Using Synthetic Records to Improve Automated Identification of Seizure Freedom in Clinical Text about People with Epilepsy

Stephen H. Barlow[1], Yujian Gan[1,2], Joe Davies[1], Joel S. Winston[1], James T. Teo[1,3], Mark P. Richardson[1], Ben Holgate[1],
[1]Department of Basic & Clinical Neuroscience, King's College London, United Kingdom
[2]School of Electronics, Electrical Engineering & Computer Science, Queen's University Belfast, United Kingdom
[3]Neurosciences Institute, Cleveland Clinic London, United Kingdom
Correspondence: benjamin.holgate@kcl.ac.uk

Abstract

Seizure freedom is a key clinical outcome for people with epilepsy (PWE) yet it is primarily recorded in free-text notes and letters in the United Kingdom, making it difficult to aggregate and track at scale. This paper introduces a generative LLM-based pipeline boosted by synthetic data to identify a PWE's seizure freedom status in clinicians' records. We fine-tuned seven different large language models (LLMs) with between 4-14 billion parameters using LoRA to compare models trained on synthetic records against those trained on expert annotated records. The best performing configuration, based on Qwen-2.5-14B, was trained entirely on synthetic records and used chain-of-thought (CoT) reasoning (both generated by GPT-5). This achieved an F1 score of 0.90±0.02 on double-annotated test data and outperformed the equivalent model trained on authentic clinician records, which achieved 0.87±0.04. The synthetically trained models also have the benefit of outputting their CoT reasoning process for greater decision-making transparency and can also make use of unused supervised training data for significantly increased test examples. This work has implications for monitoring a key treatment outcome for PWE automatically and at scale.

1 Introduction

Epilepsy is a chronic neurological condition characterised by recurrent seizures (Beghi, 2019). Seizure freedom is a key outcome for improving quality of life in people with epilepsy (PWE) (Poochikian-Sarkissian et al., 2008) and an important post-surgical outcome (Wieser et al., 2001). However, its definition remains inconsistent (Halford and Edwards, 2020). Engel's (1993) post-operative classification prioritises reduction of "disabling seizures", placing less emphasis on minor seizures, while Wieser et al. (2001) proposed a stricter definition requiring complete seizure freedom without auras, aligning more closely with PWE's expectations. Clinically, seizure freedom also requires a temporal threshold, as PWE may experience seizure recurrence after years without them. For example, the UK Driving and Vehicle Licensing Agency (DVLA) requires 12 months of seizure freedom for licence reinstatement (DVLA, 2026). Other studies have used six- and twelve-month thresholds (Hsieh et al., 2023; Choi et al., 2014). Thus, seizure freedom is both a definitional and temporal concept.

In the UK, difficult-to-treat epilepsy is managed in outpatient (ambulatory) settings, where consultation summaries (outpatient letters) are recorded as unstructured or semi-structured free text. This format limits automated information extraction. Advances in natural language processing (NLP), particularly large language models (LLMs), have enabled extraction of complex contextual information from electronic health records (EHRs) (Jerfy et al., 2024; Jin et al., 2025), including temporality (Yuan et al., 2024), which is essential for determining seizure freedom.

Accordingly, this study develops an LLM-based seizure-freedom classifier trained entirely on synthetic epilepsy letters which provides generated explanations. We also examine how training data origin (synthetic vs. authentic), class distribution, and synthetic chain-of-thought reasoning affect model performance. While this paper explores an epilepsy-specific application, we believe this approach has the potential to improve other imbalanced clinical information extraction tasks.

2 Related Work

Previous studies have applied NLP to extract seizure freedom-status from epilepsy records. Xie et al. (2022) used BERT-style models (Devlin et al., 2019) to extract seizure freedom and frequency, defining seizure freedom as no seizures since the last clinic visit or for at least 12 months.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 20–30

Due to model input length limits, they used paragraph-level classification. Fernandes et al. (2024) found performance declined without reliable paragraph selection, and later work by Xie et al. (2023) introduced a bespoke paragraph extractor to aggregate document-level predictions. However, seizure freedom is inherently a document-level construct, true or false at the time of writing. Modern generative LLMs with longer context windows could perform direct document-level classification without rule-based passage selection, which could in turn better resolve contradictory statements. Related work on seizure frequency also reflects these broader NLP developments, progressing from rule-based methods (Fonferko-Shadrach et al., 2019) to encoder-only models (Xie et al., 2022), in-context learning (Holgate et al., 2024), and parameter-efficient fine-tuning for document-level approaches (Holgate et al., 2025).

These studies usually rely on relatively small annotated datasets, reflecting expert time constraints and limiting representation of rarer contexts. Data imbalance is intrinsic, as seizure-free PWE attend fewer appointments. Synthetic data offers a potential solution, enabling larger training datasets while reserving expert-annotated data for evaluation. Teacher–student approaches, where large LLMs generate synthetic training data for smaller models, are a potential solution. Phi-4 (Abdin et al., 2024), for example, was trained primarily on synthetic GPT-4 output. Synthetic data has also addressed healthcare challenges like Šuvalov et al. (2025) creating an Estonian clinical named entity recognition system using GPT-generated records. However, it remains unclear whether fully synthetic training can capture temporal clinical concepts like seizure freedom. Additionally, approaches such as DeepSeek's synthetic reasoning chains (Guo et al., 2025; Liu et al., 2025) may enhance transparency by addressing clinician concerns about "black box" decisions (Salvi et al., 2025) by providing interpretable reasoning in model output.

