A Clinical Guideline-Grounded Hybrid Agentic Framework for Holistic Epilepsy Management

Duy Khoa Pham[1,5], Dinesh Giritharan[3], Guilherme Camargo de Oliveira[1], Bao Quoc Vo[5], Karin Verspoor[4], Meng Law[3], Patrick Kwan[3], Zongyuan Ge[1,2], and Deval Mehta[2,4(✉)]

[1] AIM for Health Lab, Faculty of IT, Monash University, Melbourne, Australia
[2] Department of Data Science & AI, Faculty of IT, Monash University, Melbourne, Australia
[3] School of Translational Medicine, Monash University, Melbourne, Australia
[4] School of Computing Technologies, RMIT University, Melbourne, Australia
[5] Department of Computing Technologies, Swinburne University of Technology, Melbourne, Australia
deval.mehta1092@gmail.com, deval.mehta@monash.edu

Abstract. Epilepsy is a chronic neurological disorder requiring multi-faceted management, including seizure detection, syndrome diagnosis, prognostication, antiseizure medication recommendation, epileptogenic zone localization, and surgical outcome prediction. Although numerous deep learning approaches have been developed for individual tasks, these models are typically siloed and modality-specific (e.g., EEG for seizure detection, MRI for localization), failing to reflect the multidisciplinary nature of real-world epilepsy care, where epileptologists, neuroradiologists, neurosurgeons, neuropsychologists and neuropsychiatrists jointly interpret heterogeneous evidence to guide decisions. In this work, we propose a clinical guideline-grounded hybrid multi-agent framework for holistic epilepsy management. Heterogeneous patient data is processed through modality-specific discriminative and generative models, where textual interpretations from generative agents are combined with structured predictions from discriminative models as auxiliary guidance. This aggregated evidence is passed to a central orchestrating agent grounded in international epilepsy guidelines, which evaluates multi-modal findings within structured clinical pathways and performs iterative cross-agent coordination for evidence-informed decision-making. We evaluate our framework across two datasets spanning six epilepsy management tasks and also introduce a publicly available multi-modal, multi-task epilepsy benchmark. Results demonstrate that integrating discriminative evidence with guideline-grounded generative coordination yields more reliable and comprehensive decisions compared to conventional LLM-based and task-specific baselines. Our dataset and code is available at URL.

Keywords: epilepsy · AI agents · LLMs

1 Introduction

Epilepsy is a chronic neurological disorder characterized by recurrent seizures due to abnormal brain activity, affecting over 60 million people worldwide [18]. Notably, more than one-third of patients develop drug-resistant epilepsy, failing to achieve seizure control despite trials of multiple appropriate medications [13]. Given its lifelong and heterogeneous nature, epilepsy requires continuous, multi-stage management – from seizure and syndrome diagnosis to epileptogenic zone (EZ) localization and treatment decisions such as anti-seizure medication (ASM) selection or surgical intervention. These decisions depend on integrating heterogeneous data, including electroencephalography (EEG), magnetic resonance imaging (MRI), genomics, and diverse clinical factors, necessitating coordinated multi-modal analysis.

Numerous automated methods have been developed for individual epilepsy tasks (Fig 1-a), including seizure type classification from EEG [3] and video [16,10], lesion localization from MRI [2,23], EZ prediction from semiology reports [28], and ASM response prediction from clinical and imaging data [20]. More recently, foundation models (FMs) and large language models (LLMs) have been explored for modality-specific analysis and reasoning (Fig 1-b), such as EEG-based report generation [11], automated event detection from health records [9], and explanation-driven ASM recommendation [22]. However, despite these advances, existing methods remain siloed, limited to single tasks and specific modalities, failing to reflect the coordinated, multi-task nature of real-world epilepsy care.

More broadly, the medical AI community has seen rapid growth in generative multi-agent systems that coordinate role-specific LLMs for complex clinical tasks [6,12,25,14,4]. These frameworks decompose problems into specialized agents that interact through iterative message passing to produce unified decisions (MDAgents [12]), demonstrating promise in multi-step diagnostic reasoning [4], collaborative image interpretation [6], and even simulating hospital workflows [14]. While these systems enable complex coordination, purely generative LLMs may suffer from instability, hallucinations, and weak grounding in quantitative modality-specific signals, particularly in numerical reasoning [15]. In contrast, discriminative models provide robust, task-optimized representations from structured data. This motivates us to develop a **hybrid paradigm** (Fig 1-c) that combines the reliability of discriminative modeling with the coordination and reasoning capabilities of generative agentic systems.

