Knowledge Tree Driven Contextualized Instruction Tuning of Foundation Models for Epilepsy Drug Recommendation

Duy Khoa Pham[1,9], Deval Mehta[1,2](✉), Yiwen Jiang[1,2], Daniel Thom[3], Richard Shek-kwan Chang[3], Mohammad Nazem-Zadeh[3], Emma Foster[3], Timothy Fazio[4,5], Sarah Holper[6], Karin Verspoor[7], Jiahe Liu[1], Duong Nhu[1], Sarah Barnard[3], Terence O'Brien[3], Zhibin Chen[3], Jacqueline French[8], Patrick Kwan[3], and Zongyuan Ge[1,2]

[1] AIM for Health Lab, Faculty of IT, Monash University, Melbourne, Australia
[2] Faculty of Engineering, Monash University, Melbourne, Australia
[3] School of Translational Medicine, Monash University, Melbourne, Australia
[4] Clinical Informatics Centre, Royal Melbourne Hospital, Melbourne, Australia
[5] Department of Medicine, University of Melbourne, Melbourne, Australia [6] Royal Melbourne Hospital, Melbourne, Australia
[7] School of Computing Technologies, RMIT University, Melbourne, Australia
[8] NYU Langone Health, New York City, New York, United States
[9] Department of Computing Technologies, Swinburne University of Technology, Melbourne, Australia
deval.mehta@monash.edu

Abstract. Epilepsy affects over 50 million people worldwide, with anti-seizure medications (ASMs) as the primary treatment for seizure control. However, ASM selection remains a "trial and error" process due to the lack of reliable predictors of effectiveness and tolerability. While machine learning approaches have been explored, existing models are limited to predicting outcomes only for ASMs encountered during training and have not leveraged recent biomedical foundation models for this task. This work investigates ASM outcome prediction using only patient MRI scans and reports. Specifically, we leverage biomedical vision-language foundation models and introduce a novel contextualized instruction-tuning framework that integrates expert-built knowledge trees of MRI entities to enhance their performance. Additionally, by training only on the four most commonly prescribed ASMs, our framework enables generalization to predicting outcomes and effectiveness for unseen ASMs not present during training. We evaluate our instruction-tuning framework on two retrospective epilepsy patient datasets, achieving an average AUC of **71.39** and **63.03** in predicting outcomes for four primary ASMs and three completely unseen ASMs, respectively. Our approach improves the AUC by **5.53** and **3.51** compared to standard report-based instruction tuning for seen and unseen ASMs, respectively. Our code, MRI knowledge tree, prompting templates, and TREE-TUNE generated instruction–answer tuning dataset are available at the link.

Keywords: instruction tuning · epilepsy · drug recommendation

1 Introduction

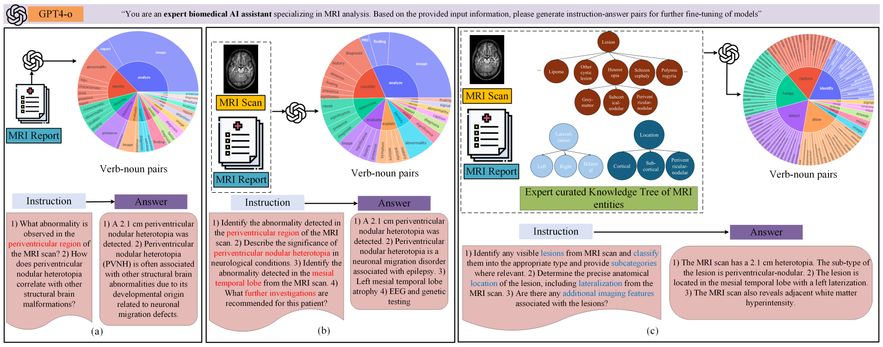
Epilepsy is one of the most common neurological disorders, affecting over 50 million people worldwide [26]. Antiseizure medications (ASMs) are the first-line treatment for newly diagnosed patients, aiming to achieve seizure freedom, yet only about half of newly diagnosed patients achieve seizure control with their initial ASM [3]. ASM selection primarily depends on epilepsy type, clinical history and diagnostic findings. However, since ASMs exhibit similar efficacy for specific epilepsy types [19], ASM selection often follows a "trial and error" approach, requiring sequential trials of different ASMs if seizures persist, which results in multiple ineffective treatments before identifying the right ASM [4].

The epilepsy research community has explored machine learning (ML) and deep learning (DL) for ASM recommendation by identifying patterns linking patient health data to ASM outcomes [5]. Early studies used simple ML models such as kNN and random forests to predict the response to ASM based on genetic data [21,23]. Recent works have incorporated the clinical history, demographics, and risk factors of patients, employing SVM, XGBoost, and MLP to predict their ASM outcomes [6,11]. More recently, researchers have tapped into applying ML for analyzing routine electroencephalogram (EEG) data to predict the most effective ASM to achieve seizure free outcome [8,22].

