Neural Signals Generate Clinical Notes in the Wild

Jathurshan Pradeepkumar[1] **Zheng Chen**[2*] **Jimeng Sun**[1*]
[1]University of Illinois Urbana-Champaign [2]SANKEN, Osaka University
{jp65,jimeng}@illinois.edu {chenz}@sanken.osaka-u.ac.jp

Abstract

Generating clinical reports that summarize abnormal patterns, diagnostic findings, and clinical interpretations from long-term EEG recordings remains labor-intensive. We present CELM, the first clinical EEG-to-Language foundation model capable of summarizing long-duration, variable-length EEG recordings and performing end-to-end clinical report generation at multiple scales. CELM integrates pretrained EEG foundation models with language models to enable scalable multimodal learning. We curate a large-scale clinical EEG dataset containing 9,922 reports paired with approximately 11,000 hours of EEG recordings from 9,048 patients to train CELM, and release the benchmark pipeline, with automated report-structuring, to facilitate future research. Experimental results show that CELM consistently outperforms existing methods across all evaluation settings. Importantly, we further conduct human evaluation with clinical experts, demonstrating that CELM generates reports that are more clinically coherent, diagnostically reliable, and better aligned with expert interpretation. We release our model code and benchmark construction pipeline at https://github.com/Jathurshan0330/CELM.

1 Introduction

Electroencephalography (EEG) records long-term, real-time neuronal activity with millisecond-level temporal resolution and is widely used for neurological diagnosis [1, 2, 3]. However, writing clinical reports to summarize diagnostic findings over EEG recordings remains a labor-intensive process. Typically, neurologists visually inspect EEG signals to assess brain activity and then compose clinical reports that document various phenotypes and their clinical interpretations [4]. This task requires substantial domain knowledge and extensive clinical experience, which incur long-term training costs and require continual updates as new EEG phenotypes or waveforms are identified [5].

Limitations of prior works. Although several deep learning approaches have been proposed for EEG report generation [5, 6, 7], several limitations remain. Conceptually, many existing methods formulate report generation as a phenotype classification task followed by text decoding. Such pipelines fail to model EEG-to-report generation in an end-to-end manner, leading to an objective misalignment between *classification* and *generation*. Methodologically, they rely on short EEG contexts (spanning only several to tens of seconds) and fixed-context templates that construct checklist-style mappings between EEG phenotypes and predefined clinical notes. They fail to model long-term reasoning required for interpreting long-duration EEG data, further preventing multi-granularity report generation. Practically, due to the limited resource, these methods are developed as task-specific models, each tailored to a narrow reporting objective [6]. However, clinical practice requires multi-level reporting, where neurologists routinely construct overall EEG summaries and section-wise outputs, including impressions, interpretations, and event or seizure annotations.

Present work. To fill the gap, we present CELM, to the best of our knowledge, the first **C**linical EEG-to-language foundation model (**ELM**) that unlocks meaningful interpretation of raw clinical EEG

*Corresponding authors

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

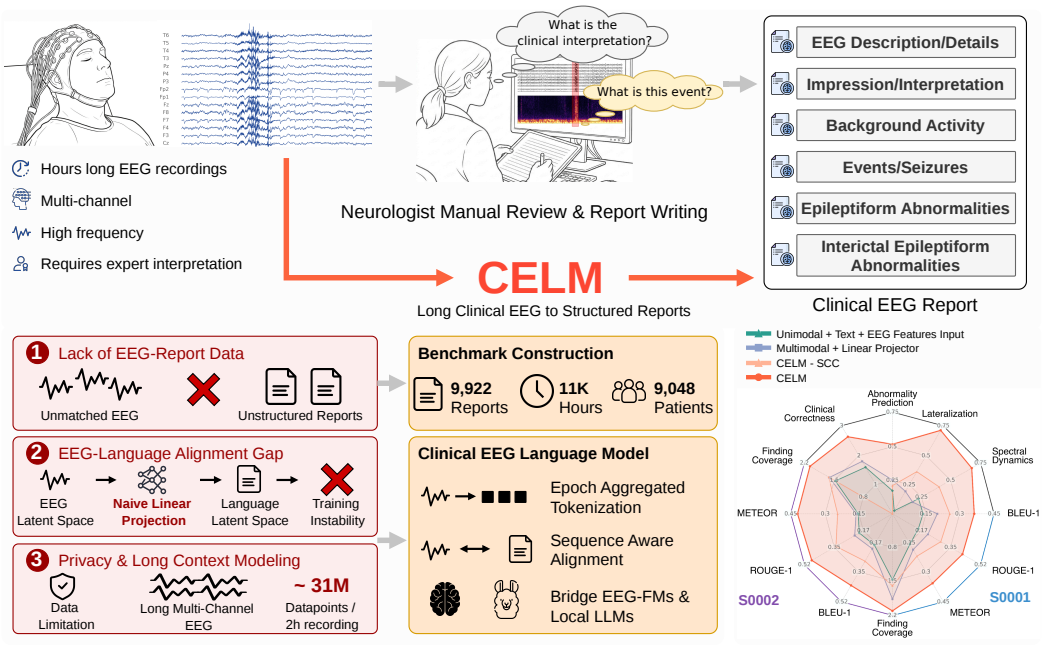

Figure 1: Overview of CELM, a clinical EEG-to-language foundation model (ELM) for translating long-duration EEG recordings into structured clinical reports. CELM addresses key challenges in EEG report generation, including limited EEG-report data, EEG-language alignment, and long-context EEG modeling. The framework constructs a large-scale EEG-report benchmark and enables direct generation of structured clinical reports from raw EEG recordings.

signals and enable report generation at multiple scales. Our motivation stems from recent advances in multimodal foundation models that enable text generation from non-text modalities (e.g., images and speech) [8, 9]. CELM hence shifts the paradigm from conventional EEG classification toward language-centric EEG understanding and describing, with three key innovations: ❶ A report structuring pipeline that automatically decomposes clinical notes and generates hierarchical structured reports, spanning EEG description, background activity, epileptiform abnormalities, events/seizures, and impressions/interpretations. ❷ A novel pretrained foundation model that processes EEG recordings of up to *three hours*, learns multi-grained representations from local patterns to global temporal contexts, and translates representations into text for end-to-end report generation. This is achieved by three core components: (i) epoch-aggregated tokenization that produces compact representations from variable-length recordings, (ii) sequence-aware alignment that captures inter-epoch temporal dependencies, and (iii) prompt fusion for conditional report generation. Importantly, our model design harnesses the representation capabilities of pretrained EEG foundation models and LLMs while enabling flexible generation across multiple clinical settings. ❸ A large-scale EEG-report benchmark curated from the largest clinical EEG corpora to date, the Harvard Electroencephalography Database [10, 11] (HEEDB), comprising 9,922 clinical reports paired with over 11,000 hours of EEG recordings from 9,048 patients, enabling CELM training and facilitating future model development.

To rigorously evaluate CELM, we benchmark it against state-of-the-art (SOTA) methods across five real-world report drafting tasks. Moreover, we conduct **human evaluation** with six clinical experts to assess the quality and clinical reliability of the generated reports. Extensive experiments demonstrate that CELM not only achieves significant improvements across quantitative metrics, but also produces reports that are better aligned with human interpretation. Our contributions are as follows:

- We introduce a *novel EEG-to-text foundation model* capable of learning representations from long-duration EEG recordings (up to three hours) and directly generating clinical reports.
- We propose *an automated report structuring pipeline* for raw EEG-text clinical data, which integrates seamlessly with the HEEDB to produce our benchmark dataset.
- Beyond benchmarking CELM, we also conduct *human evaluation* to assess whether the generated reports are more diagnostically reliable. Results show that CELM is statistically significantly more clinically coherent than those produced by prior methods and general language models.

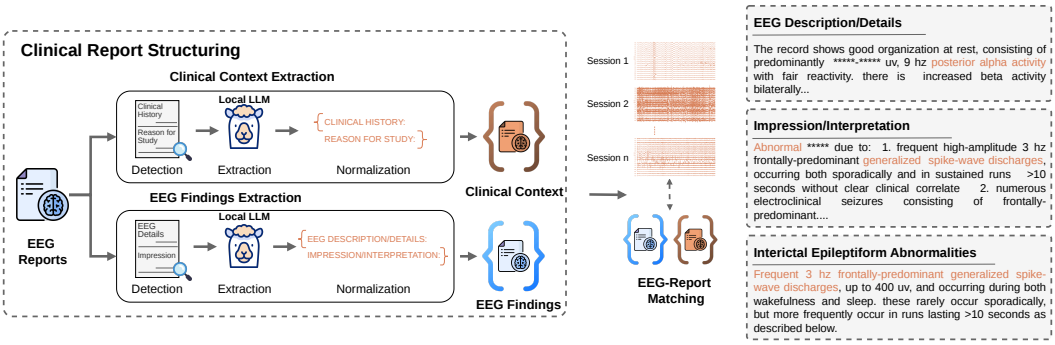

Figure 2: EEG–Report benchmark construction pipeline, including clinical report structuring (Section 3), matching reports to EEG sessions, and examples of standardized report sections.