Our contribution: We propose **EPI-GUIDE**, a clinically Guideline-grounded hybrid multi-agent framework for holistic Epilepsy management. To the best of our knowledge, it is among the first to apply hybrid discriminative-generative model coupling along with guideline-grounded agentic AI for multi-task epilepsy management. Our framework incorporates the below innovations:

- A hybrid system coupling modality-specific discriminative models with generative LLM agents, where structured discriminative predictions provide auxiliary guidance to the agentic system for final decision-making.
- A central orchestrating agent enabling structured multi-turn interactions to resolve inter-modality inconsistencies.

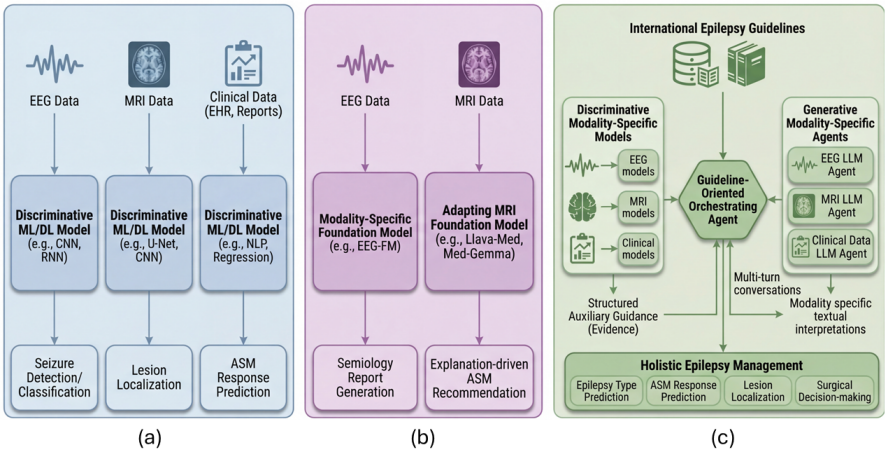

Fig. 1. Comparative overview. (a) *Siloed discriminative models* trained on modality-specific data for individual tasks; (b) *Siloed foundation models* applied to single-task, single-modality settings; (c) *Our proposed hybrid framework* coupling discriminative models and generative agents for multi-modal, multi-task holistic epilepsy management.

– Guideline grounding of the orchestrator using established international epilepsy guidelines to evaluate multimodal outputs within structured clinical pathways, enabling evidence-informed decision-making.

We evaluate our framework across two datasets spanning six epilepsy management tasks and will **publicly release a curated multi-modal, multi-task epilepsy dataset** to support future research in this direction. Our results demonstrate that combining discriminative evidence with generative coordination, particularly when augmented with clinical guidelines, leads to more reliable and comprehensive epilepsy decision-making.

2 Methodology

Our overall framework (Fig. 2), comprises three main components: (1) modality-specific discriminative and generative models; (2) multimodal evidence aggregation and task-driven guideline retrieval; and (3) a guideline-grounded orchestrating agent that evaluates integrated evidence for final decision-making.

2.1 Modality-specific Models and Evidence Aggregation

Clinical Practice & Multi-Modality Multi-task Problem Formulation: Epilepsy care is inherently multidisciplinary, with epileptologists, neuroradiologists, neurosurgeons, neuropsychiatrists and neuropsychologists integrating clinical history, EEG interpretation, neuroimaging findings and cognitive assessments to formulate diagnostic and treatment decisions. To emulate this workflow, let