The International League Against Epilepsy recommends the use of Magnetic Resonance Imaging (MRI) for newly diagnosed epilepsy patients, as it can detect structural lesions that can guide treatment decisions and epilepsy management. Early MRI findings help clinicians determine whether to escalate ASM trials or consider adjunct therapies like lesion resection [13]. Lesions like focal cortical dysplasia and mesiotemporal sclerosis have well-documented links to clinical outcomes and recognized patterns of drug responsiveness or resistance [1]. Research on MRI-based ASM outcome prediction has primarily relied on manually extracted features and simple ML models or CNNs, limiting their predictive power [12,14,24]. Moreover, existing studies operate in a closed-set setting, simply encoding ASMs as categorical input variables during training, which limits their ability to predict outcomes for unseen ASMs.

To the best of our knowledge, no prior work has explored recent biomedical foundation models [15,28] for ASM outcome prediction. These models, trained on large-scale biomedical images and text, have demonstrated strong medical context understanding. However, they have struggled with custom and unseen downstream tasks, diverse demographics, and open-set questions [20, 27]. To address this, the broader vision and natural image research communities have adopted instruction tuning [2,16] as an efficient fine-tuning approach to improve model performance for custom downstream tasks.

Our contributions: In this work, we explore ASM outcome prediction using MRI scans and reports. Our main contributions are: 1) **Foundation Model Adaptation-** We adapt biomedical vision-language foundation models to assess their efficacy for the task of ASM outcome prediction; 2) **Knowledge Tree Driven Contextualized Instruction Tuning (TREE-TUNE)-** To enhance the performance of foundation models, we construct a contextualized instruction-



"You are an **expert biomedical AI assistant** specializing in MRI analysis. Based on the provided input information, please generate instruction-answer pairs for further fine-tuning of models"

(a) Instruction: 1) What abnormality is observed in the periventricular region of the MRI scan? 2) How does periventricular nodular heterotopia correlate with other structural brain malformations?

(a) Answer: 1) A 2.1 cm periventricular nodular heterotopia was detected. 2) Periventricular nodular heterotopia (PVNH) is often associated with other structural brain abnormalities due to its developmental origin related to neuronal migration defects.

(b) Instruction: 1) Identify the abnormality detected in the periventricular region of the MRI scan. 2) Describe the significance of periventricular nodular heterotopia in neurological conditions. 3) Identify the abnormality detected in the mesial temporal lobe from the MRI scan. 4) What further investigations are recommended for this patient?

(b) Answer: 1) A 2.1 cm periventricular nodular heterotopia was detected. 2) Periventricular nodular heterotopia is a neuronal migration disorder associated with epilepsy. 3) Left mesial temporal lobe atrophy 4) EEG and genetic testing

(c) Instruction: 1) Identify any visible lesions from MRI scan and classify them into the appropriate type and provide subcategories where relevant. 2) Determine the precise anatomical location of the lesion, including lateralization from the MRI scan. 3) Are there any additional imaging features associated with the lesions?

(c) Answer: 1) The MRI scan has a 2.1 cm heterotopia. The sub-type of the lesion is periventricular-nodular. 2) The lesion is located in the mesial temporal lobe with a left laterization. 3) The MRI scan also reveals adjacent white matter hyperintensity.

Fig. 1. Instruction-tuning approaches: (a) MRI report-based (b) MRI scan and report-based (c) **Proposed TREE-TUNE**: Knowledge Tree with MRI scan and report-based. Includes verb-noun pairs and an instruction-answer sample for each approach.

tuning dataset. Using an expert-built MRI knowledge tree, we categorize MRI reports and scans, generating data-specific instruction-answer pairs with GPT-4o for improved fine-tuning; 3) **Molecular Feature Integration-** To enable outcome predictions for unseen ASMs, we encode ASMs as Simplified Molecular Input Line Entry System (SMILES) [25] representations, enabling models to learn molecular features instead of treating ASMs as categorical variables.

We benchmark our approach on two retrospective epilepsy datasets, achieving a **5.53** AUC improvement over standard text-based instruction tuning for four primary ASMs. Notably, our model attains **63.03** AUC for predicting outcomes of three unseen ASMs it has never encountered during training.

2 Proposed Method

Our approach has two components: (1) Constructing a contextualized instruction-tuning dataset and (2) Fine-tuning foundation models on the built dataset.

