although we experimented with changing some hyperparameters, such as the number of training epochs, batch size, and learning rate, we found none of these had any significant impact on the quality of the LLMs' answers. We set the number of epochs at three, the batch size at one, and the learning rate at 0.0002. In other words, the most influential factor in improving output was the size of the training dataset, followed by the prompt structure. For LoRA, we set the *r* value at 64, the alpha at 16, and the dropout rate at 0.1.

Post-processing: Despite fine-tuning the LLMs on our annotated dataset, the models' raw answers often needed to be cleaned up by a post-processing algorithm. The raw answers from the original model were typically variable, with a best-case answer being exactly what was asked by the prompt questions (e.g., '0', '2', or '1000'), a mixed answer (e.g., '11 to 16 seizures per month'), to outright nonsensical (e.g., '123456789' or 'He also showed some difficulties'). The raw answers from the fine-tuned LLMs were, however, generally more in line with what was required, typically generating an answer as either a decimal or integer with no (or little) text. Yet the LLM's construction – or attempt at construction – of a decimal was often confused with more than one decimal point (e.g., '2.00.0000'). As a consequence of the LLMs not being able to generate an answer in exactly the required format 100% of the time, we wrote a rules-based algorithm that either corrected the answer format where reasonably clear (e.g., '2.00.0000' becomes '2') or changed to a '0' if completely unclear (e.g., '123456789').

Model Selection: We began by running the four LLMs that we tested on different sized balanced datasets in order to create training datasets ranging from 300 to 650 EHRs in increments of 50, as outlined above. During fine-tuning each LLM was trained on the training dataset and also given separate evaluation and test datasets. At this stage we identified Mistral's two models as being the

best performing, followed by the Qwen 2.5 model, and the Llama 3 model. Overall, the best performing model was Ministral-8B-Instruct-2410.

We then tried various experiments to optimize the output of Ministral-8B-Instruct-2410. The most significant factors influencing the quality of the LLM's answers were the size of the training dataset (in general, more observations improved the answers) and the prompt structure. We determined that when the training dataset consisted of about 550 EHRs or more, the F1 score on our preferred method of evaluation reached about 0.80 or more.

3 Results

3.1 Performance Evaluation Methods

We used a confusion matrix to calculate recall, precision, the F1 score, and accuracy to evaluate an LLM's performance. We used a test dataset that each LLM had not seen during its training process. However, we devised two different methods of calculation, what we called the *purist* method and the *pragmatic* method. In the first method we used fuzzy logic, or the setting of soft (rather than hard) numerical boundaries between each of the eight temporal seizure frequency categories, on the basis that the temporal distinctions are arbitrary and our objective was to determine changes in a patient's seizure frequency over time.

The *purist* method set a high bar by calculating how well the LLM performed on eight temporal categories of seizure frequency. However, we treated this method more as a theoretical (rather than true) guide of performance, given the inconsistency of seizure information written by doctors and nurses in the outpatient letters, and the often inherent ambiguity of their language. Under this method, one seizure per year (specific target 0.08) equated to a range of $0 < x \leq 0.16$, one seizure per six months (specific target 0.17) was $0.16 < x \leq 0.18$, more than one seizure per six months but less than one per month (mid-point target ≈ 0.33) was $0.18 < x \leq 0.99$, one per month (specific target 1) was $0.99 < x \leq 1.1$, more than one seizure per month but less than one per week was $1.1 < x \leq 3.9$, one per week (specific target 4) was $3.9 < x \leq 4.1$, more than one per week but less than daily was $4.1 < x \leq 29$, and one or more per day was $29 < x \leq 999$ (999 being 1 below the 'fudge' figure of '1000'). In addition, we tested the model strictly against the other two categories: seizures with no information about frequency ('1000'); and no information about seizures ('0').

By contrast, the *pragmatic* method set a lower bar and reflected our broader objective to determine whether LLMs are good at extracting information about a patient's seizure frequency in such a way to reveal if their seizures are improving over time or not. In this method, we bifurcated the output into two temporal categories, infrequent and frequent seizures. Infrequent ranged from one seizure per year to one per month, which equated to a range of $0 < x \leq 1.1$. While frequent ranged from more than one per month to one or more per day, which equated to $1.1 < x \leq 999$. The two non-temporal categories remained as above. The threshold between infrequent and frequent had an empirical (rather than clinical) justification, in that our chosen demarcation line spread the number of observations in both categories more evenly, to avoid the frequent category significantly outweighing the infrequent category.