Synthetic clinical text has also been explored in epilepsy where Goldenholtz et al. (2025) used LLM-generated notes to simulate a clinical trial. This study was entirely *in silico* and thus highlights the need to evaluate synthetic training against real-world clinician-authored notes, which this study addresses.

3 Methods

Seizure freedom (SF) classification is framed in this work as a binary classification task using generative language modelling. We compare authentic and synthetic training datasets to examine the effect on performance for classifying seizure-freedom. For the synthetic training approach, we also compare how additional fine-tuning on CoT reasoning-chains affects performance and outputs. For the authentic letters we also tested whether under sampling the majority class resulted in better performance due to the pronounced class imbalance in this task.

3.1 Seizure Freedom Definition

Seizure freedom is considered an important outcome of epilepsy treatment, yet definitions differ. We sought to find a compromise between a the stricter 12-month definition of seizure freedom (DVLA, 2026; Choi et al., 2014) and Xie et al.'s (2022) last clinic approach. Hsieh et al.'s (2023) six-month definition of seizure freedom provided a clinically relevant outcome while providing enough training and testing examples (as 12-months seizure free was much less frequently seen in our dataset than six-months). Accordingly, in this work seizure freedom is defined as:

> *"The person with epilepsy has suffered no epileptic seizures of any kind for a period of six months or more."*

Figure 1 shows some artificial examples of ambiguous examples of SF-positive and SF-negative letters by this definition. These characteristics informed the need to use LLMs with longer context windows (the number of text tokens an LLM can process at a time) as relevant information could be in any part of (sometimes very long) letters, and a degree of temporal reasoning is required to ascertain the correct SF-status.

3.2 Authentic Dataset

We randomly sampled and annotated for SF status 2,114 epilepsy letters from King's College Hospital (London, UK) with a team of five epilepsy experts. These were retrieved using the information retrieval system CogStack (Jackson et al., 2018). The letters were written between 1 January 2013 and 30 September 2023. The documents comprise doctor's and nurse's reports with the vast majority referring to outpatient clinic visits. A

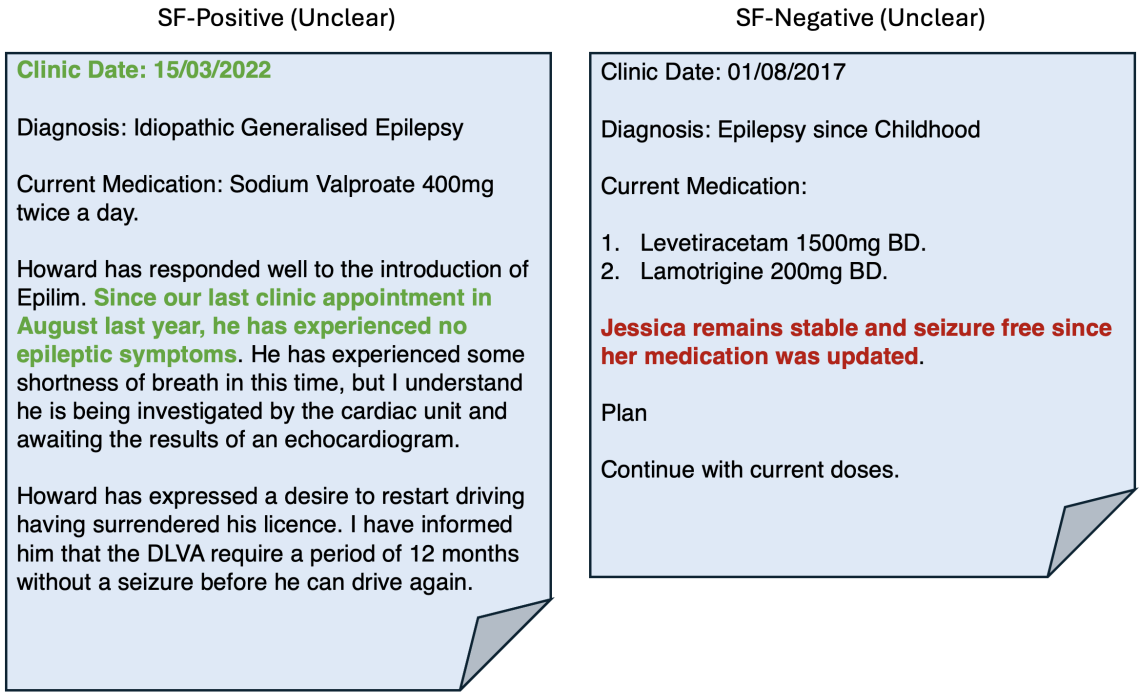

Figure 1: Fictional clinician letters for epilepsy demonstrating ambiguous SF-positive and SF-negative status. Letters can be much longer and more detailed in practice. Both examples show how temporal reasoning is required to make the correct classification (either relating information to provided dates, or recognising there is missing temporal information). The negative example demonstrates how we consider all examples that are not explicitly seizure free (for six months or more) as negative examples.

subset of 300 reports were dual annotated independently by epileptologists to test inter-annotator agreement and subsequently a gold-standard test set was produced via consensus between the experts. The overall agreement for seizure freedom was 0.85 using Cohen's Kappa, representing 'near perfect' (Cohen, 1960) or 'strong' (McHugh, 2012) agreement. SF-positive status was observed to be rare in the hospital letters, around 6% of the total. Therefore, due to limited resources for annotation, the bulk of the letters were annotated by one person. This motivated both the creation of the dual-annotated subset (for greater confidence in evaluation) and the need to create synthetic letters for training.