2 Related Work

EEG-to-Language Modeling. Existing work at the intersection of EEG and natural language can be broadly categorized into two paradigms: (1) EEG-to-language decoding and (2) text-enhanced EEG representation learning. The goal of EEG-to-language decoding is to reconstruct textual content from concurrent EEG recordings of subjects reading or imagining speech [12, 13]. This line of work spans both invasive approaches using electrocorticography (ECoG) [14, 15] and non-invasive methods based on scalp EEG [13, 16, 17]. However, these methods assume precise EEG–text alignment, whereas clinical EEG consists of heterogeneous events embedded in long, continuous recordings spanning hours to days [10, 11]. The second paradigm leverages clinical notes as auxiliary supervision to enhance EEG representation learning, rather than decoding language directly from neural activity. Inspired by vision–language pretraining frameworks [18], a recent work [19] aligns EEG data with clinical report text in a shared feature space. However, it remains focused on discriminative objectives and does not support automated clinical report generation.

Clinical EEG Report Generation. An early attempt at EEG report generation, EEGtoText [5], proposed a two-stage pipeline that first classifies EEG phenotypes, and then generates report text via a text decoder conditioned on the classified labels. However, this relies on intermediate phenotyping as a bottleneck, limiting the capabilities to capture nuanced clinical findings beyond predefined categories. While some attempts jointly learn EEG encoders and text decoders [6], they rely on fixed segmentation and template-based generation [7]. As a result, these methods fall short of end-to-end clinical report generation from long-duration EEG recordings. We address the above challenges by introducing the first family of EEG–Language foundation models.

EEG Foundation Models. Recent years have witnessed rapid advances in EEG foundation models. BENDR [20], BIOT [1], LaBraM [21], TFM-Tokenizer [22], EEGPT [23], LUNA [24], REVE [25], and CBraMod [26] introduce increasingly scalable tokenization and representation learning frameworks that achieve strong transfer on diverse EEG tasks. These models are predominantly encoder-only and are optimized for classification tasks, leaving generative objectives largely unexplored. Our `CELM` is designed for clinical EEG report generation, and notably, it is fully compatible with existing foundation models, which can be directly leveraged as EEG encoders.

3 EEG-Report Benchmark Construction

Currently, no publicly available benchmark provides structured EEG–report pairs for training and evaluating EEG-to-text models. In this work, we leverage two resources from the *Brain Data Science Platform*: (1) the Harvard Electroencephalography Database v4.1 [11, 10], which contains approximately 12K clinical reports paired with about 100K hours of EEG recordings from 10,886 patients across multiple hospital sites, and (2) the Electronic Health Records (EHR) Repository with corresponding unstructured neurology reports. The EHR repository links reports to one or more EEG sessions via temporal alignment between EDF start times and report timestamps. Therefore, we construct an EEG–report benchmark using a dedicated processing pipeline (Figure 2).

Clinical Report Structuring. Neurology reports can be broadly decomposed into two types of sections: (1) *Clinical Context*, which contains information such as patient history and monitoring indications that cannot be inferred from EEG signals, and (2) *EEG Findings*, which describe EEG

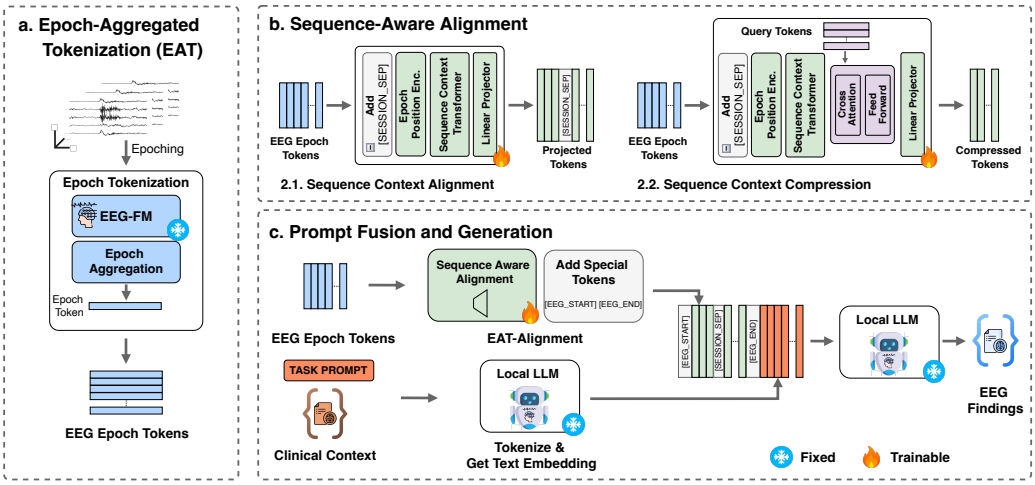

Figure 3: The proposed Clinical EEG Language Model (CELM) comprises (a) Epoch-Aggregated Tokenization, (b) Sequence-Aware Alignment, and (c) Prompt Fusion and Generation.

observations and clinical interpretations that are directly inferable from EEG recordings. A key challenge lies in reliably extracting and structuring these sections without information loss. Substantial variability in reporting styles and formatting makes manual extraction infeasible at scale. We therefore propose a three-stage automated pipeline to extract and structure reports.

- **Section Detection:** We curate a comprehensive list of candidate section headers under each category and detect their presence using string matching.
- **Section-wise Extraction:** For each detected section, we employ a local LLM (Meta-Llama-3-8B-Instruct) to extract the corresponding text using simple copy-based prompts. This detect-then-segment strategy avoids hallucinations and cross-section contamination that arise when extracting multiple sections jointly.
- **Canonical Normalization:** We standardize the extracted sections into canonical categories, including EEG description/details, impressions/interpretations, background activity, epileptiform abnormalities, and events/seizures.

EEG Preprocessing. Matched EEG recordings are preprocessed using a standard pipeline. Signals are first band-pass filtered between 0.1–75 Hz and notch filtered at 60 Hz to remove power-line interference. The recordings are then resampled to 200 Hz to ensure compatibility with existing EEG-only foundation models. Each recording is segmented into non-overlapping 10-second windows, which is the typical temporal context used by current EEG foundation architectures. Finally, EEG channels are standardized to a 22-channel montage following the International 10-20 system. While the dataset contains reports matched to multiple EEG sessions, we filter to samples with single-session matches for training, validation, and evaluation.

4 CELM: Clinical EEG Language Model

This section introduces our `CELM`, consisting of three key components: *Epoch-Aggregated Tokenization*, *Sequence-Aware Alignment*, and *Prompt Fusion and Generation*. An effective model must identify and aggregate clinically relevant information from long-term EEG recordings to support report generation at multiple levels. To this end, Section 4.1 discusses the unique challenges of clinical ELMs and provides an overview of the model architecture and forward process. Sections 4.2, 4.3, and 4.4 then present the details of the three components.

4.1 Architecture and Forward Process

Clinical EEG introduces three unique challenges that prevent the direct adoption of existing multimodal paradigms, such as dictionary-based alignment or contrastive learning with fixed-length epochs [19]. Here, we design `CELM` with three corresponding steps:

Epoch-Aggregated Tokenization addresses the extreme temporal scale of clinical EEG. EEG sessions span hours of high-frequency, multi-channel signals, resulting in at least $\sim 31.7\text{M}+$ data points per recording, far exceeding the context limits of modern LLMs [27, 28, 29].
Sequence-Aware Alignment mitigates the loss of temporal context during alignment. EEG signals are inherently sequential, and naive projection into the LLM embedding space fails to preserve long-range temporal dependencies required for clinical interpretation.
Prompt Fusion and Generation addresses weak EEG–text correspondence. Clinical reports aggregate findings over entire recordings without explicit temporal grounding, requiring the model to synthesize coherent clinical narratives from distributed EEG evidence.

Overall, given an input EEG recording, `CELM` processes the signal through above three stages. The recording is first transformed into compact epoch-level tokens via *Epoch-Aggregated Tokenization*, which reduces the temporal scale. These tokens are then projected into the language space through *Sequence-Aware Alignment*, enabling long-range temporal dependencies to be maintained across epochs. Finally, *Prompt Fusion and Generation* conditions the language model on the aligned EEG representations and optional clinical context to generate structured reports.