2.1 Construction of Contextualized Instruction Tuning Dataset

Knowledge Tree of MRI entities : Our in-house neurologists and clinicians developed a hierarchical knowledge tree to categorize and annotate entities in MRI reports. This taxonomy is derived from clinically and systematically validated anatomical locations, lesion types, and imaging features [10]. The tree has three primary nodes: *Element* – Includes types of lesions, malformations, tumors, and vascular abnormalities; *Location* – Represents the anatomical positioning of elements; and *Other Features* – Covers imaging characteristics such as atrophy, calcification, mass effect and others. The tree has a depth of 4 levels from the root with *Element* being the most complex, comprising four primary subcategories and over 50 distinct subtype elements. *Location* includes three subcategories with 15 distinct positions, while *Other Features* covers 14 distinct imaging attributes. A partial view of the knowledge tree is shown in Fig 1(c), with the full tree available in our code repository **here**.

Contextualized Data-specific Generation of Instruction-Answer Pairs Traditional instruction-answer pair generation follows two strategies: 1) *Text based generation*, where LLMs like GPT-4 generate synthetic pairs from captions or reports [15] (Fig 1(a)). 2) *Multimodal generation*, where both images and text are provided to multimodal LLMs (e.g., GPT-4v) for enhanced quality [2] (Fig.1(b)). In contrast, our method **TREE-TUNE** (Fig. 1(c)) enhances instruction generation by incorporating three inputs into GPT-4o: an MRI scan, its corresponding report, and the MRI knowledge tree. GPT-4o, prompted as an "Expert Biomedical AI Assistant specializing in MRI analysis" (temp = 0.1), then generates contextualized instruction-answer (IA) pairs based on observed MRI entities from scan and report and their hierarchical placement in our built knowledge tree. Our TREE-TUNE generated 4,896 IA pairs across 274 patient cases (median: 24 IA pairs per patient; average length of instruction: 49.1 words).

For example, if an MRI report mentions "periventricular nodular heterotopia" and "white matter hyperintensity", the knowledge tree links them to their nodes: *Lesion→Heterotopia→Periventricular-nodular* and *Other Features* respectively (Fig 1(c)). This enables GPT-4o to generate more nuanced instructions (highlighted in blue in Fig 1(c)), producing more context-aware answers than standard prompting, which embeds terms directly into the instructions (highlighted in red in Fig 1(a),(b)). Similarly, presence of a lesion in the "left mesial temporal lobe" in the MRI scan links it to *Location→Lateralization* nodes, resulting in knowledge-informed instructions and more case-specific responses (compare question 3) and question 2) from Fig 1(b) and (c)).

By integrating the knowledge tree with MRI scans and reports, **TREE-TUNE** enhances instruction specificity and contextuality. The verb-noun pair distribution (Fig. 1) shows that our approach yields a richer noun set per verb than standard methods, effectively capturing MRI entities and generating more refined instructions, as further analyzed in the Results section.

2.2 Instruction-Tuning of Foundation Models

To evaluate the impact of our contextualized instruction-tuning dataset, we finetune foundation models following LLaVA [17] (Fig 2). Given an MRI scan X_I, a pretrained biomedical vision encoder extracts visual features $Z_I = g(X_I)$, which are projected into the language embedding space H_I via a trainable matrix W_I. The ground truth ASM, X_{ASM}, is represented in SMILES format [25] and processed as $Z_{ASM} = d(X_{ASM})$ using the pretrained graph-based drug encoder MoLeR [18], which models molecular structures as graphs and encodes the atoms, bonds, and other structural features of the ASM drug molecule. ASM embeddings Z_{ASM} are then projected into the language space H_{ASM} using another trainable matrix W_{ASM}.

For each MRI-ASM pair, we generate a contextualized instruction-answer set using the knowledge tree, MRI scans, and reports. These pairs $\{(X_q^1, X_a^1),$