This project and its data usage operated under UK Health Research Authority (HRA) London South East Research Ethics Committee approval (reference18/LO/2048 and renewed 24/LO/0057) granted to the King's Electronic Records Research Interface (KERRI) with data research opt-out. Individual consent for participation was waived by the KERRI committee at King's College Hospital (KCH) for purposes of evaluating NLP for Epilepsy (approved February 2021) with local institutional oversight.

3.3 Synthetic Dataset

The aforementioned class imbalance motivated the use of synthetic training examples. As part of a wider synthetic data project 15,099 synthetic epilepsy letters were generated by GPT-5 to reflect a range of different seizure frequency and seizure freedom occurrences in clinical epilepsy letters (Gan et al., 2026). These were created from 10 manually composed letters (to mimic a UK style of epilepsy reporting) and 74 templates of different letter characteristics (different seizure frequencies, no seizures mentioned, seizure free etc.). Once these letters were generated, they were fed back to GPT-5 where it was prompted to ascertain the correct seizure frequency/seizure freedom-status and provide a description of how it came to that conclusion. This functions as a quality-assurance step for the synthetic data as any letters that GPT-5 does not correctly categorise were discarded. From these explanations 454 synthetic letters were found that were correctly described by GPT-5 as implying SF-positive status. These were extracted along with 454 SF-negative examples to create a balanced synthetic training dataset. This approach mitigates the class imbalance issue present in a true clinical distribution of

letters. Two variations on this synthetic dataset were trialled: Synthetic and Synthetic-CoT. For the latter we include the reasoning chains generated by GPT-5 as part of the output for supervised training to test if it improves model performance. This also has the benefit of providing text explanations for the classifications, something that could potentially assist clinicians (Miao et al., 2024). This project respects OpenAI's Terms of Use. This study is conducted for academic, non-commercial research purposes. We do not release distilled model weights and do not position the distilled model as a deployable substitute for OpenAI services; results are reported for methodological analysis.

3.4 Dataset Splits, Evaluation, and Prompting

We define four training datasets: Authentic, Authentic-Balanced, Synthetic and Synthetic-CoT. The Authentic training set uses 1,814 annotated letters (all but the 300 double-annotated letters). The Authentic-Balanced dataset uses the 115 SF-positive examples that are not included in the gold-standard test set alongside a matching 115 SF-negative examples. This matches the distribution of the Synthetic and Synthetic-CoT training datasets which contain 454 SF-positive and 454 SF-negative examples.

We used two test datasets. The gold-standard test set consisted of the 300 examples with consensus annotation from two experts and both authentic and synthetic models used this for testing. The silver-standard test set consisted of all 2,114 annotated letters and was used for additional testing of the synthetically trained models. This allowed for a better indication of long-term performance. The authentic models were not tested on this as they were explicitly trained on all but the 300 gold-standard examples in this dataset.

As seizure freedom extraction is framed as a binary classification task, we used precision, recall, F1 score, and accuracy for evaluation. Due to the pronounced class imbalance, we prioritised macro-averaged F1 over accuracy to compare the different models. This ensured we evaluated both the majority and minority classes equally and avoided making decisions due to deceptively high accuracy scores caused by the class imbalance.

The prompt used for training and inference is demonstrated in Figure 2. The additional supervision signal of GPT-5 reasoning chains were prepended to the desired output in the Synthetic-CoT dataset. Figure 2 also demonstrates how models trained on the Synthetic-CoT dataset generate temporal reasoning before making a final classification decision. The other training approaches used the same input prompt but only provided a 'Seizure free: yes/no' output.

Base Model	Parameters	Type
Gemma-3	4B	VLM
MedGemma	4B	VLM
Qwen-2.5-7B	7B	LLM
Lingshu	7B	VLM
Llama-3.1-Instruct	8B	LLM
Ministral-Instruct-2410	8B	LLM
Qwen-2.5-14B	14B	LLM

Table 1: Comparison of the LLMs used in this study. 'B' represents 'Billion'.

3.5 LLMs and Training Approach

We selected seven open-weight LLMs with between 4 billion to 14 billion parameters for fine-tuning on the SF extraction task (Table 1). Some of these were vision-language models (VLMs), but only their language capabilities were utilised for this task. The selection represents both general and clinically specialised models and were chosen as they represented a diverse selection of different size LLMs with different pre-training corpora and model architectures. We also trained two simpler, logistic regression classifiers using term frequency-inverse document frequency (TF-IDF) encodings to serve as baselines, one on the 'Authentic' dataset, and one on the balanced 'Synthetic' dataset.