3.2 Model Performance

As shown in Table 2, the best-performing LLM, Ministral-8B-Instruct-2410, achieved its highest F1 score on the pragmatic method of 0.81 (purist method 0.68) with a training dataset of 650 EHRs, and a corresponding accuracy rate of 0.68 (0.52). As Appendix A illustrates, the F1 score on the pragmatic method rose beyond the 0.70 level once the training dataset became greater than 500 EHRs. While this might imply that the bigger the training dataset, the more effective the fine-tuning and the better the answers, this may not necessarily be the case. The F1 score dipped at 600 training observations but then rose to a new high at 650. Further research is required with even larger training datasets to investigate in more depth.

On the other hand, the results suggest that recall is not dependent on the size of the training data. Recall was consistently high, ranging from 0.86 to 1.00 on almost all training dataset sizes (with one exception). In other words, this Ministral model was proficient at correctly estimating seizure frequency.

By contrast, the results imply that precision is dependent on the size of the training dataset. The Ministral model required more than 500 training observations to improve precision – the same size needed to trigger an uplift in the F1 score. Nevertheless, precision remained the model's weak spot, achieving a best result of only 0.71 at 650

Fine-tuned LLM: Best F1 Scores

	Purist Method				Pragmatic Method			
LLM	**recall**	**precision**	**F1**	**acccuracy**	**recall**	**precision**	**F1**	**acccuracy**
Ministral-8B-Instruct-2410	0.91	0.54	0.68	0.52	0.93	0.71	0.81	0.68
Mistral-Nemo-Instruct-2407	1.00	0.48	0.65	0.48	1.00	0.64	0.78	0.64
Qwen2.5-7B-Instruct	0.60	0.31	0.47	0.32	0.71	0.62	0.66	0.51
Llama-3.1-8B-Instruct	0.20	0.35	0.26	0.22	0.22	0.39	0.28	0.23

Fine-tuned LLM: Mean Over 3 Runs and F1 Standard Deviation

	Purist Method				Pragmatic Method			
LLM	**recall**	**precision**	**F1 (SD)**	**acccuracy**	**recall**	**precision**	**F1 (SD)**	**acccuracy**
Ministral-8B-Instruct-2410	0.91	0.51	0.66 (0.02)	0.49	0.93	0.69	0.79 (0.01)	0.66
Mistral-Nemo-Instruct-2407	0.99	0.45	0.62 (0.05)	0.45	0.99	0.62	0.76 (0.02)	0.61
Qwen2.5-7B-Instruct	0.38	0.38	0.38 (0.09)	0.26	0.48	0.63	0.53 (0.12)	0.39
Llama-3.1-8B-Instruct	0.22	0.26	0.22 (0.05)	0.19	0.22	0.27	0.23 (0.05)	0.20

Table 2: Comparative performance evaluation of fine-tuned LLMs with same training dataset of 650 EHRs.

training observations on the pragmatic method, which was still comparatively low. This points to the model still 'hallucinating' on too many occasions, despite our attempts to minimize false positives through various techniques, in particular, fine-tuning, prompt engineering, setting a very low temperature, and adjusting the proportions of the balanced dataset.

The second-best performing LLM was the other Mistral model, Mistral-Nemo-Instruct-2407, which achieved a top F1 score of 0.78 on the pragmatic method, followed by Qwen2.5-7B-Instruct (0.66) and Llama-3.1-8B-Instruct (0.28) (see Table 2). Appendix B shows the comparative performance evaluation of the original LLMs -- that is, the non-fine-tuned models – which is much lower.

4 Discussion

Fine-tuning improved the F1 score of our best-performing LLM, Ministral-8B-Instruct-2410, by at least three times based on a training dataset of 650 EHRs. The F1 score of the fine-tuned model when evaluated by the purist method, 0.68, was three times that of the F1 score of the untrained model, 0.22. And the F1 score of the fine-tuned model when evaluated by the pragmatic method, 0.81, was 3.7 times that of the original model, also 0.22. This demonstrates that fine-tuning is an effective technique to improve the capacity of LLMs to identify the frequency of a patient's seizures in unstructured, free-text EHRs.

Both Mistral models performed at a high standard on this seizure frequency task, with only a 3 percentage points difference in their best F1 scores. However, there was a significant drop-off of 15 percentage points for the Qwen2.5 F1 score, and a 53 percentage points slide for the Llama 3.1 model, which did not perform well at all on this task.