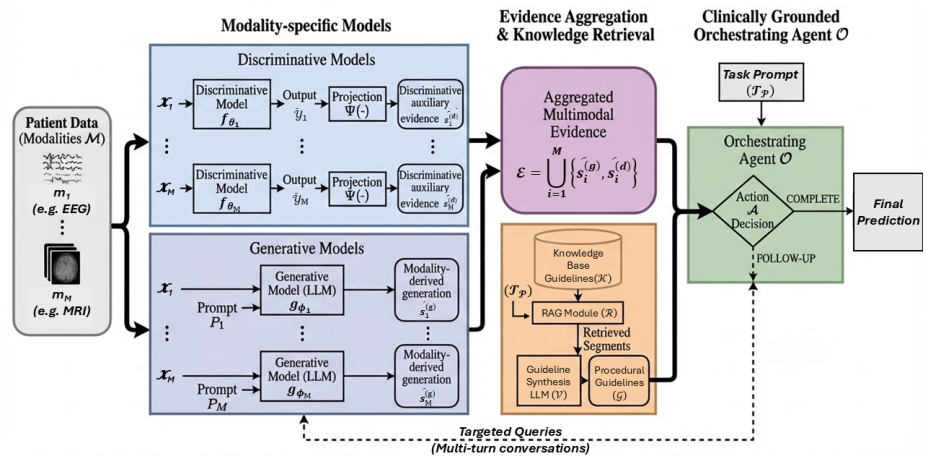

Fig. 2. Methodology overview: EPI-GUIDE is a hybrid framework that integrates evidence from discriminative and generative models, coordinated by an epilepsy guideline–grounded orchestrating agent for multi-modal, multi-task decision-making.

$\mathcal{M} = \{m_1, \ldots, m_M\}$ denote the set of available patient modalities (e.g., EEG, MRI, ...), and $\mathcal{T} = \{t_1, \ldots, t_T\}$ the set of epilepsy management tasks (e.g., epilepsy type classification, EZ localization, ...). Given inputs from $\mathcal{M}$, the objective is to generate predictions across $\mathcal{T}$. We therefore define two complementary model families: modality-specific discriminative models and modality-specific generative models.

Discriminative Models: For each modality $m_i \in \mathcal{M}$, we define a discriminative model $f_{\theta_i} : \mathcal{X}_i \to \hat{\mathcal{Y}}_i$, parameterized by θ_i, where $\mathcal{X}_i$ denotes the modality-specific input space and $\hat{\mathcal{Y}}_i$ the task-specific prediction space. Each model is trained using supervised objectives appropriate to the target task $t \in \mathcal{T}$ (for e.g., cross-entropy for epilepsy type classification). The resulting predictions $\hat{\mathcal{Y}}_i$ may include class probabilities, regression scores, or confidence estimates depending upon the task $t_i \in \mathcal{T}$ (for e.g., classification, risk estimation, ...).

To integrate with the language-based agentic framework, these structured outputs are converted into textual evidence via

$$\tilde{s}_i^{(d)} = \Psi(\hat{\mathcal{Y}}_i), \tag{1}$$

where $\Psi(\cdot)$ denotes an LLM-based transformation and $\tilde{s}_i^{(d)} \in \mathcal{S}^{(d)}$ denotes modality-specific discriminative evidence used as auxiliary guidance in our framework.

Generative Models: In parallel, we define a generative model route through:

$$\tilde{s}_i^{(g)} = g_{\phi_i}(\mathcal{X}_i, P_i), \quad \tilde{s}_i^{(g)} \in \mathcal{S}^{(g)}, \tag{2}$$

where g_{ϕ_i} is an LLM initialized and prompted with modality-specific instructions P_i to produce textual clinical interpretations. Finally, the discriminative $\tilde{s}_i^{(d)}$ and generative $\tilde{s}_i^{(g)}$ outputs are then aggregated as:

$$\mathcal{E} = \bigcup_{i=1}^{M} \left\{ \tilde{s}_i^{(g)}, \tilde{s}_i^{(d)} \right\}, \tag{3}$$

forming the unified evidence passed to the orchestrating agent. This mirrors clinical multidisciplinary team (MDT) practice, isolating our contributions of hybrid modeling and guideline grounding from fusion depth (Sec. 4). Missing modalities in MME (MRI: 94/306, EEG: 71/306) complicate latent fusion, which we leave to future work.

2.2 Clinically Grounded Orchestrating Agent

Epilepsy management follows structured diagnostic and treatment pathways defined by international guidelines and multidisciplinary consensus. Motivated by this workflow, we introduce a guideline-grounded orchestrating agent built on Retrieval-Augmented Generation (RAG) to analyze the aggregated evidence.

Our knowledge base $\mathcal{K}$ is **curated by expert neurologists** and comprises **66** documents on epilepsy guidelines from the International League Against Epilepsy (ILAE) [26,1], National Institute for Health and Care Excellence (NICE) [19], and epilepsy textbooks, covering epilepsy type classification, drug-resistant epilepsy, ASM selection, surgical candidacy, and prognosis. The corpus is indexed into a vector database and segmented into semantic chunks. Given a task query $\mathcal{T}_\mathcal{P}$, the RAG module $\mathcal{R}$ retrieves relevant guideline segments, which are synthesized into a procedural guideline $\mathcal{G} = \mathcal{V}(\mathcal{R}(\mathcal{K}))$, where $\mathcal{V}(\cdot)$ denotes LLM-based guideline synthesis.