4.2 Epoch-Aggregated Tokenization

The fundamental bottleneck in developing ELMs is representing long EEG recordings within the LLM context constraints. A typical two-hour clinical EEG recording at 200 Hz across 22 channels yields $7,200 \times 200 \times 22 \approx 31.7$M data points, far exceeding current LLM capacities, making raw sample-level tokenization infeasible. A common alternative is mini-window tokenization, leveraging pretrained EEG encoders that typically tokenize 1-second segments. However, naively applying this granularity to full recordings produces ≈ 158K tokens per session, still well beyond the context limits of most local LLMs [27, 28, 29]. To overcome this limitation while harnessing pretrained EEG representations, we propose Epoch-Aggregated Tokenization (see Figure 3a).

Given an EEG recording session $\mathbf{X} \in \mathbb{R}^{N \times C \times T}$, where N, C, and T denote the number of epochs, channels, and time points per epoch (typically 10 seconds) respectively, we first tokenize 1s windows within each epoch using a pretrained EEG encoder. The resulting representations are then aggregated across windows and channels to produce a single token per epoch, yielding a compact sequence of epoch tokens $\mathbf{E}_{\text{eeg}} \in \mathbb{R}^{N \times D_{\text{eeg}}}$, where D_{eeg} is the embedding dimension. The aggregation strategy is encoder-dependent. For example, CBraMod applies pooling [26], whereas TFM-Tokenizer [22] uses `[CLS]` pooling. This tokenization compresses the tokens by up to $C \times T$ ($\sim 220\times$ compared to window tokenization), enabling EEG recordings to fit within LLM context limits.

4.3 Sequence-Aware Alignment

To enable EEG-conditioned clinical report generation, EEG representations must be aligned with the text embedding space of LLMs. Existing vision-language models such as LLaVA achieve this through simple linear projection [30], but this strategy does not generalize well to EEG (Section 5.6) due to its sequential and long-range temporal structure. Unlike static images, EEG signals contain temporal dynamics across epochs that must be preserved during alignment. This introduces two key challenges: modeling long temporal dependencies and scaling to very long sequences. To address these challenges, we propose two sequence-aware alignment strategies (Figure 3b): (1) Sequence Context Alignment, which preserves full temporal structure, and (2) Sequence Context Compression, which compresses EEGs into fixed-length latent representations for memory-efficient alignment.

Sequence Context Alignment (SCA). Given a sequence of tokens $\mathbf{E}_{\text{eeg}}$, we augment the sequence with position encodings and `[SESSION_SEP]` tokens to differentiate EEG sessions. A lightweight linear-attention transformer [31] is then applied to capture the temporal structure across epochs. The resulting representations are subsequently projected through a linear layer into the language embedding space, yielding $\mathbf{H}_{\text{eeg}} \in \mathbb{R}^{N \times D_{\text{llm}}}$, where D_{llm} denotes the LLM embedding dimension.

Sequence Context Compression (SCC). To address the memory and scalability challenges of long EEG sequences, we explore a compression strategy inspired by Perceiver architectures [32, 33]. After encoding temporal structure among epoch tokens, a fixed set of learnable query tokens $\mathbf{Q} \in \mathbb{R}^{L \times D_{\text{eeg}}} (L < N)$ attend to the full sequence of epoch tokens via cross-attention. This compresses

the variable-length EEG sequence into a fixed number of tokens, which are subsequently projected into the language embedding space, enabling efficient alignment for recordings of arbitrary duration.

4.4 Prompt Fusion and Generation

CELM generates clinical reports by conditioning a local LLM on EEG-derived representations jointly with clinical context when available (see Figure 3c). Given projected EEG tokens $\mathbf{H}_{\text{eeg}}$, we augment the sequence with special tokens [EEG_START] and [EEG_END] to distinguish EEG representations from text inputs. A task-specific prompt that specifies the target report sections, together with optional clinical context (e.g., reason for study or clinical history), is tokenized using the LLM tokenizer to obtain text embeddings $\mathbf{H}_{\text{prompt}}$. The final input sequence is formed by concatenation,

$$\mathbf{H}_{\text{in}} = [\texttt{[EEG_START]}; \mathbf{H}_{\text{eeg}}; \texttt{[EEG_END]}; \mathbf{H}_{\text{text}}] \tag{1}$$

which is processed autoregressively to generate a structured EEG report text. We employ instruction tuned Qwen3-4B [29] as our base LLM, CBraMod [26] as our EEG encoder and perform supervised fine-tuning using the next-token prediction objective:

$$\mathcal{L} = -\sum_{t=1}^{T} \log P_\theta(y_t \mid H_{\text{input}}, y_{<t}) \tag{2}$$

During training, both the EEG encoder and LLM remain frozen, only the Sequence-Aware Alignment module is updated, enabling efficient adaptation while preserving pretrained representations.

5 Experiments and Results

Datasets. The EEG-Report benchmark is constructed using data from two hospital sites in the Harvard Electroencephalography Database v4.1 [11, 10]: Massachusetts General Hospital (MGH, S0001) and Brigham and Women's Hospital (BWH, S0002). In total, through our pipeline, we curated $12,290$ clinical EEG reports matched to one or more EEG sessions from $10,886$ patients. For model training and evaluation, we restrict our analysis to reports paired with a single EEG session.
❒ *S0001:* This site includes diverse EEG visit types (routine EEG, epilepsy monitoring unit (EMU), portable EEG, complex EEG). We filtered only routine EEG and EMU samples paired with single sessions, excluding long recordings exceeding $10,000$ seconds to prevent memory issues during training. We curated $5{,}049$ reports from $4{,}669$ patients.
❒ *S0002:* From the available visit types (EEG and portable EEG), we retained only standard EEG visits with single corresponding sessions, yielding $4,873$ reports from $4,379$ patients.
For both sites, we performed patient-wise 60/20/20 splits into training, validation, and test sets. Additional dataset details are provided in Appendix A.

Baselines. Due to the lack of existing EEG multimodal models for clinical EEG report generation, we compare CELM against three types of baselines: (1) *Unimodal + text-only input*: only clinical context is provided as input without EEG data, providing a lower bound that quantifies hallucination and isolates performance when the model does not attend to neural signals, (2) *Unimodal + text and EEG features input*: the clinical context is augmented with channel-wise spectral band power features (delta, theta, alpha, beta, gamma, Appendix B.1) extracted from each EEG session, overcoming context length limitation that prevents providing raw EEG arrays directly, (3) *Multimodal + linear projector*: a linear projector maps pretrained EEG embeddings from CBraMod [26] into the language model's embedding space, enabling direct conditioning on continuous EEG signals. We evaluate LLMs from three model families, including LLaMA-3 [27], Gemma-3 [28], and Qwen-3 [29], across multiple parameter scales, and include MedGemma [34] as a domain-specific baseline (more details at Appendix B). We report lexical and semantic metrics [35, 36, 37] along with manual qualitative analysis, LLM-as-judge and a expert evaluation study to evaluate report generation performance.

5.1 Report Generation Performance

Table 1 compares the report generation performance of CELM on test samples from both sites containing patient history. It compares against unimodal baselines that use either text-only or text with handcrafted EEG features as input, and finetuned multimodal baselines that use EEG foundation model embeddings. We report results for two CELM variants: CELM-SCC, which uses sequence context

Table 1: Report generation performance comparison on samples with clinical context across two hospital sites (S0001 and S0002). We report mean ± standard deviation for ROUGE-1, METEOR and BERTScore. The best results are highlighted in orange, and the best baseline per category is highlighted in blue. Relative improvements are also reported, and ⋆ indicates finetuning.