Early experiments revealed that in-context learning was not sufficient for adequate performance on this task so fine-tuning was performed using low-rank adaptation (LoRA) (Hu et al., 2021). Due to the extra VRAM (Video Random Access Memory) Qwen-2.5-14B required for fine-tuning, we used 8-bit quantization using Bitsandbytes (Dettmers et al., 2021). The optimisation task was autoregressive language modelling, where the model is trained to predict the next token of text from previous text. The appropriate instruction templates and tokenizers were used for each model tested.

No specific hyperparameter tuning took place in order to compare synthetic and authentic approaches without biasing training towards either

Chain-of-Thought Fine-Tuning Prompt

Input:

"You are a professional neuroscientist. The following is a document written by a healthcare professional. Analyze the text and ascertain whether the patient referred to in the letter has been seizure free for six months or more. Think through the problem step-by-step. Some things to consider include: Ascertain if the letter contains information about seizures, ascertaining if they have had recent seizures, and if they have not had seizures for six months or more. Finally, if you think the subject has been seizure free for six months or more return 'Seizure free: yes', otherwise return 'Seizure free: no'.

Document:

[LETTER]"

Output:

"The letter explicitly states: "She reports ongoing absence of seizures since February 2023." The letter date is 02 October 2025. Calculate duration of seizure freedom from Feb 2023 to Oct 2025: from Feb 2023 to Feb 2024 = 12 months; Feb 2024 to Feb 2025 = 12 months; Feb 2025 to Oct 2025 = 8 months. Total = 32 months. This exceeds 6 months, so she is seizure free, and the duration is multiple years. The allowed format for multi-year seizure freedom is 'seizure free for multiple year' when exact number in years isn't required. ***Seizure free: yes"***

Figure 2: The main prompt used for the Synthetic-CoT training strategy. The Authentic, Authentic-Balanced and Synthetic training datasets used the same prompt but "Seizure free: yes/no" only as the desired output. [LETTER] signifies where the authentic or synthetic letter would be inserted into the prompt.

one. All models were trained for three epochs with a batch size of four. We used a learning rate of 0.0001 with a cosine schedule and a warmup ratio of 0.1. LoRA rank, alpha and dropout were set at 8, 16 and 0 respectively. Greedy decoding was used for text generation to ensure a deterministic answer. For synthetic data generation GPT-5's temperature was set (via API) to 0.

4 Results

Table 2 shows the distribution of SF-positive and SF-negative examples in the different datasets used for training and testing. The extracted clinical data (Authentic, Authentic-Silver and Authentic-Gold) demonstrated a pronounced class imbalance that is rectified by under sampling in the Authentic-Balanced, Synthetic, and Synthetic-CoT training datasets. There were only 21 SF-positive letters in the gold standard dataset but this is increased to 136 in the silver standard test set, demonstrating the benefit of having additional annotated data available for testing.

Figure 3 shows how synthetic and authentic training datasets affect performance on the gold-standard test set. Synthetic training tended to perform as well, if not better, than authentic training. For authentic training, using a balanced dataset did not help and, in some cases, seemed to hinder the model. The use of CoT reasoning chains in training improved performance on certain models but not others. All the LLM-based approaches dramatically outperformed the TF-IDF/logistic regression model which only achieved an F1 score of 0.48 (in contrast to F1 scores of 0.74-0.91 for the LLM-based pipelines).

Figure 4 explores the performance of synthetically trained models further on the larger silver-standard test set. This provides a more thorough demonstration of synthetically trained model performance with 2,114 test examples as compared to 300 (Table 2). It was observed again that benefits for training with CoT reasoning chains were dependent on the individual base LLM. The Qwen models seemed to find the additional CoT reasoning supervision beneficial on the larger test dataset with Qwen-2.5-14B model's F1 score improving to 0.85 from 0.84, and Qwen-2.5-7B improving to 0.84 from 0.82. The best performing model was Qwen-2.5 14B with CoT reasoning and this was also a strong performer on the gold standard subset with an F1 score of 0.88 (Figure 3). In

Dataset	Use	No. Letters	SF-Positive	SF-Negative
Authentic	Training	1814	115	1699
Authentic-Balanced	Training	230	115	115
Synthetic	Training	908	454	454
Synthetic-CoT	Training	908	454	454
Authentic-Silver	Synthetic Test	2114	136	1978
Authentic-Gold	Synth and Authentic Test	300	21	279

Table 2: The different training and test sets used in this work and their distribution of SF-positive and negative examples.