Both Mistral models were also stable and consistent across multiple fine-tuning runs: their average F1 scores under the pragmatic method across three runs were only 2 percentage points below that of their respective top F1 scores; and the standard deviation of their F1 scores across 3 runs was only 1% or 2%. Stability is important in medical research. By contrast, Qwen2.5's F1 score was highly variable with a standard deviation of 12%.

Our study also demonstrates that some of the most recent LLMs have a capacity for mathematical reasoning. The Ministral models, in particular, were adept at identifying the frequency of a patient's seizures from the raw text, which could be anything from annually to daily or more, then converting that frequency to a standardized time period of per month, both in terms of decimals and integers. Indeed, Qwen2.5 was designed in part specifically to achieve "state-of-the-art performance" in mathematical tasks (Yang et al., 2025), and Llama 3's design had a partial focus on "mathematical reasoning performance" (Grattafiori et al., 2024), while the Mistral AI Team claims its Ministral 8B model achieves superior results to Llama 3.1 8B on a mathematical benchmark (Mistral AI Team, 2024b), which accords with our experience.

We can also postulate whether the LLMs we tested, especially the Ministral models, have some in-depth knowledge of medicine in general and

epilepsy in particular in their original, non-fine-tuned form. On the one hand, the comparatively low F1 scores of the original models compared to the much higher F1 scores of the fine-tuned models imply that may not be the case. On the other hand, the models' ability to quickly pick up the logic from the annotated training dataset to identify and calculate seizure frequency in a standardized format suggests it might be the case.

If the latter, it would support the findings of a recent study that tested three well-known LLMs – GPT-4, Bard, and Claude 2; admittedly not models that we used – on epilepsy practice examinations (Habib et al., 2024). These LLMs achieved mean scores of 72%, 65%, and 67%, respectively, compared to anecdotal reports suggesting the passing score for the examinations was approximately 70%.

"We found that LLMs scored well on the epilepsy practice examinations, did not appear to rely on memorization, and could logically explain the reasons for a correct answer," said the authors. "However, they occasionally hallucinated logic for incorrect answers." Their latter point matched our experience with too many false positives and a comparatively lower precision, even with our best-performing model and optimal training dataset.

Minimizing hallucinations in medical research is a common problem (Kim et al., 2025). Hallucinations are defined as responses from LLMs that are inaccurate or have fabricated information. This could affect clinical decisions and patient safety. Algorithms tend to hallucinate when providing answers to questions that have a high complexity, when there is insufficient or biased training data for a topic, or when a dataset is particularly noisy. All of these are common problems in medical research, especially with data collected from medical reports and diaries. Fine-tuning a general LLM is one way to mitigate these effects but it is not necessarily a complete solution (Zuo and Jiang, 2025). As a result, hallucinations may still occur after fine-tuning.

One possible solution is Retrieval Augmented Generation (RAG), which has gained popularity in medical contexts in recent years (Li et al., 2024; Halamka 2023). RAG involves taking a pre-trained LLM but not fine-tuning it. Instead, a prompt is given to the algorithm which then uses its training and augments it by looking up information from a corpus of documents, either from a public or private source. This can reduce the effect of hallucinations by essentially performing a cross-check. RAG warrants investigation in further research of our study.

5 Conclusion

Fine-tuning is an efficient method to optimize the extraction of seizure frequency data from unstructured, free-text medical records by LLMs. Moreover, we found that some of the most recent LLMs demonstrated an impressive ability for mathematical reasoning, in this case not only calculating the frequency of a patent's epilepsy seizures from a text, but also converting that calculation into a standardized temporal format of the number of seizures per month. Prompt engineering is also critical to fine-tuning an LLM for this task. However, hallucinations and the associated problem of too many false positives remain an issue, and further research is required here. Nevertheless, this study, by achieving an F1 score of 0.81 from our best-performing model, shows that fine-tuning an LLM provides a new and innovative way of extracting seizure frequency data from EHRs that in turn enables better analysis of the effects of ASMs in the treatment of epilepsy and therefore improved patient outcomes.

Limitations

This study has three main limitations. First, the confidential nature of the medical records used for the training dataset means the model outputs are not reproduceable by research teams outside the hospital where the authors worked. Second, the confidential records meant 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. Third, we were restricted in what sized LLMs we could use by the computing power generated by our GPU platform (eight Nvidia V100 GPUs).

Ethical Considerations

The confidential EHRs of patients had to remain within the hospital's secure IT network. As a consequence, the study's researchers could only access the data and input it into LLMs via the hospital's IT network.