	S0001			S0002		
Method	**ROUGE-1**	**METEOR**	**BERTScore**	**ROUGE-1**	**METEOR**	**BERTScore**
	Unimodal + Text Only Input					
Gemma-3-1b-it	0.2112 ± 0.1408	0.1567 ± 0.1238	0.6461 ± 0.3014	0.2330 ± 0.0876	0.1319 ± 0.0565	0.7058 ± 0.2423
Gemma-3n-e2b-it	0.2410 ± 0.1722	0.1875 ± 0.1725	0.7848 ± 0.0765	0.2318 ± 0.0600	0.1181 ± 0.0472	0.7683 ± 0.0560
Gemma-3-4b-it	0.2857 ± 0.1590	0.2147 ± 0.1668	0.7941 ± 0.0922	0.2775 ± 0.0558	0.1506 ± 0.0514	0.7851 ± 0.0730
Gemma-3n-e4b-it	0.2416 ± 0.1778	0.1816 ± 0.1786	0.7847 ± 0.0747	0.1998 ± 0.0649	0.0993 ± 0.0490	0.7635 ± 0.0525
Medgemma-4b-it	0.2196 ± 0.1924	0.1529 ± 0.1835	0.6683 ± 0.2994	0.2280 ± 0.0734	0.1198 ± 0.0542	0.7576 ± 0.1375
Llama-3.2-1b-instruct	0.0919 ± 0.0744	0.0529 ± 0.0519	0.5580 ± 0.3284	0.0954 ± 0.0869	0.0494 ± 0.0504	0.5139 ± 0.3501
Llama-3.2-3b-instruct	0.2224 ± 0.1755	0.1843 ± 0.1864	0.7116 ± 0.2311	0.2346 ± 0.0877	0.1430 ± 0.0689	0.7261 ± 0.1919
Llama-3.1-8b-instruct	0.2511 ± 0.1872	0.1839 ± 0.1788	0.6994 ± 0.2605	0.2655 ± 0.0766	0.1480 ± 0.0581	0.7502 ± 0.1531
Meta-llama-3-8b-instruct	0.2673 ± 0.1731	0.1946 ± 0.1809	0.7392 ± 0.2055	0.2588 ± 0.0586	0.1416 ± 0.0481	0.7702 ± 0.0817
Qwen3-1.7b	0.1378 ± 0.1557	0.1102 ± 0.1316	0.3856 ± 0.3941	0.1189 ± 0.1296	0.0735 ± 0.0821	0.3650 ± 0.3893
Qwen3-4b-instruct-2507	0.3038 ± 0.1460	0.2418 ± 0.1440	0.7975 ± 0.0852	0.2690 ± 0.0448	0.1622 ± 0.0434	0.7867 ± 0.0178
	Unimodal + Text + EEG Features Input					
Gemma-3-1b-it	0.0086 ± 0.0524	0.0064 ± 0.0462	0.0311 ± 0.1512	0.1755 ± 0.1157	0.0954 ± 0.0692	0.5785 ± 0.3384
Gemma-3n-e2b-it	0.0116 ± 0.0605	0.0088 ± 0.0543	0.0415 ± 0.1756	0.2190 ± 0.0840	0.1082 ± 0.0546	0.7175 ± 0.2153
Gemma-3-4b-it	0.0093 ± 0.0594	0.0078 ± 0.0558	0.0247 ± 0.1384	0.2886 ± 0.0458	0.1738 ± 0.0476	0.7883 ± 0.0418
Gemma-3n-e4b-it	0.0397 ± 0.0606	0.0231 ± 0.0392	0.2498 ± 0.3570	0.0357 ± 0.0903	0.0180 ± 0.0462	0.1073 ± 0.2678
Medgemma-4b-it	0.2100 ± 0.1919	0.1529 ± 0.1832	0.6172 ± 0.3319	0.1829 ± 0.1128	0.1000 ± 0.0702	0.5979 ± 0.3260
Llama-3.2-1b-instruct	0.0820 ± 0.0727	0.0473 ± 0.0479	0.4978 ± 0.3562	0.1016 ± 0.0773	0.0517 ± 0.0448	0.5903 ± 0.3062
Llama-3.2-3b-instruct	0.1844 ± 0.1611	0.1751 ± 0.1871	0.6441 ± 0.2898	0.2309 ± 0.0903	0.1540 ± 0.0743	0.7377 ± 0.1557
Llama-3.1-8b-instruct	0.2151 ± 0.1693	0.1844 ± 0.1647	0.6930 ± 0.2289	0.2256 ± 0.0904	0.1509 ± 0.0682	0.6895 ± 0.2171
Qwen3-1.7b	0.2051 ± 0.1453	0.1683 ± 0.1304	0.6063 ± 0.3215	0.2049 ± 0.1021	0.1413 ± 0.0724	0.6345 ± 0.2944
Qwen3-4b-instruct-2507	0.2058 ± 0.1192	0.2067 ± 0.1079	0.7249 ± 0.1840	0.2638 ± 0.0341	0.2017 ± 0.0285	0.7704 ± 0.0125
	Multimodal + Linear Projector + CBraMod⋆					
Gemma-3-4b-it	0.2688 ± 0.1448	0.2020 ± 0.1579	0.7904 ± 0.0798	0.3084 ± 0.0491	0.2002 ± 0.0482	0.7879 ± 0.0729
Medgemma-4b-it	0.0494 ± 0.1806	0.0441 ± 0.1719	0.0886 ± 0.2601	0.1491 ± 0.1349	0.0729 ± 0.0701	0.4492 ± 0.3899
Llama-3.2-3b-instruct	0.0568 ± 0.1311	0.0396 ± 0.1172	0.2402 ± 0.3583	0.2425 ± 0.0837	0.1398 ± 0.0610	0.7333 ± 0.1900
Llama-3.1-8b-instruct	0.0063 ± 0.0428	0.0050 ± 0.0369	0.0189 ± 0.1169	0.0385 ± 0.0823	0.0209 ± 0.0458	0.1533 ± 0.3011
Qwen3-4b-instruct-2507	0.2819 ± 0.1132	0.2444 ± 0.1093	0.7921 ± 0.0417	0.2998 ± 0.0425	0.2143 ± 0.0349	0.7852 ± 0.0153
`CELM-SCC`	**0.3843** ± 0.1876	**0.2889** ± 0.1866	**0.7949** ± 0.1783	**0.4487** ± 0.1283	**0.3232** ± 0.1261	**0.8229** ± 0.1341
`CELM`	**0.5565** ± 0.1683	**0.4734** ± 0.1941	**0.8697** ± 0.0901	**0.6408** ± 0.1494	**0.5597** ± 0.1728	**0.8755** ± 0.1521
Improvement	+83.18%	+93.70%	+9.05%	+107.78%	+161.18%	+11.06%

compression to compress and produce a fixed number of projected EEG tokens, and `CELM`, which uses non-compressed EEG representations via sequence context alignment. Incorporating EEG features yields modest gains over text-only on S0002 (0.1622 → 0.2017) except for LLaMA-3.1-8B, which fails due to its 8K context limit. On S0001, performance degrades due to greater multi-section report complexity. Finetuned multimodal baselines, which align EEG foundation embeddings with the language space via linear projection, achieve the best baseline performance. However, the marginal gains relative to the additional training overhead underscore the fundamental challenge of EEG-language alignment.

Table 2: Zero-context report generation performance on S0001 and S0002. The best results are highlighted in orange, and the best baseline is highlighted in blue.

	S0001		S0002	
Method	**ROUGE-1**	**METEOR**	**ROUGE-1**	**METEOR**
	Unimodal + Text + EEG Features Input			
Gemma-3-4b-it	0.20 ± 0.09	0.15 ± 0.07	0.09 ± 0.13	0.05 ± 0.08
Medgemma-4b-it	0.17 ± 0.12	0.10 ± 0.07	0.13 ± 0.10	0.07 ± 0.06
Llama-3.2-3b-it	0.14 ± 0.11	0.10 ± 0.08	0.20 ± 0.09	0.12 ± 0.06
Qwen3-4b-it	0.16 ± 0.07	0.16 ± 0.06	0.23 ± 0.08	0.17 ± 0.06
	Multimodal + Linear Projector + CBraMod			
Gemma-3-4b-it	0.24 ± 0.05	0.17 ± 0.05	0.29 ± 0.06	0.19 ± 0.05
Medgemma-4b-it	0.03 ± 0.09	0.02 ± 0.06	0.18 ± 0.11	0.08 ± 0.05
Llama-3.2-3b-it	0.02 ± 0.06	0.01 ± 0.04	0.13 ± 0.12	0.07 ± 0.07
Qwen3-4b-it	0.25 ± 0.05	0.22 ± 0.04	0.29 ± 0.03	0.21 ± 0.03
`CELM-SCC`	**0.27** ± 0.17	**0.21** ± 0.14	**0.39** ± 0.12	**0.26** ± 0.10
`CELM`	**0.51** ± 0.16	**0.41** ± 0.17	**0.56** ± 0.20	**0.49** ± 0.21
Improvement	+104%	+86%	+93%	+133%

`CELM` outperforms all baselines across all metrics, with statistically significant gains ($p < 0.001$, paired Wilcoxon signed-rank test) and an approximately two-fold improvement in ROUGE-1 and METEOR over the strongest baseline on both sites (e.g., S0002: 0.3084 → 0.6408). Although `CELM-SCC` outperforms the baselines, its performance gap relative to the non-compressed model (0.4487 vs 0.6408) highlights a key trade-off between compression and performance for ELMs. This motivates future research in compression and alignment strategies that preserve clinically relevant information while scaling to longer EEG sequences.