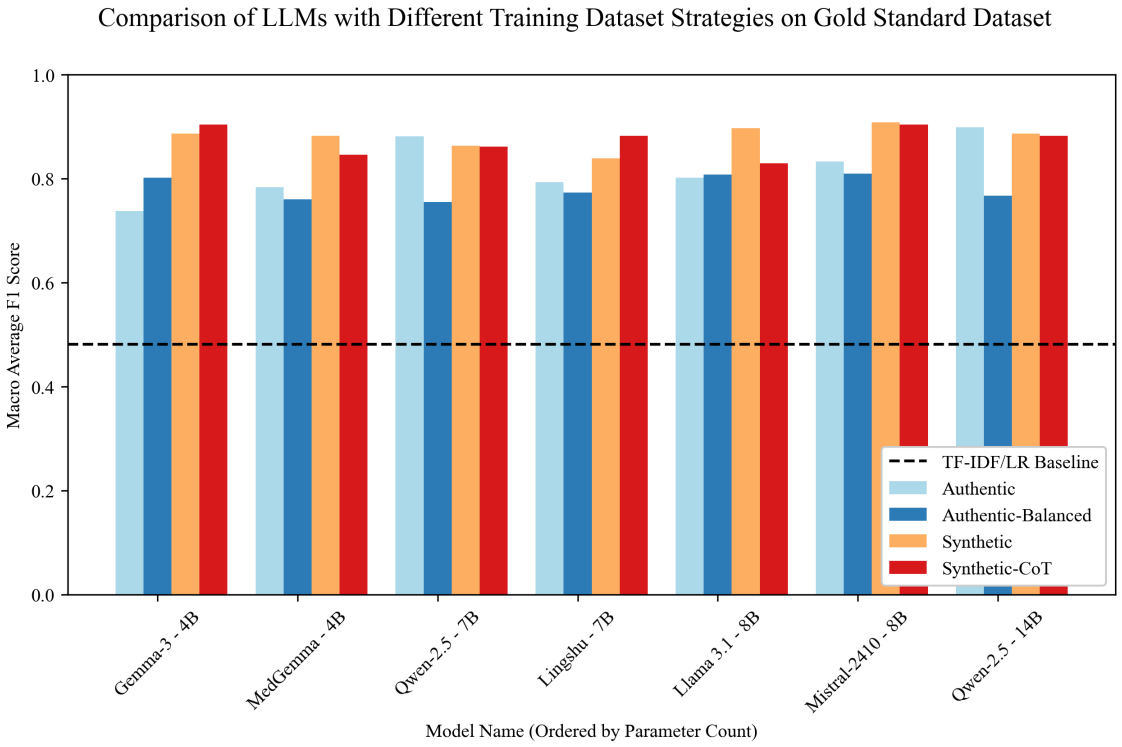

Figure 3: Comparison of the different LLMs trialled when trained on four different dataset strategies. Results are on the gold-standard test set. A TF-IDF logistic regression classifier (trained on the authentic training dataset) is provided as a baseline.

contrast, LLama 3.1-8B performed better without CoT reasoning supervision with its F1 score performance on the silver standard test set dropping to 0.77 from 0.83. The synthetically-trained TF-IDF/logistic regression baseline performs poorly and only achieves an F1 score of 0.24.

As Qwen-2.5-14B was the best performing LLM, when taking into account results on both the gold and silver test sets, Table 3 shows its performance in more detail when trained three times with different random seeds. The Synthetic-CoT training set resulted in the best performing model pipeline on both the gold and silver-standard datasets with F1 scores of 0.90±0.02 and 0.85±0.02 respectively.

5 Discussion

Our best model configuration for classifying seizure freedom status, using Qwen-2.5-14B and the Synthetic-CoT training strategy, achieved F1 scores of 0.90±0.02 on the gold standard test and 0.85±0.02 on the larger silver standard test dataset. The strong performance on both suggests it will be robust to different examples when used in practice. This model was trained entirely on synthetic data using a CoT reasoning approach for training. This means it generates the LLM-reasoning steps the model took to come to a final answer as part of the output. It is difficult to directly compare our results with Xie et al. (2022; 2023) and Fernandes et al.'s (2024) approaches to seizure freedom classification as the models were trained on different datasets and use different definitions of seizure freedom. However, our best model configuration compares favourably to the 0.83-0.88 F1 scores reported by Xie et al. (2022; 2023) (who also had a gold-standard test set of 300 examples) and 0.38 reported by Fernandes et al. (2024). The lower performance of the latter was attributed to using Xie et al.'s (2022) model without fine-tuning on their own letters. This demonstrates the issue of model generalization to external data. The goal of our research was to maximise performance for a stringent definition of seizure freedom rather than tackle broader seizure frequency concerns. Seizure freedom is an important outcome in epilepsy treatment so ensuring that we automatically identify PWE who have had at least six months of seizure freedom at a point in time was the priority. Although we explore a specific clinical use case in this paper, we believe the results of these experiments are also evidence to support using synthetic training data for other medical NLP tasks.

A key finding from this work is that a successful seizure freedom classifier can be trained using only synthetically generated training data. This provides two key benefits. It allows all expert annotated examples to be used for testing, and reduces privacy concerns by not using any sensitive data for fine-tuning. The latter makes it impossible for the model to disclose this data from leakage (Zhong et al., 2025), or adversarial attack (Rahman, 2023). Our best overall model fine-tuned Qwen-2.5-14B on synthetic letters and CoT reasoning. The Qwen 2.5 models were trained with an emphasis on reasoning (Yang et al., 2024), which is a potential reason they benefited from the CoT training signal, particularly on the silver-standard test set (Figure 4). Llama 3.1, in contrast, was not specifically optimised for reasoning which could explain why performance worsened for this LLM using CoT training signal. We identify two potential reasons for how a synthetically trained model can outperform an authentically trained one. Firstly, the increased training examples that the synthetic approach allows provides greater supervision signal. Secondly, there are likely less ambiguous synthetic examples (as these have been verified by GPT-5) so the model is less likely to overfit to them. These factors could also interact, as an ambiguous authentic example will have a disproportionately large effect due to the smaller number of training examples overall.