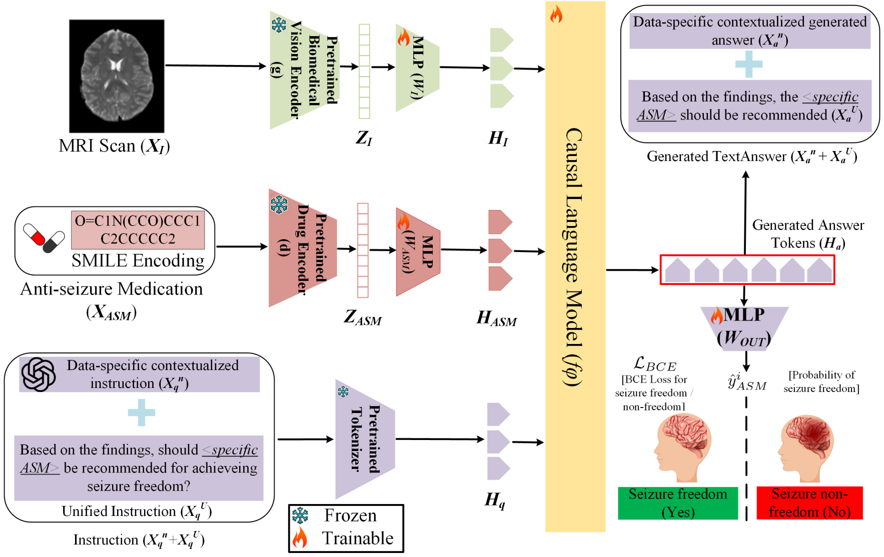

MRI Scan (X_I)
Pretrained Biomedical Vision Encoder (g)
Z_I
MLP (W_I)
H_I
O=C1N(CCO)CCC1 C2CCCCC2
SMILE Encoding
Anti-seizure Medication (X_{ASM})
Pretrained Drug Encoder (d)
Z_{ASM}
MLP (W_{ASM})
H_{ASM}
Data-specific contextualized instruction (X_q'')
Based on the findings, should <specific ASM> be recommended for achieveing seizure freedom?
Unified Instruction (X_q^U)
Instruction ($X_q''+X_q^U$)
Pretrained Tokenizer
H_q
Frozen
Trainable
Causal Language Model ($f\varphi$)
Data-specific contextualized generated answer (X_a'')
Based on the findings, the <specific ASM> should be recommended (X_a^U)
Generated TextAnswer ($X_a''+X_a^U$)
Generated Answer Tokens (H_a)
MLP (W_{OUT})
$\mathcal{L}_{BCE}$
[BCE Loss for seizure freedom / non-freedom]
$\hat{y}^i_{ASM}$
[Probability of seizure freedom]
Seizure freedom (Yes)
Seizure non-freedom (No)

Fig. 2. Our overall framework for instruction-tuning of models to generate answer and predict response outcome for an input MRI scan,an ASM drug, and instruction prompt.

$(X_q^2, X_a^2), \ldots (X_q^N, X_a^N)\}$ serve as ground truth for fine-tuning, where N represents the number of pairs per sample. The instruction-answers are structured sequentially, ending with a "unified instruction-answer pair" $((X_q^U, X_a^U))$ (Fig 2) prompting the binary seizure-free response outcome (Yes/No) for the given ASM. The combined instructions are tokenized into embeddings H_q via a pretrained tokenizer [15] and the language model f_ϕ (Vicuna [7]) is fine-tuned on the generated tokens using its original autoregressive objective. Following [17], for an instruction-answer pair of length L, we compute the probability of the target answer X_a by:

$$p(\mathbf{X}_a|\mathbf{X}_I, \mathbf{X}_{ASM}, \mathbf{X}_q) = \prod_{i=1}^{L} p_\phi(x_i \mid \mathbf{X}_I, \mathbf{X}_{ASM}, \mathbf{X}_{q,<i}, \mathbf{X}_{a,<i}) \tag{1}$$

where ϕ are trainable parameters of the language model, $X_{q,<i} = (X^n_{q,<i} + X^U_{q,<i})$, and $X_{a,<i} = (X^n_{a,<i} + X^U_{a,<i})$ represent the combined n^{th} instruction-answer pair and the unified instruction-answer pair tokens before the current predicted token x_i.

In addition to the autoregressive training objective, we use binary cross entropy loss to optimize the prediction of response outcome for an ASM. As shown in Fig 2, we pass the generated answer tokens H_a to a trainable layer W_{OUT} followed by a $sigmoid()$ function to compute the probability of response outcome:

$$\hat{y}^i_{ASM} = \sigma(W_s H_a + b_s) = \frac{1}{1 + e^{-(W_s H_a + b_s)}} \tag{2}$$

where W_s and b_s are the trainable parameters of W_{OUT}, H_a are the generated answer tokens, and $\hat{y}^i_{ASM}$ is the predicted probability of seizure freedom/non-freedom for MRI scan X_I and the ASM X_{ASM}. We then apply the standard BCE loss between the predicted outcome probability and the ground truth [1,0] corresponding to seizure freedom / non-freedom.

Our framework uses direct fine-tuning, freezing the pretrained biomedical vision encoder g_θ and drug encoder d_ω, while updating only the projection matrices W_I (image), W_{ASM} (drug), W_{OUT} (outcome) and the language model f_ϕ. We optimize the network using both autoregressive and BCE loss with equal weighting. During inference, we only instruct the framework with the unified instruction X^U_q and obtain the generated answer and the outcome response in the manner described earlier.