5.2 Zero-Context Report Generation

To assess our model's ability to generate clinical reports solely from EEG signals, without relying on clinical context, we define a zero-context report generation task in which only EEG data is provided

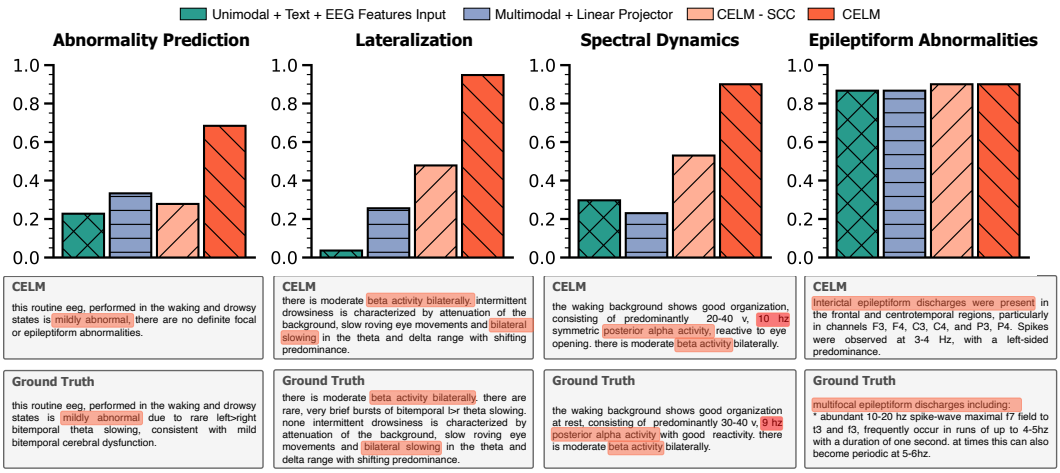

Figure 4: Manual qualitative analysis across four clinically relevant dimensions: abnormality prediction, lateralization, spectral dynamics, and epileptiform abnormalities. Representative example cases are provided below each dimension.

as input. This experiment also studies the potential language-model bias, in which reports could be generated by inferring plausible clinical context while ignoring EEG inputs. We conduct this experiment on both sites, using only the EEG signal as input, and summarize the results in Table 2. `CELM` outperforms all baselines across all metrics, demonstrating that it effectively extracts and utilizes clinically relevant information from EEG recordings.

5.3 Manual Qualitative Analysis

While lexical metrics capture overlap, they fail to assess clinical correctness, and no reliable automated alternative exists. To address this, we manually inspect 30 randomly sampled reports from both sites and evaluate correctness across four clinically relevant dimensions. Results and representative examples are presented in Figure 4. ❒ **Abnormality prediction.** We evaluate whether the generated report correctly classifies the recording as normal/abnormal. `CELM` achieves higher accuracy than all baselines. However, it frequently misclassifies mildly abnormal recordings as normal. One important observation is that even when the model accurately describes abnormal signal patterns (e.g., "frequent bilateral but right-predominant temporal sharply contoured theta and delta slowing"), it often fails at decision-making. This suggests a lack of domain knowledge, in which injecting neurology guidelines presents a promising approach. ❒ **Lateralization.** We assess whether the generated report correctly identifies the hemisphere in which brain activities occur. `CELM` performs significantly better than all baselines on this task. Notably, the unimodal baseline with EEG features and text input captures frequency information but fails to localize brain activity spatially. In contrast, `CELM` and other baselines benefit from the learned channel embeddings of EEG foundation models, which enable better lateralization. ❒ **Spectral dynamics.** We study whether the generated report accurately captures frequency band behavior, which is a critical aspect of EEG interpretation. Interestingly, the unimodal baseline with handcrafted EEG features outperforms the multimodal baseline using linearly projected CBraMod embeddings, indicating that task-specific features can partially enable LLMs to infer specific EEG-related tasks. However, crafting such features is labor-intensive, motivating the use of foundation model embeddings. While naive linear projection fails to leverage these embeddings effectively, `CELM` achieves strong performance, demonstrating the importance of proper alignment strategies. ❒ **Epileptiform abnormalities.** We evaluate whether the generated report correctly identifies epileptiform activity. While all models perform reasonably well on this dimension, we note that epileptiform events are rare and highly imbalanced in the dataset.

5.4 Human Expert Evaluation Study

To further strengthen our evaluation, we conducted an expert study in which six domain experts (clinicians and BME researchers) assessed report quality through an anonymized survey. Each expert evaluated 20 randomly sampled cases, with each case presenting the ground-truth clinical

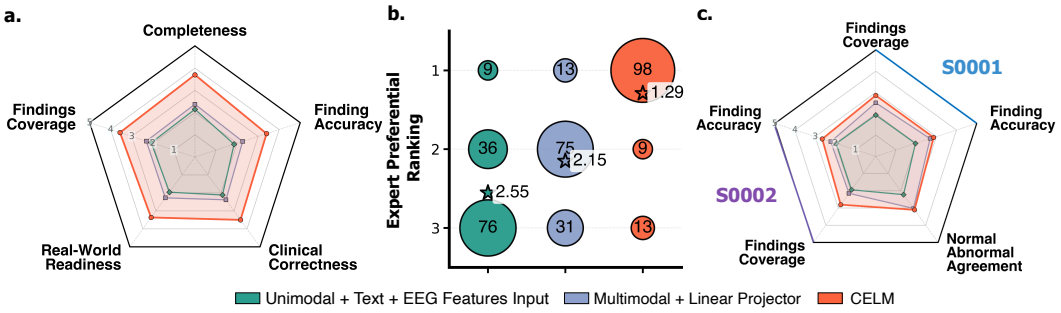

Figure 5: (a) Human expert evaluation study results, (b) Human expert preferential ranking for each report. (c) LLMs-as-a-judge results

report alongside three anonymized generated reports from two different baseline categories and CELM. Experts rated each report on a 5-point scale (1: very poor to 5: very good) across five clinically relevant dimensions: (1) Clinical Correctness - accuracy compared to the ground truth, (2) Findings Coverage - coverage of key findings identified relative to ground truth, (3) Finding Accuracy - accuracy of attributes such as lateralization, frequency, and morphology, (4) *Real-World Readiness* - suitability for clinical use, and (5) *Completeness* - completeness of the report, whether all relevant aspects are covered. Results are summarized in Figure 5a, where CELM significantly outperforms both baselines across all five dimensions (Wilcoxon signed-rank test, $p < 0.001$ for all comparisons). Experts also provided their preferred rankings across the reports, as shown in Figure 5b, where CELM obtains a weighted-average rank of 1.29, with 98 votes for rank 1. Inter-rater reliability was assessed using three measures: (1) ICC(2,k) was 0.79 overall (95% CI 0.74–0.82), with per-dimension values of 0.82 for Finding Accuracy (95% CI 0.75–0.89), 0.79 for Findings Coverage (95% CI 0.70–0.87), 0.79 for Clinical Correctness (95% CI 0.70–0.87), 0.76 for Real-World Readiness (95% CI 0.65–0.84), and 0.76 for Completeness (95% CI 0.65–0.84), all in the good range [38], (2) Adjacent agreement (within ± 1 point on the 5-point scale) was 74%-80% across dimensions, and (3) Kendall's W indicated statistically significant concordance across all dimensions ($W = 0.40$–0.49, $p < 0.001$).

LLM-as-judge. Human expert evaluation is labor-intensive and difficult to scale. We conducted an LLM-as-judge evaluation of generated reports using the Qwen2.5-72B-Instruct-AWQ model (Figure 5c). Although closed LLMs (e.g., GPT-4, Claude) correlate well with human judgments, data-usage agreement prohibits their use. We define three evaluation metrics, as in the expert study, and CELM consistently outperforms.

5.5 EEG Grounding Control Analysis

Clinical EEG reports often follow standardized structures and templates, raising the possibility that CELM learns site-specific reporting conventions rather than clinically relevant EEG representations. While Section 5.2 demonstrates that CELM can generate reports without clinical context, we further investigate its dependence on EEG signals through five inference-time perturbation experiments on the S0002 test set, without any retraining (Table 3): ❒ **Mismatch**: pairing each reference report with a randomly selected EEG recording from a different patient, ❒ **Channel permutation**: randomly permuting EEG channels across the montage, ❒ **Shuffle epochs**: randomly permuting the temporal order of EEG epochs, ❒ **Gaussian noise**: replacing EEG signals with Gaussian noise scaled to the original signal's standard deviation, and ❒ **Repeat single epoch**: repeating a randomly selected epoch to match the original recording length.

Table 3: Report generation performance under inference time EEG perturbations in S0002. We report mean ± standard deviation with the relative drop %.

Perturbation	ROUGE-1	Drop %	BERTScore	Drop %
CELM	**0.641** ± 0.149		**0.876** ± 0.152	
Mismatch	0.390 ± 0.322	↓ 39.2%	0.544 ± 0.437	↓ 37.8%
Channel permutation	0.437 ± 0.328	↓ 31.7%	0.593 ± 0.430	↓ 32.3%
Shuffle epochs	0.394 ± 0.341	↓ 38.5%	0.531 ± 0.447	↓39.4%
Gaussian noise	0.390 ± 0.341	↓39.1%	0.526 ± 0.447	↓39.9%
Repeat single epoch	0.423 ± 0.326	↓ 33.8%	0.581 ± 0.432	↓33.7%

All five perturbations significantly degrade performance ($p < 0.01$, Wilcoxon signed-rank test), with relative reductions of 32-39% in ROUGE-1 and 32-40% in BERTScore. The substantial decline in BERTScore indicates that the degradation extends beyond lexical overlap. Epoch shuffling produces performance losses comparable to replacing EEG signals with noise, suggesting that CELM is sensitive to temporal organization. Similarly, channel permutation causes substantial degradation,

indicating reliance on spatial channel information, consistent with its lateralization performance (Figure 4). Collectively, these results provide evidence that CELM relies on EEG recordings rather than site-specific reporting templates alone.