Synthetic data has been proposed as a solution to data bottlenecks when training healthcare models (McDuff et al., 2023). However, it is non-trivial creating examples which replicate authentic documents faithfully (Smolyak et al., 2024). Epilepsy letters are a good case study for synthetic data approaches as concepts like seizure freedom and frequency require dates (clinic dates, medication changes, etc.) to remain in letters for training. This is because dates and sentences referring to them may be the only reference to seizure status in the text. This can be considered private information and some de-identification models will remove them by default. Synthetic examples overcome this limitation by allowing student models to train on temporal relationships without jeopardising patient privacy. The reasoning chain demonstrated in Figure 2 shows how both the teacher and student models make use of this

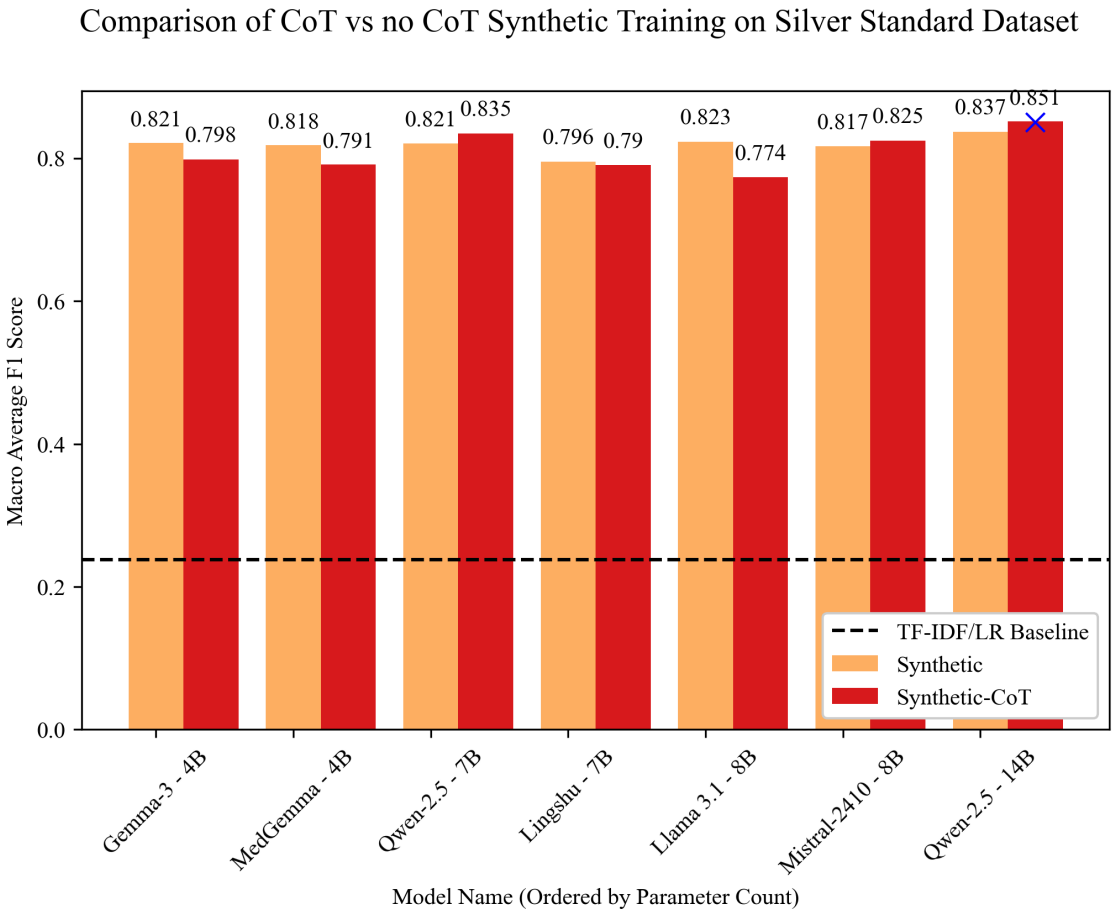

Figure 4: Comparison of Synthetic and Synthetic-CoT training strategies on the silver-standard test set. The blue cross represents the best performing model by macro average F1 score. Exact F1-score values are stated above the bars. A TF-IDF logistic regression classifier (trained on the synthetic training dataset) is provided as a baseline.

Training Set	Test Set	Accuracy	F1	Precision	Recall	Support
Authentic	Gold	0.972±0.009	0.874±0.049	**0.952±0.011**	0.824±0.062	300
Auth-Balanced	Gold	0.926±0.015	0.793±0.026	0.739±0.026	**0.909±0.012**	300
Synthetic	Gold	0.969±0.002	0.88±0.007	0.883±0.025	0.881±0.028	300
Synth-CoT	Gold	**0.973±0.007**	**0.896±0.023**	0.904±0.037	0.89±0.016	300
Synthetic	Silver	0.954±0.007	0.829±0.014	0.799±0.027	0.87±0.013	2114
Synth-CoT	Silver	***0.961±0.004***	***0.847±0.019***	***0.826±0.016***	***0.873±0.03***	2114