2.3 Training & Implementation Details

To construct the instruction-tuning datasets, we used the GPT-4o API (Jan–Feb 2025). No data augmentation was applied to MRI scans. Fine-tuning started with a learning rate of 1×10^{-4}, using a cosine scheduler, 5-epoch warm-up, and a constant schedule type. Training ran for 100 epochs with a batch size of 16 per device using the AdamW optimizer. Following prior ASM recommendation studies [11], we evaluated performance using AUC, precision, and recall. Additionally, semantic similarity matching assessed the fine-tuned language model's responses. Experiments were conducted on NVIDIA A100 and A6000 GPUs.

3 Datasets and Splits

We retrospectively collected two private datasets: **ASM-ED1**, which is available upon request for research purposes only, and **ASM-ED2**, which could be curated from the Human Epilepsy Project [9]. Both datasets originate from distinct cohorts of newly diagnosed epilepsy patients. Seizure control was assessed one year after starting the first ASM, with success defined as seizure freedom while maintaining the same ASM.

ASM Distribution - Patients were administered with one of the seven ASMs: carbamazepine, lamotrigine, levetiracetam, oxcarbazepine, phenytoin, topiramate, or valproate. The most frequently used were levetiracetam, lamotrigine, oxcarbazepine, and carbamazepine, having 250, 79, 37, and 24 cases in ASM-ED1, and 116, 41, 39, and 25 cases in ASM-ED2. The remaining three ASMs had 54 cases only in ASM-ED1. ASM-ED1 included 444 patients (112 seizure-free, 332 not), while ASM-ED2 had 247 patients (62 seizure-free, 185 not).

Train and Test splits - We use 5-fold cross-validation with an 80/20 train-validation split for both datasets. ASM-ED1 is used to evaluate unseen ASM recommendation, as it includes all seven ASMs. Training is limited to the four most common ASMs (levetiracetam, lamotrigine, oxcarbazepine, and carbamazepine), while valproate, topiramate, and phenytoin are encountered only in testing.

4 Results

We analyze the lexical composition of our instruction-tuning dataset and also benchmark various biomedical foundation models (vision and vision-language) alongside different instruction-tuning approaches for ASM outcome prediction.

4.1 Lexical Analysis of Instruction-Tuning Approaches

We analyzed the lexical composition of instructions from different generation approaches (Table 1). TREE-TUNE shows the richest set of nouns and adjectives, significantly surpassing MRI report- or scan-based methods. Its significantly higher noun percentage captures more epilepsy-specific entities (e.g., lesion-encephalocele-gliosis, lateralization-right), while significant proportion of adjectives enhance descriptive detail, improving characterization of epilepsy-related MRI findings, as also shown in the verb-noun pair chart (Fig. 1).

Table 1. Comparison of lexical composition (in terms of %) across all instructions for the different instruction tuning approaches on both datasets combined.

Dataset / Lexical	noun	adjective	adverb	verb	numerical	preposition
MRI report [15]	34.1	16.3	1.5	12.9	0.7	11.9
MRI scan + report [2]	36.9	12.3	0.8	15.6	0.1	9.4
TREE-TUNE (Ours)	39.9	20.9	1.6	15.6	0.7	3.6

4.2 ASM Outcome Prediction Performance

We evaluate ASM outcome prediction in two settings: **seen ASMs** (trained) and **zero-shot unseen ASMs** (not seen during training). Table 2 shows seen ASM results, comparing vision-only, vision-language, and instruction-tuning methods, including the naïve prompt "Describe the findings in the MRI scan". Instruction-tuning consistently outperforms other methods, with our approach achieving the highest AUC (76.45 (p=0.012)) on ASM-ED1, 63.03 (p=0.041) on unseen ASMs,

Table 2. Seizure outcome prediction performance for popular seen ASMs.