5.6 Alignment Module Ablation

The alignment module is critical for ELMs, and naive linear projection has been shown to be insufficient. We ablate four alignment strategies: (1) *Linear projector*, projection via a single linear layer, (2) *Perceiver projector*, cross-attention with learnable queries followed by feedforward (similar to Perceiver resampler in Flamingo [39], and analogous to the Q-Former in BLIP-2 [40]) and our proposed (3) *SCA* and (4) *SCC* projectors (Section 4.3). As shown in Figure 6, SCA achieves the best performance. Training dynamics (Figure 6b) reveal that Linear and Perceiver projectors overfit, with increasing validation loss despite decreasing training loss, whereas SCA and SCC exhibit stable convergence. This confirms that modeling inter-epoch temporal dependencies prior to projection is essential for effective alignment.

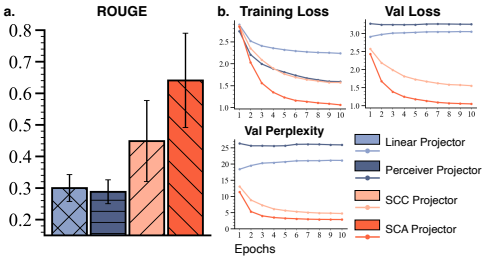

Figure 6: (a) Report generation performance of different alignment module variants. (b) Training dynamics of each variant, including training loss, validation loss, and perplexity curves.

6 Broader Impact, Limitations and Implications for Future Work

Limitations and Implications for Future work. In this work, we demonstrate the feasibility of end-to-end clinical EEG–language modeling for generating structured reports directly from long-duration EEG recordings. The strong performance gains over competitive baselines highlight the potential of ELMs for clinical report generation. However, several limitations remain. ☞*Failure modes.* CELM often misclassifies mildly abnormal EEGs as normal, even when correctly describing abnormalities, and struggles with rare findings such as interictal epileptiform abnormalities. ☞*Evaluation limitations.* ELM development is constrained by the lack of rigorous benchmarks and clinically grounded evaluation protocols, as common text generation metrics mainly capture lexical similarity rather than clinical correctness. ☞*Memory scalability.* Memory remains a key bottleneck. While our approach supports EEG recordings of up to approximately three hours, further advances in memory-efficient representations are needed to scale to longer recordings. ☞*Human-in-the-loop potential.* ELMs offer opportunities for human-in-the-loop workflows, where clinicians guide generation via targeted prompts instead of fixed templates, enabling better alignment with clinical intent.

Broader Impact and Intended Use Statement. This work advances machine learning research by introducing the first clinical EEG–language model that integrates long-duration clinical EEG signals with large language models to generate structured clinical reports. The proposed approach has the potential to support clinical workflows. However, this work is intended only as a research contribution rather than a deployable clinical system. CELM is an assistive drafting tool that needs neurologist review, and must not be used for autonomous reporting or diagnosis. Any real-world application would require extensive validation, regulatory approval, and careful oversight by experts.

7 Conclusion

We introduced CELM, the first clinical EEG-to-language foundation model for automated report generation from long-duration EEG recordings. CELM addresses three core challenges in this setting: (1) representing hour-scale EEG within LLM context limits via Epoch-Aggregated Tokenization, (2) preserving long-range temporal dependencies through Sequence-Aware Alignment, and (3) enabling flexible, multi-scale report generation via Prompt Fusion. We further presented a scalable EEG–Report benchmark construction pipeline and conducted extensive evaluations across hospital sites, report sections, and zero-context settings. Empirical results show that CELM consistently outperforms strong baselines. Overall, this work establishes EEG-to-language modeling as a distinct and promising research direction at the intersection of multimodal learning and clinical neurophysiology, and we hope that CELM and the accompanying benchmark construction pipeline will facilitate future advances in long-context EEG modeling and clinically grounded EEG language systems.

Acknowledgments, Disclosures and Ethics Statement

Funding Transparency Statement. This project has been funded by the Jump ARCHES endowment through the Health Care Engineering Systems Center. The authors declare no competing interests.

Ethics Statement. This retrospective study uses the de-identified Harvard EEG Database v4.1 [10, 11], accessed via BDSP with credentialed access, CITI training, and a Data Use Agreement under the BDSP Credentialed Health Data License 1.5.0. All data used in this study were de-identified by the dataset providers to protect participants' protected health information (PHI) and the authors had no access to identifiable information and no interaction with patients. All data processing and experiments were performed using locally hosted LLMs, without transmitting data to external APIs. We release only the benchmark construction pipeline and modeling code; clinical data and derived reports are not shared, and access to model checkpoints will be gated. `CELM` and all released artifacts are intended for research use only and not for clinical deployment, consistent with the dataset's data-use terms. The expert evaluation involved research collaborators and peers who voluntarily rated de-identified model outputs without compensation.

References

[1] Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.

[2] Rikuto Kotoge, Zheng Chen, Tasuku Kimura, Yasuko Matsubara, Takufumi Yanagisawa, Haruhiko Kishima, and Yasushi Sakurai. Evobrain: Dynamic multi-channel EEG graph modeling for time-evolving brain networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[3] Haohui Jia, Zheng Chen, Lingwei Zhu, Rikuto Kotoge, Jathurshan Pradeepkumar, Yasuko Matsubara, Jimeng Sun, Yasushi Sakurai, and Takashi Matsubara. ODEBrain: Continuous-time EEG graph for modeling dynamic brain networks. In *International Conference on Learning Representations*, pages 106978–106996, 2026.

[4] William O Tatum IV, Olga Selioutski, Juan G Ochoa, Heidi Munger Clary, Janna Cheek, Frank W Drislane, and Tammy N Tsuchida. American clinical neurophysiology society guideline 7: guidelines for eeg reporting. *The Neurodiagnostic Journal*, 56(4):285–293, 2016.

[5] Siddharth Biswal, Cao Xiao, M Brandon Westover, and Jimeng Sun. Eegtotext: learning to write medical reports from eeg recordings. In *Machine learning for healthcare conference*, pages 513–531. PMLR, 2019.

[6] Woonghee Lee, Jaewoo Yang, Doyeong Park, and Younghoon Kim. Automated clinical impression generation for medical signal data searches. *Applied Sciences*, 13(15), 2023.

[7] Kang Yin and Hye-Bin Shin. Neurolex: A lightweight domain language model for eeg report understanding and generation. *10.48550/arXiv.2511.12851*, 2025.

[8] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, Karthik Padthe, Vasu Sharma, Hu Xu, Xiaoqing Ellen Tan, Megan Richards, Samuel Lavoie, Pietro Astolfi, Reyhane Askari Hemmat, Jun Chen, Kushal Tirumala, Rim Assouel, Mazda Moayeri, Arjang Talattof, Kamalika Chaudhuri, Zechun Liu, Xilun Chen, Quentin Garrido, Karen Ullrich, Aishwarya Agrawal, Kate Saenko, Asli Celikyilmaz, and Vikas Chandra. An introduction to vision-language modeling, 2024.

[9] Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Matthew Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in VLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[10] Chenxi Sun, Jin Jing, Niels Turley, Callison Alcott, Wan-Yee Kang, Andrew J Cole, Daniel M Goldenholz, Alice Lam, Edilberto Amorim, Catherine Chu, et al. Harvard electroencephalography database: A comprehensive clinical electroencephalographic resource from four boston hospitals. *Epilepsia*, 2025.

[11] S Zafar, T Loddenkemper, JW Lee, A Cole, D Goldenholz, J Peters, A Lam, E Amorim, C Chu, S Cash, et al. Harvard electroencephalography database (version 4.1). *Brain Data Science Platform*, 2025.

[12] Christian Herff, Dominic Heger, Adriana De Pesters, Dominic Telaar, Peter Brunner, Gerwin Schalk, and Tanja Schultz. Brain-to-text: decoding spoken phrases from phone representations in the brain. *Frontiers in neuroscience*, 8:141498, 2015.

[13] Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5350–5358, 2022.

[14] Joseph G Makin, David A Moses, and Edward F Chang. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature neuroscience*, 23(4):575–582, 2020.

[15] Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Foram Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, et al. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

[16] Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. Dewave: Discrete encoding of eeg waves for eeg to text translation. *Advances in Neural Information Processing Systems*, 36:9907–9918, 2023.