Table 3: Performance of Qwen-2.5-14B when LoRA fine-tuned with the four different training strategies on both test sets. Each model was trained three time with three random seeds, and the mean and standard deviation is reported. F1, precision, and recall metrics are macro-averaged. Support signifies the number of examples in the test sets. Bold text represent the best performing on the gold-standard test set, and bold plus italics represents the best performing on the silver-standard test set.

temporal information to arrive at a classification. By using a high-quality reasoning model as the teacher model, like GPT-5, we demonstrate that student models can learn to generate reasoning chains that justify the classification of SF-status. This can not only improve performance of the model (depending on the student LLM used), but also provides a human interpretable explanation for the clinician, something that has been identified as important step towards building confidence in AI solutions for healthcare (Miao et al., 2024). The reasoning chains also provided another benefit by serving as a verifier for quality when creating the synthetic letters. This combination of synthetic letters with CoT reasoning appears to be a powerful combination for developing epilepsy-specific systems where temporal features are directly related to patient outcomes.

The synthetic training data approach also has advantages when it comes to evaluation. Many classification models are tested on a small number of examples (Hou et al., 2024), which reduces the support for any classification metrics presented. Not requiring annotations for training means all annotations can go towards verifying the performance of models. The synthetically trained models in this project were evaluated on seven times the number of samples than was possible for the authentically trained models, providing greater confidence in them.

Although the largest model tested was the most effective, it should be noted that all the models tested reliably extracted SF-status from reports (Figures 3 and 4). Many hospitals do not have extensive computational resources, so having a smaller, 4 billion parameter LLM as a viable option is encouraging for clinical translation. It is also worth noting that the smaller models seemed to particularly benefit from the additional training examples provided by the synthetic letters. This is illustrated in Figure 3 where Gemma and MedGemma performed comparatively poorly when trained on authentic letters compared to synthetic letters (on the gold-standard test set). As multimodal AI becomes more prevalent, it was also encouraging to see that the VLMs (Gemma, MedGemma, and Lingshu) did not perform noticeably worse than LLMs when operating using text only. This suggests there is little drawback in training foundation models for multimodality outside of data resource and compute constraints.

For future work, it would be useful to apply synthetic-only training approaches to other clinical NLP tasks. Identifying people with rare phenotypes (for both epilepsy and other conditions) in electronic health records would be a valuable test for this approach. We also plan to implement the best performing model from this project as a feature extractor for an epilepsy prognostic tool, where a PWE's seizure freedom status is likely to have predictive value.

6 Conclusion

We developed an LLM-based pipeline to classify PWE's clinical reports for seizure freedom status. A configuration using Qwen-2.5-14B trained on a dataset consisting entirely of synthetically generated clinician letters from GPT-5 with CoT reasoning chains performed the best. This model offered key advantages over the traditional supervised approach. Most notably the CoT from the teacher model, GPT-5, can be used as supervision signal and allow the student model, to generate CoT text for better performance and greater transparency in its answers. Also, not needing annotated data for training allowed the model to be evaluated on a much larger test set. This pipeline has the potential to be used for monitoring the progress of PWE at scale and determining if medication changes or other interventions result in this key clinical outcome being achieved.

Limitations

There were three main limitations. Firstly, we were unable to perform external testing for the authentic models due to only having annotations from one hospital. Secondly, the silver-standard test set was only single-annotated and although we found these annotations to be reliable, there could be more annotation mistakes in this test set when compared to the gold-standard test set. Finally, these are document-level classification models, and there are cases where relevant temporal information could be spread across multiple documents which this method will not take into account.

Acknowledgements

This document was produced as part of a research project funded by the Epilepsy Research Institute (grant ref 2209) and by an investigator-initiated research grant from Angelini Pharma.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, and Piero Kauffmann. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Ettore Beghi. 2019. The epidemiology of epilepsy. *Neuroepidemiology*, 54(2):185–191.

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

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2021. 8-bit optimizers via block-wise quantization. *ArXiv*, abs/2110.02861.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

DVLA. 2026. Epilepsy and driving. `https://www.gov.uk/epilepsy-and-driving` [Accessed: 2026-02-27].

Jerome Engel, Jr., P.C. Van Ness, T. Rasmussen, and L.M. Ojemann. 1993. Outcome with respect to epileptic seizures. In *Engel, J. Jr (Ed.), Surgical Treatment of the Epilepsies*, pages 609–621. Raven Press, New York.

Marta Fernandes, Aidan Cardall, Lidia M. V. R. Moura, Christopher McGraw, Sahar F. Zafar, and M. Brandon Westover. 2024. Extracting seizure control metrics from clinic notes of patients with epilepsy: A natural language processing approach. *Epilepsy Research*, 207:107451.

Beata Fonferko-Shadrach, Arron S. Lacey, Angus Roberts, Ashley Akbari, Simon Thompson, David V. Ford, Ronan A. Lyons, Mark I. Rees, and William Owen Pickrell. 2019. 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*, 9(4):e023232.

Yujian Gan, Stephen H. Barlow, Ben Holgate, Joe Davies, James T. Teo, Joel S. Winston, and Mark P. Richardson. 2026. Reproducible synthetic clinical letters for seizure frequency information extraction. *Preprint*, arXiv:2603.11407.