The orchestrator operates on aggregated multimodal evidence $\mathcal{E}$ as:

$$\mathcal{O} : (\mathcal{E}, \mathcal{G}, \mathcal{T}_P) \to \mathcal{A}, \quad \mathcal{A} \in \{\text{FOLLOW-UP}, \text{COMPLETE}\}. \tag{4}$$

where $\mathcal{A}$ is the action and $\mathcal{T}_P$ a task prompt. The orchestrator computes a confidence-weighted concordance score between classifier and agent predictions: if ≥ 0.7 and consistent with $\mathcal{G}$, $\mathcal{A}$ is COMPLETE. Otherwise, a targeted FOLLOW-UP addresses the most uncertain disagreement, for up to three rounds. Final predictions use the classifier ($\geq 85\%$), the generative interpretation ($< 50\%$), or a confidence-weighted vote. Missing modalities trigger a no-data branch with weight renormalization over available evidence.

2.3 Implementation Details

Discriminative models: We train modality-specific discriminative predictors with weighted cross-entropy (inverse class frequency) for all classification tasks. *Clinical Text/Records:* TF–IDF (max_features = 5000, bigrams) with XGBoost, and PubMedBERT [7] fine-tuned with a lightweight classification head. *MRI/EEG:* ImageNet-pretrained ResNet-50 [8] and MedSigLIP-448 [24] (frozen vision encoder; trained head). Input images are resized to 224×224 (ResNet) or 448×448 (MedSigLIP). For EEG, we additionally fine-tune an EEG foundation model (REVE-base [5]) with task-specific heads.

Generative models: We use MedGemma-27B-text-it [24] for clinical text and MedGemma-1.5-4B-it [24] for MRI/EEG inputs (temperature = 0.1, max tokens = 2048), each initialized with modality-specific prompts.
Guideline-grounded orchestrator and RAG: We use GPT-OSS-120b [21]; temperature = 0.1, tokens=2048) as the orchestrating agent. The guideline corpus is indexed with bge-base-en-v1.5 embeddings [27] and retrieved by cosine similarity w.r.t the task prompt. The orchestrator may perform up to three multi-turn rounds, marking a case COMPLETE once its patient-level concordance score reaches 0.7. The 8×A100-80GB configuration is used to host GPT-OSS-120B locally; the discriminative models are lightweight, and the modular design allows swapping in more efficient or hosted models without changing the framework. Details about learning rate, optimizer, number of epochs and other hyperparameters for each individual model is available in our code repository.

3 Datasets

We benchmark with a standard 5-fold cross validation for two datasets. On the MME cohort (306 patients, five tasks, missing modalities), folds are stratified by epilepsy type and modality availability (has_mri × has_eeg), ensuring balanced folds with no patient overlap.
1. Curated Multi-modal Multi-task Epilepsy (MME) Dataset: Our curated cohort (to be released) includes 306 epilepsy patients, with MRI available for 94 and EEG for 71 cases. It contains five epilepsy management tasks, each as a three-class problem: (1) **Epilepsy type** (Focal: 216, Generalized: 30, Other: 58; n=304), (2) **Seizure type** (Focal Onset: 150, Generalized Onset: 57, Unknown/Other: 94; n=301), (3) **EZ localization** (Temporal: 151, Extratemporal: 64, Multifocal: 27; n=242), (4) **ASM response** (Drug-Resistant: 151, Responsive: 34, On Treatment: 57; n=242), and (5) **Surgical outcome** (Seizure-Free: 55, Improved: 19, No Improvement: 59; n=133). Due to real-world clinical variability, not all patients contain complete annotations for every task, reflecting the heterogeneous and longitudinal nature of epilepsy management.
2. Human Intracerebral Stimulation HD-EEG Dataset [17]: We additionally benchmark on the public simultaneous intracerebral stimulation and HD-EEG dataset [17], comprising 7 subjects and 61 stimulation sessions with paired 256-channel HD-EEG, sEEG recordings, and structural MRI. We define two tasks: (1) **EZ localization** (Temporal/Frontal/Parieto-Occipital), derived from stimulation-site coordinates; and (2) **Stimulation intensity classification** (Low $\leq$0.3 mA / High $\geq$0.5 mA).

4 Results

Benchmark on MME and HD-EEG Datasets Tables 1 and 2 compare EPIGUIDE against modality-specific discriminative and ensemble models as well as generative baselines on both our MME and the public HD-EEG dataset [17]. On the MME cohort, generative models underperform (MedGemma-27B: 61.8%

Table 1. Accuracy (%) on five epilepsy management tasks on our MME dataset. Results are reported as mean$_{\pm \text{std}}$ across 5 folds or multiple evaluation rounds. T: clinical text; M: MRI; E: EEG. **Bold** - best performance, underline - second best.