[17] Jinzhao Zhou, Yiqun Duan, Fred Chang, Thomas Do, Yu-Kai Wang, and Chin-Teng Lin. Belt-2: Bootstrapping eeg-to-language representation alignment for multi-task brain decoding. *arXiv preprint arXiv:2409.00121*, 2024.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[19] Tidiane Camaret Ndir, Robin T Schirrmeister, and Tonio Ball. Eeg-clip: Learning eeg representations from natural language descriptions. *Frontiers in Robotics and AI*, 12:1625731, 2025.

[20] Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

[21] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024.

[22] Jathurshan Pradeepkumar, Xihao Piao, Zheng Chen, and Jimeng Sun. Single-channel eeg tokenization through time-frequency modeling. *arXiv preprint arXiv:2502.16060*, 2025.

[23] Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024.

[24] Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. Luna: Efficient and topology-agnostic foundation model for eeg signal analysis. *arXiv preprint arXiv:2510.22257*, 2025.

[25] Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. Reve: A foundation model for eeg–adapting to any setup with large-scale pretraining on 25,000 subjects. *arXiv preprint arXiv:2510.21585*, 2025.

[26] Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. *arXiv preprint arXiv:2412.07236*, 2024.

[27] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[28] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[29] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[31] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[32] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[33] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

[34] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[35] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.

[36] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[37] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[38] Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

[39] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[40] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PmLR, 2023.

[41] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

Appendix

Contents

A Details of EEG-Report Benchmark **14**

B Additional Experiment Details **15**

B.1 Band Power Features 15

B.2 Baselines and Prompts 15

B.3 Evaluation Metrics 16

C Implementation Details **18**

C.1 Model Details and Hyperparameters 18

C.2 Training Details 18

C.3 Prompts 18

D Additional Experiment Results **20**

D.1 Extended Report Generation Performance on S0001 20

D.2 Extended Report Generation Performance on S0002 20

D.3 Additional Results on Alignment Module Ablation 22

D.4 EEG Encoder Ablation 22

D.5 Performance Analysis by Report Section 22

D.6 Cross-Site Generalization 24

D.7 Score Distribution 24

D.8 SCA vs. SCC by Recording Duration 26

D.9 Benchmark Construction Quality Validation 26

D.10 Qualitative Analysis and Case Studies 26

E Human Expert Evaluation Protocol **30**

F Model Card **31**

A Details of EEG-Report Benchmark

This section provides detailed information on the EEG–Report benchmark constructed in our study. Table 4 summarizes key statistics, including the number of EEG–report pairs, patient counts, and total EEG recording duration for each site. Figure 7 further illustrates dataset distributions across EEG session duration, patient demographics (age and gender), and EEG report sections.

Table 4: Statistics of the EEG–Report benchmark.

Site	# of Paired EEG & Reports	# of Patients	Total Duration (hrs)
S0001	5049	4669	6639
S0002	4873	4379	4367

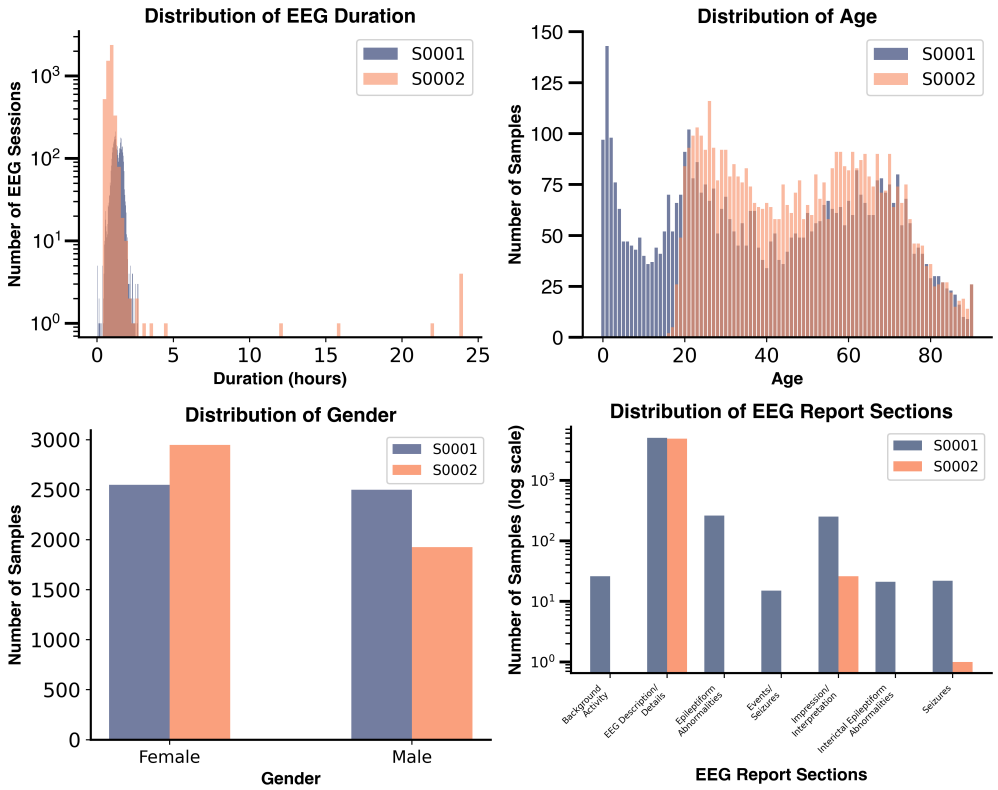

Figure 7: Dataset statistics of the filtered and constructed EEG-Report Benchmark used in our study

B Additional Experiment Details

B.1 Band Power Features

This section describes the extraction of handcrafted EEG features used in the unimodal text + EEG feature baselines. To construct EEG features that fit current context-length constraints, we extract band-power features from EEG recordings. For each 10-second multichannel EEG segment, we compute channel-wise band power using Welch's method by first estimating the power spectral density and then integrating it within five EEG frequency bands: delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–80 Hz). To further reduce input sequence length for LLMs with limited context windows, we optionally pool consecutive epochs by concatenating their time axes prior to spectral estimation, yielding a more compact representation. The resulting band power values are formatted as structured text and provided as input to the LLMs (see example in Figure 9).

Table 5: Summary of LLM Baselines

LLM	Family	Context Length
Gemma-3-1b-it	Gemma 3	32K
Gemma-3n-E2B-it	Gemma 3	32K
Gemma-3n-E4B-it	Gemma 3	32K
Gemma-3-4b-it	Gemma 3	128K
Medgemma-4b-it	Gemma	128K
Llama-3.2-1B-Instruct	Llama-3.2	128K
Llama-3.2-3B-Instruct	Llama-3.2	128K
Llama-3.1-8B-Instruct	Llama-3.1	128K
Meta-Llama-3-8B-Instruct	Llama-3	8K
Qwen3-1.7B	Qwen 3	32K
Qwen3-4B-Instruct-2507	Qwen 3	256K

B.2 Baselines and Prompts

In this section, we provide additional details on the baseline models and the prompts used in our experiments. Table 5 summarizes all LLMs considered in our study along with their respective context lengths[2], further underscoring the practical challenges of directly conditioning LLMs on long-duration EEG recordings. Figures 8 and 9 present

[2]Context length information is extracted from their respective technical reports and Hugging Face `https://huggingface.co/`

the prompts used for the unimodal text-only and unimodal text + EEG feature baselines, respectively. For clarity, we provide examples for certain prompt sections using bounding boxes.

Unimodal + Text Only Prompt

You are an expert clinical neurophysiologist specializing in EEG interpretation and clinical report generation.

TASK
Your task is to generate the specified sections of a formal clinical EEG report using only the provided:

- Patient history
- EEG description

EEG SECTION DESCRIPTIONS
[STANDARDIZED_SECTION_DESCRIPTIONS]

```
e.g. EEG DESCRIPTION/DETAILS: Detailed narrative of EEG findings
including background activity, sleep stages, physiologic variants,
and abnormalities observed during the recording period.
```

GUIDELINES

- Generate only the sections listed in **SECTIONS TO BE GENERATED**.
- Do NOT generate any additional sections.
- Do NOT repeat the same section more than once.
- Only generate the output in the JSON format and do not include any other text or explanation.

OUTPUT FORMAT (STRICT)
Return ONLY the following JSON structure, with no preamble, explanation, or markdown:

```
json
{"report_sections": [
    {"section_name": "Name of the section as given in SECTIONS TO BE
                      GENERATED",
     "section_text": "Generated text for the section in string"},
    ...
]}
```

SECTIONS TO BE GENERATED
[SECTION_NAMES]

```
e.g. ['EEG DESCRIPTION/DETAILS']
```

PATIENT HISTORY AND EEG DESCRIPTION
[PATIENT_HISTORY_AND_EEG_DESCRIPTION]

```
e.g. age: 77.0, gender: Female, indication: ***** y.o. female with
history of afib, l mca stroke, and headaches, presenting with episodes
of altered consciousness concerning for seizure vs. syncope.
none  pertinent medications: keppra, neurontin, seroquel.  none
```

Now generate the EEG report.