Daniel M. Goldenholz, Shira R. Goldenholz, Sara Habib, and M. Brandon Westover. 2025. Inductive reasoning with large language models: A simulated randomized controlled trial for epilepsy. *Epilepsy Research*, 211:107532.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, and Xiao Bi. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jonathan J. Halford and Jonathan C. Edwards. 2020. Seizure freedom as an outcome in epilepsy treatment clinical trials. *Acta Neurologica Scandinavica*, 142(2):91–107.

Ben Holgate, Joe Davies, Shichao Fang, Joel Winston, James Teo, and Mark Richardson. 2025. Fine-tuning llms to extract epilepsy seizure frequency data from health records. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 44–55.

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.

Rui Hou, Joseph Y. Lo, Jeffrey R. Marks, E. Shelley Hwang, and Lars J. Grimm. 2024. Classification performance bias between training and test sets in a limited mammography dataset. *PLOS ONE*, 19(2):e0282402.

Jason K. Hsieh, Francesco G. Pucci, Swetha J. Sundar, Efstathios Kondylis, Akshay Sharma, Shehryar R. Sheikh, Deborah Vegh, Ahsan N. Moosa, Ajay Gupta, Imad Najm, Richard Rammo, William Bingaman, and Lara Jehi. 2023. Beyond seizure freedom: Dissecting long-term seizure control after surgical resection for drug-resistant epilepsy. *Epilepsia*, 64(1):103–113.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Richard Jackson, Ismail Kartoglu, Clive Stringer, Genevieve Gorrell, Angus Roberts, Xingyi Song, Honghan Wu, Asha Agrawal, Kenneth Lui, Tudor Groza, Damian Lewsley, Doug Northwood, Amos Folarin, Robert Stewart, and Richard Dobson. 2018. Cogstack - experiences of deploying integrated information retrieval and extraction services in a large national health service foundation trust hospital. *BMC Med Inform Decis Mak*, 18(1):47.

Aadit Jerfy, Owen Selden, and Rajesh Balkrishnan. 2024. The growing impact of natural language processing in healthcare and public health. *INQUIRY:*

The Journal of Health Care Organization, Provision, and Financing, 61:00469580241290095.

Myeong Jin, Sang-Min Choi, and Gun-Woo Kim. 2025. Comcare: A collaborative ensemble framework for context-aware medical named entity recognition and relation extraction. *Electronics*, 14(2).

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, and Chen Dong. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Daniel McDuff, Theodore Curran, and Achuta Kadambi. 2023. Synthetic data in healthcare. *arXiv preprint arXiv:2304.03243*.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22(3):276–82.

Jing Miao, Charat Thongprayoon, Supawadee Suppadungsuk, Pajaree Krisanapan, Yeshwanter Radhakrishnan, and Wisit Cheungpasitporn. 2024. Chain of thought utilization in large language models and application in nephrology. *Medicina*, 60(1):148.

Sonia Poochikian-Sarkissian, Souraya Sidani, Richard Wennberg, and Gerald M. Devins. 2008. Seizure freedom reduces illness intrusiveness and improves quality of life in epilepsy. *Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques*, 35(3):280–286.

Mohamed. A. Rahman. 2023. A survey on security and privacy of multimodal llms - connected healthcare perspective. In *2023 IEEE Globecom Workshops (GC Wkshps)*, pages 1807–1812.

Massimo Salvi, Silvia Seoni, Andrea Campagner, Arkadiusz Gertych, U. Rajendra Acharya, Filippo Molinari, and Federico Cabitza. 2025. Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare. *International Journal of Medical Informatics*, 197:105846.

Daniel Smolyak, Margrét V Bjarnadóttir, Kenyon Crowley, and Ritu Agarwal. 2024. Large language models and synthetic health data: progress and prospects. *JAMIA open*, 7(4):ooae114.

Hendrik Šuvalov, Mihkel Lepson, Veronika Kukk, Maria Malk, Neeme Ilves, Hele-Andra Kuulmets, and Raivo Kolde. 2025. Using synthetic health care data to leverage large language models for named entity recognition: Development and validation study. *J Med Internet Res*, 27:e66279.

H. G. Wieser, W. T. Blume, D. Fish, E. Goldensohn, A. Hufnagel, D. King, M. R. Sperling, H. Lüders, and T. A. Pedley. 2001. Ilae commission report. proposal for a new classification of outcome with respect to epileptic seizures following epilepsy surgery. *Epilepsia*, 42(2):282–6.

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 H. Scheid, Danmeng Wei, Micah Weitzman, Ramya Muthukrishnan, Joongwon Kim, and 3 others. 2022. Extracting seizure frequency from epilepsy clinic notes: a machine reading approach to natural language processing. *Journal of the American Medical Informatics Association*, 29(5):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, and Colin A. Ellis. 2023. Long-term epilepsy outcome dynamics revealed by natural language processing of clinic notes. *Epilepsia*, 64(7):1900–1909.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Back to the future: Towards explainable temporal reasoning with large language models.

Xiaoying Zhong, Siyi Li, Zhao Chen, Long Ge, Dongdong Yu, Shijia Wang, Liangzhen You, and Hongcai Shang. 2025. Considerations for patient privacy of large language models in health care: Scoping review. *J Med Internet Res*, 27:e76571.