Method	Mod.	Epil. Type	Sz. Type	EZ Loc.	ASM Resp.	Surg. Out.	Overall
Discriminative Models							
TF-IDF + XGBoost	T	$78.3_{\pm 4.7}$	$70.4_{\pm 2.4}$	$81.8_{\pm 1.5}$	$76.4_{\pm 5.4}$	$68.4_{\pm 3.7}$	$75.1_{\pm 3.5}$
PubMedBERT [7]	T	$83.9_{\pm 1.7}$	$65.1_{\pm 4.2}$	$83.5_{\pm 3.6}$	$75.2_{\pm 2.9}$	$80.5_{\pm 4.8}$	$77.6_{\pm 3.4}$
ResNet-50 [8]	M	$71.8_{\pm 7.9}$	$57.8_{\pm 5.7}$	$79.7_{\pm 8.7}$	$78.4_{\pm 5.0}$	$56.1_{\pm 5.7}$	$68.8_{\pm 6.6}$
MedSigLIP-448 [24]	M	$82.6_{\pm 10.0}$	$61.1_{\pm 6.1}$	$74.0_{\pm 13.2}$	$79.9_{\pm 11.8}$	$79.3_{\pm 4.7}$	$75.4_{\pm 9.2}$
ResNet-50 [8]	E	$69.1_{\pm 8.4}$	$56.5_{\pm 11.6}$	$65.3_{\pm 10.4}$	$\mathbf{87.1}_{\pm 6.5}$	$40.7_{\pm 16.9}$	$63.7_{\pm 10.8}$
MedSigLIP-448 [24]	E	$74.7_{\pm 13.2}$	$52.2_{\pm 11.8}$	$73.1_{\pm 13.2}$	$\mathbf{87.1}_{\pm 3.2}$	$75.3_{\pm 12.8}$	$72.5_{\pm 10.8}$
Weighted Ensemble	T+M+E	$81.9_{\pm 1.2}$	$69.4_{\pm 1.8}$	$84.3_{\pm 1.5}$	$78.9_{\pm 1.4}$	$79.0_{\pm 1.1}$	$78.7_{\pm 1.3}$
Stacking Ensemble	T+M+E	$82.2_{\pm 1.1}$	$70.1_{\pm 1.7}$	$83.1_{\pm 1.4}$	$78.5_{\pm 1.3}$	$75.9_{\pm 1.6}$	$77.9_{\pm 1.4}$
Generative Models							
MedGemma-27B [24]	T	$72.4_{\pm 4.5}$	$53.2_{\pm 5.1}$	$67.8_{\pm 4.8}$	$63.2_{\pm 5.3}$	$52.6_{\pm 6.2}$	$61.8_{\pm 5.2}$
MedGemma-4B [24]	M	$52.2_{\pm 5.8}$	$30.0_{\pm 7.2}$	$56.5_{\pm 6.1}$	$8.1_{\pm 3.4}$	$0.0_{\pm 0.0}$	$29.4_{\pm 4.5}$
Hybrid (Ours)							
EPI-GUIDE	**T+M+E**	$\mathbf{89.5}_{\pm 1.7}$	$\mathbf{77.5}_{\pm 2.1}$	$\mathbf{91.0}_{\pm 1.5}$	$84.8_{\pm 1.9}$	$\mathbf{86.2}_{\pm 2.0}$	$\mathbf{85.8}_{\pm 1.8}$

using clinical text; VLM: 29.4% using MRI), likely due to zero-shot setting; while fine-tuned discriminative models perform better (PubMedBERT: 77.6%; MedSigLIP-448: 75.4%), with the best ensemble at 78.7%, however they aren't explainable. EPI-GUIDE achieves **85.8%** overall, outperforming PubMedBERT (+8.2), the ensemble (+7.1), and the best generative model (+24.0), with the largest gains on the data-scarce surgical outcome task (+5.7 over ensemble). EPI-GUIDE attains **84–86 wF1** and **78–80 mF1**, versus 78.0/71.3 (best ensemble) and 56.8/43.1 (MedGemma-27B zero-shot); the 6.6-point gap reflects low recall (<45%) on minority classes (Multifocal EZ, On-treatment ASM).