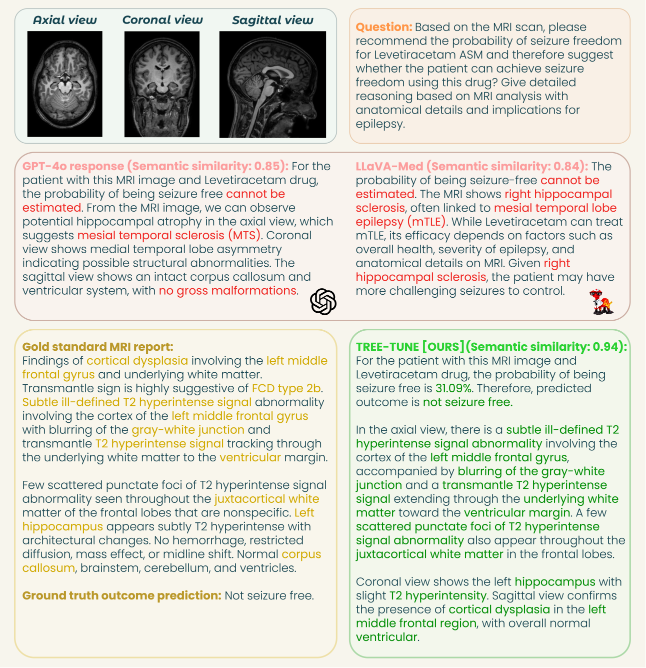
Method / Dataset / Metric	ASM-ED1			ASM-ED2		
	AUC (%)	Precision (%)	Recall (%)	AUC (%)	Precision (%)	Recall (%)
BiomedCLIP Vision [28]	$59.25_{1.15}$	$48.56_{1.12}$	$51.20_{1.14}$	$55.53_{1.30}$	$46.01_{1.20}$	$49.50_{1.30}$
LLaVA-Med Vision [15]	$60.37_{1.12}$	$49.58_{3.11}$	$52.43_{2.13}$	$58.91_{1.25}$	$48.21_{3.02}$	$51.74_{2.21}$
BiomedCLIP Vision-Language [28]	$65.00_{1.90}$	$54.02_{1.73}$	$53.01_{2.32}$	$51.86_{3.72}$	$52.00_{2.20}$	$51.01_{2.50}$
Instruction-Tuning Approaches						
Naive	$74.71_{2.13}$	$69.63_{5.15}$	$63.59_{1.24}$	$61.36_{3.65}$	$67.03_{5.53}$	$62.31_{1.65}$
MRI report [15]	$71.54_{4.35}$	$68.23_{4.76}$	$62.87_{2.83}$	$60.21_{5.12}$	$58.44_{5.97}$	$52.23_{3.44}$
MRI scan + report [2]	$75.24_{5.81}$	$69.57_{4.86}$	$65.76_{4.32}$	$61.75_{4.87}$	$61.86_{1.16}$	$60.24_{3.91}$
TREE-TUNE (Ours)	$76.45_{4.21}$	$70.52_{6.12}$	$67.85_{6.57}$	$66.33_{6.15}$	$61.94_{4.92}$	$55.42_{6.92}$
Average Increment	4.94	2.29	4.98	6.12	3.5	3.19

Table 3. Seizure outcome prediction performance for unseen ASMs.

Method / Metric	AUC (%)	Precision (%)	Recall (%)
BiomedCLIP Vision [28]	55.01	35.03	30.14
BiomedCLIP Vision-Language [28]	58.05	40.04	39.07
Instruction-tuning approaches			
Naive	60.12	45.06	48.68
MRI report [15]	59.52	51.03	59.05
MRI scan + report [2]	60.07	56.05	61.04
TREE-TUNE (Ours)	63.03	60.06	65.07
Average Increment	3.51	9.03	6.02

showing a **5.53**% average AUC gain over standard MRI report-based methods. ASM-ED2 has fewer samples and greater hospital variability, while ASM-ED1 (from only two hospitals) performs better. The challenging ASM-ED2 and unseen ASM settings highlight TREE-TUNE's strong performance.

Zero-shot prediction We evaluate TREE-TUNE's zero-shot capability on three unseen ASMs excluded from ASM-ED1 training. As shown in Table 3, our method improves AUC by **3.51**% over MRI report-based instruction tuning, outperforming others and maintaining robust generalization.



Question: Based on the MRI scan, please recommend the probability of seizure freedom for Levetiracetam ASM and therefore suggest whether the patient can achieve seizure freedom using this drug? Give detailed reasoning based on MRI analysis with anatomical details and implications for epilepsy.

GPT-4o response (Semantic similarity: 0.85): For the patient with this MRI image and Levetiracetam drug, the probability of being seizure free cannot be estimated. From the MRI image, we can observe potential hippocampal atrophy in the axial view, which suggests mesial temporal sclerosis (MTS). Coronal view shows medial temporal lobe asymmetry indicating possible structural abnormalities. The sagittal view shows an intact corpus callosum and ventricular system, with no gross malformations.

LLaVA-Med (Semantic similarity: 0.84): The probability of being seizure-free cannot be estimated. The MRI shows right hippocampal sclerosis, often linked to mesial temporal lobe epilepsy (mTLE). While Levetiracetam can treat mTLE, its efficacy depends on factors such as overall health, severity of epilepsy, and anatomical details on MRI. Given right hippocampal sclerosis, the patient may have more challenging seizures to control.