Acknowledgments

This research project was funded by Epilepsy Research Institute UK (project reference 2209), an

independent research-funding charity. Epilepsy Research Institute UK received funding from Angelini Pharma to part-support this project. Angelini Pharma distributes Cenobamate in Europe and the UK. The study funders did not play any role in data collection, data analysis, or data interpretation, writing of the manuscript, or the decision to submit the manuscript for publication. The project operated under the London South-East Research Ethics Committee approval granted to the King's Electronic Records Research Interface (KERRI) (reference 18/LO/2048 and renewed 24/LO/0057).

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.

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.

Daniel M. Goldenholz, Shira R. Goldenholz, Sara Habib, M. Brandon Westover. 2025. Inductive reasoning with large language models: A simulated randomized controlled trial for epilepsy. *Epilepsy Research*, vol. 211. https://doi.org/10.1016/j.eplepsyres.2025.107532

Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783

Sara Habib, Haroon Butt, Shira R. Goldenholz, Chi Yuan Chang, Daniel M. Goldenholz. 2024. Large Language Model Performance on Practice Epilepsy Board Examinations. *JAMA Neurology* 81(6): 660–661. DOI: 10.1001/jamaneurol.2024.0676

John Halamka. 2023. Understanding Retrieval-Augmented Generation. Mayo Clinic Platform. https://www.mayoclinicplatform.org/2023/11/02/understanding-retrieval-augmented-generation/.

Ben Holgate, Shichao Fang, Anthony Shek, Matthew McWilliam, Pedro Viana, Joel S. Winston, James T. Teo, and Mark P. Richardson. 2024. Extracting Epilepsy Patient Data with Llama 2. In Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, pages 526–535, Bangkok, Thailand. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, Rajesh Ranganath. 2019. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv:1904.05342

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Liang, Xuhai Xu, Xin Liu, Daniel McDuff, Hyeonhoon Lee, Hae Won Park, Samir Tulebaev, Cynthia Breazeal. 2025. Medical Hallucination in Foundation Models and Their Impact on Healthcare. https://arxiv.org/pdf/2503.05777

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

Anson Li, Renee Shrestha, Thinoj Jegatheeswaran, Hannah O. Chan, Colin Hong, Rakesh Joshi. Mitigating Hallucinations in Large Language Models: A Comparative Study of RAG Enhanced vs. Human-Generated Medical Templates. https://www.medrxiv.org/content/10.1101/2024.09.27.24314506v1.full.pdf

Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. 2024. Large Language Models Are Poor Clinical Decision-Makers: A Comprehensive Benchmark. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 13696–13710, Miami, Florida, USA. Association for Computational Linguistics. DOI: 10.18653/v1/2024.emnlp-main.759

Mistral AI Team. 2024a. Mistral NeMo. https://mistral.ai/news/mistral-nemo

Mistral AI Team. 2024b. Un Ministral, des Ministraux. https://mistral.ai/news/ministraux

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.

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.

Kevin Xie, William K. S. Ojemann, Ryan S. Gallagher, Russell T. Shinohara, Alfredo Lucas, Chloé E. Hill, Roy H. Hamilton, Kevin B. Johnson, Dan Roth, Brian Litt, Colin A. Ellis. June 2024. Disparities in seizure outcomes revealed by large language models. Journal of the American Medical Informatics Association, Volume 31, Issue 6, Pages 1348–1355, https://doi.org/10.1093/jamia/ocae047

An Yang, et al. 2025. Qwen2.5 Technical Report. arXiv:2412.15115

Xi Yang, Aokun Chen, Nima PourNejatian, et al. 2022. A large language model for electronic health records. *npj Digital Medicine* 5:194. https://doi.org/10.1038/s41746-022-00742-2