On the public HD-EEG dataset, classical EEG features perform poorly across both tasks (36.1–39.7%). The supervised EEG foundation model REVE-base improves to 74.5% overall, with strong stimulation intensity classification (88.1%) but limited EZ region accuracy (60.8%). EPI-GUIDE achieves **81.3%** overall, surpassing REVE (+6.8) and the Multi-Agent LLM baseline (+8.3), with gains most pronounced on EZ region classification (+14.9 over Multi-Agent LLM).

Ablation Studies for EPI-GUIDE. Table 3 evaluates each component. Removing discriminative guidance drops accuracy to 62.9%, and removing both discriminative and guideline components to 60.8% (generative baseline). Disabling guideline retrieval gives 78.8%, the 7.0-point gain splitting into +3.0 (PubMed) and +4.0 (ILAE/NICE), concentrated on EZ localization and surgical outcome. On disagreement cases (22–45%), the three-tier rule lifts macro-F1 from 27% to 65%. FOLLOW-UP adds +0.1% overall but +6.3% on cases needing ≥2 rounds. Under missing modalities, EPI-GUIDE reaches 80.1% (T), 83.4% (T+M), 82.6% (T+E), confirming text as the anchor modality.

Qualitative Evaluation by Expert Neurologists Fig. 3 presents a representative MME case and corresponding EPI-GUIDE outputs. We further conducted expert evaluation with neurologists on 15 cases spanning five clinical scenarios (classifier correction, multimodal fusion, surgical prediction, rare syn-

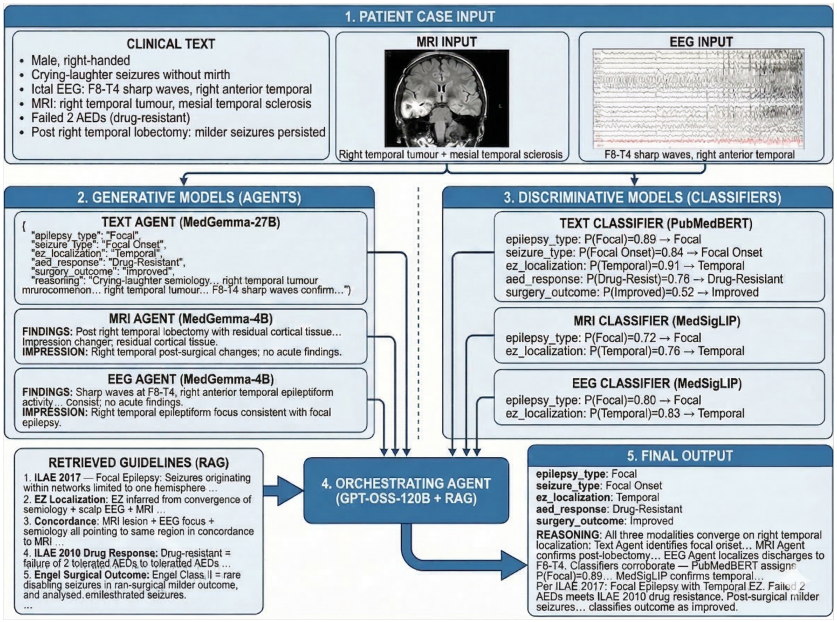

Fig. 3. A case study demonstrating the overall flow of EPI-GUIDE.

dromes, and comprehensive classification), rated on a 1–5 Likert scale for assessing prediction reliability. EPI-GUIDE achieved 93.4% task-level accuracy (57/61) compared to 54.1% for LLM-only (33/61), and received an average rating of **3.64/5** compared to LLM-only (**2.64/5**) on reliability of predictions.

Table 2. Accuracy(%) on two epilepsy management tasks on the HD-EEG dataset [17]. T: clinical text; M: MRI; E: EEG. **Bold** - best performance, underline - second best.

Method	Mod.	EZ Loc.	Stim. Int.	Overall
Discriminative Models				
GFP + XGBoost	E	$23.8_{\pm 19.0}$	$48.4_{\pm 37.2}$	$36.1_{\pm 28.1}$
BandPower + XGBoost	E	$36.7_{\pm 9.7}$	$42.7_{\pm 28.0}$	$39.7_{\pm 18.9}$
REVE (fine-tuned) [5]	E	$\underline{60.8_{\pm 14.1}}$	$88.1_{\pm 28.6}$	$\underline{74.5_{\pm 21.4}}$
Generative Models				
MedGemma Text-Only [24]	T	$44.3_{\pm 9.8}$	$61.8_{\pm 11.2}$	$53.1_{\pm 10.5}$
Multi-Agent LLM	T+M+E	$49.2_{\pm 8.5}$	$\underline{96.7_{\pm 4.2}}$	$73.0_{\pm 6.3}$
Hybrid (Ours)				
EPI-GUIDE	**T+M+E**	$\mathbf{64.1}_{\pm 5.2}$	$\mathbf{98.4}_{\pm 0.8}$	$\mathbf{81.3}_{\pm 3.5}$

Table 3. Ablation accuracy (%) of EPI-GUIDE components on MME. Disc.: discriminative auxiliary evidence; RAG: guideline-grounded retrieval. **Bold** indicates best.