Gold standard MRI report:
Findings of cortical dysplasia involving the left middle frontal gyrus and underlying white matter. Transmantle sign is highly suggestive of FCD type 2b. Subtle ill-defined T2 hyperintense signal abnormality involving the cortex of the left middle frontal gyrus with blurring of the gray-white junction and transmantle T2 hyperintense signal tracking through the underlying white matter to the ventricular margin.

Few scattered punctate foci of T2 hyperintense signal abnormality seen throughout the juxtacortical white matter of the frontal lobes that are nonspecific. Left hippocampus appears subtly T2 hyperintense with architectural changes. No hemorrhage, restricted diffusion, mass effect, or midline shift. Normal corpus callosum, brainstem, cerebellum, and ventricles.

Ground truth outcome prediction: Not seizure free.

TREE-TUNE [OURS](Semantic similarity: 0.94):
For the patient with this MRI image and Levetiracetam drug, the probability of being seizure free is 31.09%. Therefore, predicted outcome is not seizure free.

In the axial view, there is a subtle ill-defined T2 hyperintense signal abnormality involving the cortex of the left middle frontal gyrus, accompanied by blurring of the gray-white junction and a transmantle T2 hyperintense signal extending through the underlying white matter toward the ventricular margin. A few scattered punctate foci of T2 hyperintense signal abnormality also appear throughout the juxtacortical white matter in the frontal lobes.

Coronal view shows the left hippocampus with slight T2 hyperintensity. Sagittal view confirms the presence of cortical dysplasia in the left middle frontal region, with overall normal ventricular.

Fig. 3. An exemplar case to demonstrate the generated response of our TREE-TUNE.

4.3 Qualitative Analysis of TREE-TUNED vs. Standard Models

Fig 3 compares TREE-TUNE, GPT-4o, and LLaVA-Med responses to a zero-shot prompt and MRI scan. GPT-4o and LLaVA-Med mislabel findings of cortical dysplasia, juxtacortical white matter abnormalities, and gray-white junction blurring as hippocampal sclerosis and mesial temporal lobe epilepsy. In contrast, TREE-TUNE accurately identifies these features, achieves the highest semantic similarity, and predicts ASM probability aligned with the ground truth. This demonstrates superior anatomical understanding and a step towards reasoning.

5 Conclusion

This work evaluates the effectiveness of biomedical foundation models in predicting ASM outcomes from MRI scans and reports. We introduce TREE-TUNE, a novel MRI knowledge tree driven contextualized instruction-tuning framework to enhance model performance beyond standard instruction-tuning strategies. Moreover, our approach not only achieves high accuracy in recommending commonly prescribed ASMs but also gives reasonable performance for unseen ASMs. Our work marks a foundational step toward developing a reasoning-based ASM recommendation system, paving the way for personalized epilepsy management.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bernasconi, A., Cendes, F., Theodore, W.H., Gill, R.S., Koepp, M.J., Hogan, R.E., Jackson, G.D., Federico, P., Labate, A., Vaudano, A.E., et al.: Recommendations for the use of structural magnetic resonance imaging in the care of patients with epilepsy: a consensus report from the international league against epilepsy neuroimaging task force. Epilepsia **60**(6), 1054–1068 (2019)
2. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. In: European Conference on Computer Vision. pp. 370–387. Springer (2024)
3. Chen, Z., Brodie, M.J., Kwan, P.: What has been the impact of new drug treatments on epilepsy? Current opinion in neurology **33**(2), 185–190 (2020)
4. Chen, Z., Brodie, M.J., Liew, D., Kwan, P.: Treatment outcomes in patients with newly diagnosed epilepsy treated with established and new antiepileptic drugs: a 30-year longitudinal cohort study. JAMA neurology **75**(3), 279–286 (2018)
5. Chen, Z., Rollo, B., Antonic-Baker, A., Anderson, A., Ma, Y., O'Brien, T.J., Ge, Z., Wang, X., Kwan, P.: Brain health: New era of personalised epilepsy management. The BMJ **371** (2020)
6. Cheval, M., Houot, M., Chastan, N., Szurhaj, W., Marchal, C., Catenoix, H., Valton, L., Gavaret, M., Herlin, B., Biraben, A., et al.: Early identification of seizure freedom with medical treatment in patients with mesial temporal lobe epilepsy and hippocampal sclerosis. Journal of Neurology **270**(5), 2715–2723 (2023)

7. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023) **2**(3), 6 (2023)
8. Croce, P., Ricci, L., Pulitano, P., Boscarino, M., Zappasodi, F., Lanzone, J., Narducci, F., Mecarelli, O., Di Lazzaro, V., Tombini, M., et al.: Machine learning for predicting levetiracetam treatment response in temporal lobe epilepsy. Clinical Neurophysiology **132**(12), 3035–3042 (2021)
9. Fox, J., Barnard, S., Agashe, S.H., Holmes, M.G., Gidal, B., Klein, P., Abou-Khalil, B.W., French, J., Investigators, H.E.P.: Patterns of antiseizure medication utilization in the human epilepsy project. Epilepsia **64**(12), 3196–3204 (2023)
10. Hakami, T., Mcintosh, A., Todaro, M., Lui, E., Yerra, R., Tan, K.M., French, C., Li, S., Desmond, P., Matkovic, Z., et al.: Mri-identified pathology in adults with new-onset seizures. Neurology **81**(10), 920–927 (2013)
11. Hakeem, H., Feng, W., Chen, Z., Choong, J., Brodie, M.J., Fong, S.L., Lim, K.S., Wu, J., Wang, X., Lawn, N., et al.: Development and validation of a deep learning model for predicting treatment response in patients with newly diagnosed epilepsy. JAMA neurology **79**(10), 986–996 (2022)
12. Hu, Z., Jiang, D., Zhao, X., Yang, J., Liang, D., Wang, H., Zhao, C., Liao, J.: Predicting drug treatment outcomes in children with tuberous sclerosis complex–related epilepsy: A clinical radiomics study. American Journal of Neuroradiology **44**(7), 853–860 (2023)
13. Labate, A., Aguglia, U., Tripepi, G., Mumoli, L., Ferlazzo, E., Baggetta, R., Quattrone, A., Gambardella, A.: Long-term outcome of mild mesial temporal lobe epilepsy: a prospective longitudinal cohort study. Neurology **86**(20), 1904–1910 (2016)
14. Lee, H.M., Fadaie, F., Gill, R., Caldairou, B., Sziklas, V., Crane, J., Hong, S.J., Bernhardt, B.C., Bernasconi, A., Bernasconi, N.: Decomposing mri phenotypic heterogeneity in epilepsy: a step towards personalized classification. Brain **145**(3), 897–908 (2022)
15. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36** (2024)
16. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024)
17. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
18. Maziarz, K., Jackson-Flux, H., Cameron, P., Sirockin, F., Schneider, N., Stiefl, N., Segler, M., Brockschmidt, M.: Learning to extend molecular scaffolds with structural motifs. arXiv preprint arXiv:2103.03864 (2021)
19. Perucca, E., Brodie, M.J., Kwan, P., Tomson, T.: 30 years of second-generation antiseizure medications: impact and future perspectives. The Lancet Neurology **19**(6), 544–556 (2020)
20. Sepehri, M.S., Fabian, Z., Soltanolkotabi, M., Soltanolkotabi, M.: Mediconfusion: Can you trust your ai radiologist? probing the reliability of multimodal medical foundation models. arXiv preprint arXiv:2409.15477 (2024)
21. Shazadi, K., Petrovski, S., Roten, A., Miller, H., Huggins, R.M., Brodie, M.J., Pirmohamed, M., Johnson, M.R., Marson, A.G., O'Brien, T.J., et al.: Validation

of a multigenic model to predict seizure control in newly treated epilepsy. Epilepsy research **108**(10), 1797–1805 (2014)

22. Shin, Y., Hwang, S., Lee, S.B., Son, H., Chu, K., Jung, K.Y., Lee, S.K., Park, K.I., Kim, Y.G.: Using spectral and temporal filters with eeg signal to predict the temporal lobe epilepsy outcome after antiseizure medication via machine learning. Scientific Reports **13**(1), 22532 (2023)
23. Silva-Alves, M.S., Secolin, R., Carvalho, B.S., Yasuda, C.L., Bilevicius, E., Alvim, M.K., Santos, R.O., Maurer-Morelli, C.V., Cendes, F., Lopes-Cendes, I.: A prediction algorithm for drug response in patients with mesial temporal lobe epilepsy based on clinical and genetic information. PLoS One **12**(1), e0169214 (2017)
24. Wang, X., Hu, T., Yang, Q., Jiao, D., Yan, Y., Liu, L.: Graph-theory based degree centrality combined with machine learning algorithms can predict response to treatment with antiepileptic medications in children with epilepsy. Journal of Clinical Neuroscience **91**, 276–282 (2021)
25. Weininger, D.: Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. Journal of chemical information and computer sciences **28**(1), 31–36 (1988)
26. World Health Organization: Epilepsy (2024), https://www.who.int/news-room/fact-sheets/detail/epilepsy, accessed: 13 Feb. 2025
27. Xiong, Z., Wang, X., Zhou, Y., Keane, P.A., Tham, Y.C., Wang, Y.X., Wong, T.Y.: How generalizable are foundation models when applied to different demographic groups and settings? NEJM AI **2**(1), AIcs2400497 (2025)
28. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)