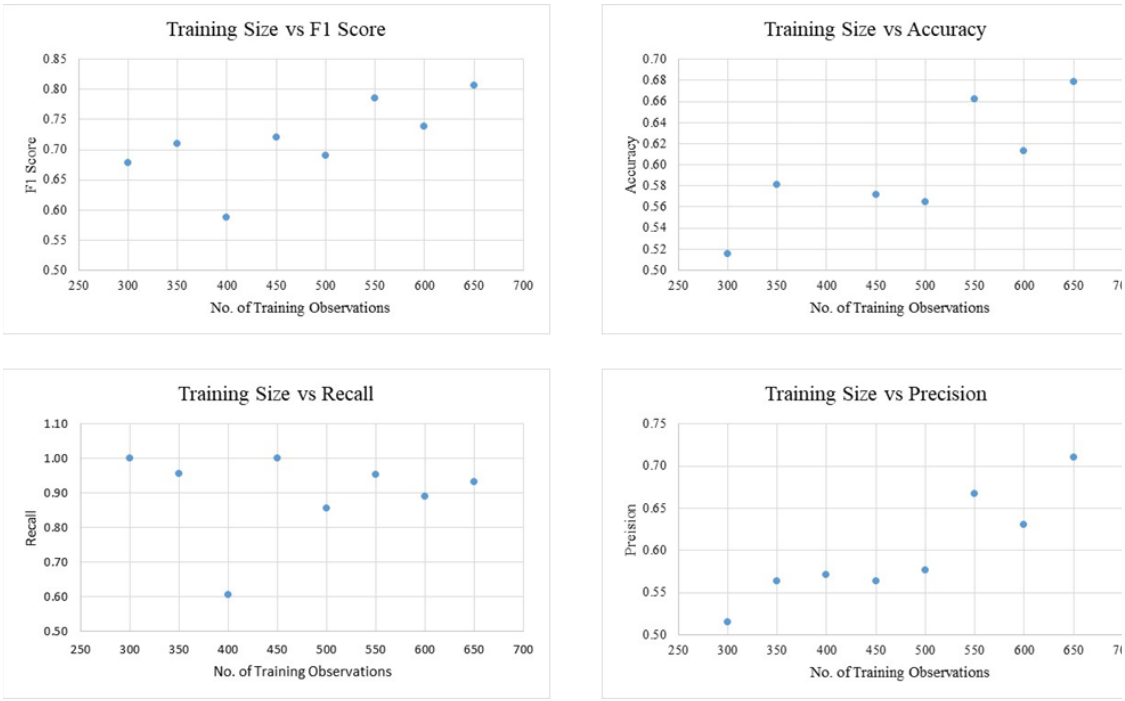
Zhichao Yang, Avijit Mitra, Sunjae Kwon, and Hong Yu. 2024. ClinicalMamba: A Generative Clinical Language Model on Longitudinal Clinical Notes. In Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 54–63, Mexico City, Mexico. Association for Computational Linguistics. DOI: 10.18653/v1/2024.clinicalnlp-1.5

WHO. Epilepsy. 2019. https://www.who.int/news-room/fact-sheets/detail/epilepsy.

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.

Kaiwen Zuo, Yirui Jiang. 2025. MedHallBench: A New Benchmark for Assessing Hallucination in Medical Large Language Models. https://arxiv.org/html/2412.18947v3

Appendix A



Appendix A: Ministral-8B-Instruct-2410 performance (pragmatic method) and size of training dataset.

Appendix B

Non-fine-tuned LLM: Best F1 Scores

	Purist Method				Pragmatic Method			
LLM	**recall**	**precision**	**F1**	**acccuracy**	**recall**	**precision**	**F1**	**acccuracy**
Ministral-8B-Instruct-2410	0.20	0.24	0.22	0.20	0.20	0.24	0.22	0.20
Mistral-Nemo-Instruct-2407	0.08	0.15	0.10	0.11	0.09	0.19	0.13	0.13
Qwen2.5-7B-Instruct	0.02	0.25	0.04	0.11	0.02	0.25	0.04	0.11
Llama-3.1-8B-Instruct	n/a	n/a	n/a	n/a	0.01	0.33	0.03	0.11

Non-fine-tuned LLM: Mean Over 3 Runs and F1 Standard Deviation

	Purist Method				Pragmatic Method			
LLM	**recall**	**precision**	**F1 (SD)**	**acccuracy**	**recall**	**precision**	**F1 (SD)**	**acccuracy**
Ministral-8B-Instruct-2410	0.20	0.24	0.22 (0.00)	0.20	0.20	0.24	0.22 (0.00)	0.20
Mistral-Nemo-Instruct-2407	0.08	0.15	0.10 (0.00)	0.11	0.09	0.19	0.13 (0.00)	0.13
Qwen2.5-7B-Instruct	0.02	0.25	0.04 (0.00)	0.11	0.02	0.25	0.04 (0.00)	0.11
Llama-3.1-8B-Instruct	n/a	n/a	n/a	n/a	0.01	0.33	0.03 (0.00)	0.11

Appendix B: Comparative performance evaluation of non-fine-tuned LLMs with same training dataset of 650 EHRs.

Note: Llama-3.1-8B-Instruct ‘n/a’ due to lack of true positives under purist method.