Configuration	Disc.	RAG	Epil. Type	Sz. Type	EZ Loc.	AED	Resp.	Surg. Out.	Mean
EPI-GUIDE (Ours)	✓	✓	**89.5**	**77.5**	**91.0**	**84.8**	**86.2**	**85.8**	
w/o RAG	✓	–	83.2	70.4	84.1	77.5	78.9	78.8	
w/o Disc.	–	✓	76.1	57.8	62.4	64.8	53.6	62.9	
w/o Both	–	–	74.3	55.2	60.3	63.2	51.1	60.8	

5 Conclusion

In this work, we present EPI-GUIDE, to our knowledge among the first hybrid, guideline-grounded agentic framework for holistic epilepsy management. Addressing the limitations of siloed, single-task models and the instability of purely generative systems, our hybrid approach integrates discriminative outputs as auxiliary evidence with guideline-grounded multi-agent coordination. EPI-GUIDE demonstrates consistent improvements across two datasets, demonstrating the promise of hybrid discriminative-generative agentic systems for robust decision-making.

Acknowledgment and Disclosure of Interests. This work was supported in part by an NVIDIA Academic Grant which provided cloud access to 8xA100 GPUs. The authors declare no competing interests.

References

1. Beniczky, S., Trinka, E., Wirrell, E., Abdulla, F., Al Baradie, R., Alonso Vanegas, M., Auvin, S., Singh, M.B., Blumenfeld, H., Bogacz Fressola, A., et al.: Updated classification of epileptic seizures: Position paper of the international league against epilepsy. Epilepsia **66**(6), 1804–1823 (2025)
2. Berger, M., Licandro, R., Nenning, K.H., Langs, G., Bonelli, S.: Artificial intelligence applied to epilepsy imaging: Current status and future perspectives. Revue Neurologique **181**(5), 420–424 (2025)
3. Boonyakitanont, P., Lek-Uthai, A., Chomtho, K., Songsiri, J.: A review of feature extraction and performance evaluation in epileptic seizure detection using eeg. Biomedical Signal Processing and Control **57**, 101702 (2020)
4. Deng, R., Martin, G., Wang, T., Zhang, G., Liu, Y., Weng, C., Wang, Y., Rousseau, J.F., Peng, Y.: Cpgprompt: Translating clinical guidelines into llm-executable decision support. arXiv preprint arXiv:2601.03475 (2026)
5. El Ouahidi, Y., Lys, J., Thölke, P., Farrugia, N., Pasdeloup, B., Gripon, V., Jerbi, K., Lioi, G.: REVE: A foundation model for EEG: Adapting to any setup with large-scale pretraining on 25,000 subjects. Advances in Neural Information Processing Systems (2025), https://brain-bzh.github.io/reve/
6. Ghezloo, F., Seyfioglu, M.S., Soraki, R., Ikezogwo, W.O., Li, B., Vivekanandan, T., Elmore, J.G., Krishna, R., Shapiro, L.: Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23431–23441 (2025)

7. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**(1), 1–23 (2021)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Holgate, B., Davies, J., Fang, S., Winston, J., Teo, J., Richardson, M.: Fine-tuning llms to extract epilepsy seizure frequency data from health records. In: Proceedings of the 24th Workshop on Biomedical Language Processing. pp. 44–55 (2025)
10. Ilyas, T., Mehta, D., Sivathamboo, S., Wijaya, I., Steele, R., Simpson, H., Millist, L., O'Brien, T., Kwan, P., Ge, Z.: Privacy-centric seizure diagnosis via relation-aware fusion of minimally-invasive modalities. In: International Workshop on Applications of Medical AI. pp. 173–183. Springer (2025)
11. Jiang, W.B., Wang, Y., Lu, B.L., Li, D.: Neurolm: A universal multi-task foundation model for bridging the gap between language and eeg signals. arXiv preprint arXiv:2409.00101 (2024)
12. Kim, Y., Park, C., Jeong, H., Chan, Y.S., Xu, X., McDuff, D., Lee, H., Ghassemi, M., Breazeal, C., Park, H.W.: Mdagents: An adaptive collaboration of llms for medical decision-making. Advances in Neural Information Processing Systems **37**, 79410–79452 (2024)
13. Kwan, P., Schachter, S.C., Brodie, M.J.: Drug-resistant epilepsy. New England Journal of Medicine **365**(10), 919–926 (2011)
14. Li, J., Lai, Y., Li, W., Ren, J., Zhang, M., Kang, X., Wang, S., Li, P., Zhang, Y.Q., Ma, W., et al.: Agent hospital: A simulacrum of hospital with evolvable medical agents. arXiv preprint arXiv:2405.02957 (2024)
15. Mahendra, R., Spina, D., Cavedon, L., Verspoor, K.: Evaluating numeracy of language models as a natural language inference task. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Findings of the Association for Computational Linguistics: NAACL 2025. pp. 8351–8376. Association for Computational Linguistics, Albuquerque, New Mexico (Apr 2025). https://doi.org/10.18653/v1/2025.findings-naacl.467, https://aclanthology.org/2025.findings-naacl.467/
16. Mehta, D., Sivathamboo, S., Simpson, H., Kwan, P., O'Brien, T., Ge, Z.: Privacy-preserving early detection of epileptic seizures in videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 210–219. Springer (2023)
17. Mikulan, E., Russo, S., Parmigiani, S., Sarasso, S., Zauli, F.M., Rubino, A., Avanzini, P., Cattani, A., Sorrentino, A., Gibbs, S., et al.: Simultaneous human intracerebral stimulation and hd-eeg, ground-truth for source localization methods. Scientific data **7**(1), 127 (2020)
18. Moshé, S.L., Perucca, E., Ryvlin, P., Tomson, T.: Epilepsy: new advances. The Lancet **385**(9971), 884–898 (2015)
19. National Institute for Health and Care Excellence: Cenobamate for treating focal onset seizures in epilepsy. Technology appraisal guidance TA753, National Institute for Health and Care Excellence (2025), https://www.nice.org.uk/guidance/ta753
20. Nazem-Zadeh, M.R., Chang, R.S.k., Barnard, S., Pardoe, H.R., Kuzniecky, R., Mishra, D., Kamkar, H., Nhu, D., Metha, D., Thom, D., et al.: Development and validation of clinico-imaging machine learning and deep learning models to predict responses to initial antiseizure medications in epilepsy. Epilepsia (2025)
21. OpenAI: gpt-oss-120b & gpt-oss-20b model card (2025), https://arxiv.org/abs/2508.10925

22. Pham, D.K., Mehta, D., Jiang, Y., Thom, D., Chang, R.S.k., Nazem-Zadeh, M., Foster, E., Fazio, T., Holper, S., Verspoor, K., et al.: Knowledge tree driven contextualized instruction tuning of foundation models for epilepsy drug recommendation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 373–383. Springer (2025)
23. Ripart, M., Spitzer, H., Williams, L.Z., Walger, L., Chen, A., Napolitano, A., Rossi-Espagnet, C., Foldes, S.T., Hu, W., Mo, J., et al.: Detection of epileptogenic focal cortical dysplasia using graph neural networks: a meld study. JAMA neurology **82**(4), 397–406 (2025)
24. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
25. Wang, Z., Wu, J., Cai, L., Low, C.H., Yang, X., Li, Q., Jin, Y.: Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. arXiv preprint arXiv:2503.18968 (2025)
26. Wirrell, E.C., Nabbout, R., Scheffer, I.E., Alsaadi, T., Bogacz, A., French, J.A., Hirsch, E., Jain, S., Kaneko, S., Riney, K., et al.: Methodology for classification and definition of epilepsy syndromes with list of syndromes: report of the ilae task force on nosology and definitions. Epilepsia **63**(6), 1333–1348 (2022)
27. Xiao, S., Liu, Z., Zhang, P., Muennighoff, N.: C-pack: Packaged resources to advance general chinese embedding (2023)
28. Yang, S., Luo, Y., Fotedar, N., Jiao, M., Rao, V.R., Ju, X., Wu, S., Xian, X., Sun, H., Karakis, I., et al.: Episemollm: A fine-tuned large language model for epileptogenic zone localization based on seizure semiology with a performance comparable to epileptologists. MedRxiv pp. 2024–05 (2024)