Figure 8: Prompt for unimodal + text only baselines.

B.3 Evaluation Metrics

We evaluate report generation performance using a comprehensive set of natural language generation metrics to evaluate different aspects of generation quality. In the main manuscript, we report ROUGE-1, METEOR and BERTScore. In Appendix D, we additionally report smoothed BLEU-1 and BLEU-4, ROUGE-2, ROUGE-L, and ROUGE-LSUM. We further include per-sample score distributions for all metrics in Appendix D.7.

Unimodal + Text + EEG Features Prompt

EEG-DERIVED STATISTICS

```
{"eeg_session_0":{"delta (0.5-4Hz) band power (dB)": [[-22.90, -21.60, ...]],
                  "theta (4-8Hz) band power (dB)": [[-25.30, -24.80, ...]],
                  "alpha (8-12Hz) band power (dB)": [[-27.50, -26.60, ...]],
                  "beta (12-30Hz) band power (dB)": [[-30.60, -30.20, ...]],
                  "gamma (30-80Hz) band power (dB)": [[-36.30, -36.10, ...]],
```

EEG CHANNELS

```
['C3', 'C4', 'O1', 'O2', 'Cz', 'F3', 'F4', 'F7', 'F8', 'Fz', 'Fp1', 'Fp2',
'Fpz', 'P3', 'P4', 'Pz', 'T3', 'T4', 'T5', 'T6', 'A1', 'A2']
```

You are an expert clinical neurophysiologist specializing in EEG interpretation and clinical report generation.

TASK

Your task is to generate the specified sections of a formal clinical EEG report using only the provided:

- Patient history
- EEG description
- EEG Channels
- EEG-derived statistics (provided above)

EEG SECTION DESCRIPTIONS

[STANDARDIZED_SECTION_DESCRIPTIONS]

```
e.g. EEG DESCRIPTION/DETAILS: Detailed narrative of EEG findings including
background activity, sleep stages, physiologic variants, and abnormalities
observed during the recording period.
```

GUIDELINES

- Generate only the sections listed in **SECTIONS TO BE GENERATED**.
- Do NOT generate any additional sections.
- Do NOT repeat the same section more than once.
- Only generate the output in the JSON format and do not include any other text or explanation.

OUTPUT FORMAT (STRICT)

Return ONLY the following JSON structure, with no preamble, explanation, or markdown:

```
json
{"report_sections": [
   {"section_name": "Name of the section as given in SECTIONS TO BE GENERATED",
    "section_text": "Generated text for the section in string"},
    ...]}
```

SECTIONS TO BE GENERATED

[SECTION_NAMES]

```
e.g. ['EEG DESCRIPTION/DETAILS']
```

PATIENT HISTORY AND EEG DESCRIPTION

[PATIENT_HISTORY_AND_EEG_DESCRIPTION]

```
e.g. age: 77.0, gender: Female, indication: ***** y.o. female with history
of afib, l mca stroke, and headaches, presenting with episodes of altered
consciousness concerning for seizure vs. syncope.
none  pertinent medications: keppra, neurontin, seroquel.  none
```

Now generate the EEG report.

Figure 9: Prompt for unimodal + text and EEG features as input baselines.

C Implementation Details

To facilitate reproducibility, we provide the complete source code for constructing the EEG–report benchmark from the Harvard Electroencephalography Database and the EHR database hosted on the Brain Data Science Platform, together with our model implementation, training scripts, and data loaders at `https://github.com/Jathurshan0330/CELM`. Additionally, we provide the detailed implementation details of our approach in this section and the prompts used for the baselines in Appendix B. The dataset utilized for this study is publicly available and can be accessed at `https://bdsp.io/content/harvard-eeg-db/4.1/` after obtaining credentialed access. Experiments were conducted using multiple NVIDIA RTX 6000 and NVIDIA A100 GPUs with 49GB and 80GB of memory.

C.1 Model Details and Hyperparameters

We use CBraMod [26] as the EEG encoder and Qwen-4B-Instruct-2507 as the local LLM backbone in our framework. CBraMod is selected due to its SOTA performance among EEG foundation models and its pretraining on large, diverse EEG datasets. In addition, our encoder ablation study comparing CBraMod with LaBraM shows that CBraMod consistently achieves better performance. We chose Qwen-4B-Instruct-2507 as the LLM backbone because it outperforms other local LLMs in unimodal baselines and reliably follows instruction prompts to generate outputs with the required structure. In Table 6, we provide the hyperparameters of the alignment modules used in our study.

Table 6: Hyperparameters of the alignment (projector) modules.

Module	Hyperparameter	Values
Linear Projector	Projector Input Dimension (CBraMod Embedding Dimension)	200
	Projector Output Dimension (Qwen-3 4B Embedding Dimension)	2560
	Bias	True
	# Trainable parameters	522,240
Perceiver Projector	# Query tokens	256
	Embedding Dimension	200
	# Attention heads	8
	# Perceiver layers (cross attention followed by feed-forward layer)	2
	Dropout	0.1
	Feed-forward multiplier ($\times$)	2
	Final linear projection	Linear(200, 2560)
	# Trainable parameters	1,219,040
Sequence Context Compression (SCC)	Transformer Encoder Type	Linear attention transformer
	Transformer # heads	8
	Transformer # layers (depth)	1
	# Query tokens	256
	# Perceiver layers	1
	# Trainable parameters	1,378,440
Sequence Context Alignment (SCA)	Transformer Encoder Type	Linear attention transformer
	# Heads	8
	# Transformer Layers	2
	Final linear projection	Linear(200, 2560)
	# Trainable parameters	1,486,240

C.2 Training Details

In our approach, the EEG encoder and LLM backbone are frozen, and only the alignment module is trained via supervised learning using a next-token prediction objective. Training details are summarized in Table 7.

Table 7: Training details

Hyperparameter	Values
Batch size	4
Gradient accumulation steps	4
Optimizer	AdamW
Learning rate	1e-4
Weight decay	0.01
β_1	0.9
β_2	0.99
LR scheduler	Linear scheduler with warmup
# of training epochs	10
Warm-up ratio	0.1
Mixed precision	bf16

C.3 Prompts

Figure 10 illustrates the prompt used in our approach along with the EEG tokens. We also provide examples of the sections in the prompt.

ELM Prompt

Input : EEG projected tokens prepended to text tokens.

EEG CHANNELS

```
['C3', 'C4', 'O1', 'O2', 'Cz', 'F3', 'F4', 'F7', 'F8', 'Fz', 'Fp1', 'Fp2', 'Fpz',
'P3', 'P4', 'Pz', 'T3', 'T4', 'T5', 'T6', 'A1', 'A2']
```

You are an expert clinical neurophysiologist specializing in EEG interpretation and clinical report generation.

TASK

Your task is to generate the specified sections (**SECTIONS TO BE GENERATED**) of a formal clinical EEG report using the above provided data of EEG recording sessions and the following information:

- Patient history
- EEG description
- EEG Channels

EEG SECTION DESCRIPTIONS

[STANDARDIZED_SECTION_DESCRIPTIONS]

```
e.g. EEG DESCRIPTION/DETAILS: Detailed narrative of EEG findings including
background activity, sleep stages, physiologic variants, and abnormalities
observed during the recording period.
```

GUIDELINES

- Generate only the sections listed in **SECTIONS TO BE GENERATED**.
- Do NOT generate any additional sections.
- Do NOT repeat the same section more than once.
- Only generate the output in the JSON format and do not include any other text or explanation.

OUTPUT FORMAT (STRICT)

Return ONLY the following JSON structure, with no preamble, explanation, or markdown:

```
json
{"report_sections": [
    {"section_name": "Name of the section as given in SECTIONS TO BE
                      GENERATED",
     "section_text": "Generated text for the section in string"},
    ... ]}
```

SECTIONS TO BE GENERATED

[SECTION_NAMES]

```
e.g. ['EEG DESCRIPTION/DETAILS']
```

PATIENT HISTORY AND EEG DESCRIPTION

[PATIENT_HISTORY_AND_EEG_DESCRIPTION]

```
e.g. age: 77.0, gender: Female, indication: ***** y.o. female with history
of afib, l mca stroke, and headaches, presenting with episodes of altered
consciousness concerning for seizure vs. syncope.
none  pertinent medications: keppra, neurontin, seroquel.  none
```

Now generate the EEG report.

Figure 10: Prompt for CELM.

D Additional Experiment Results

D.1 Extended Report Generation Performance on S0001

Table 8 presents a comprehensive performance comparison of EEG report generation on the S0001 site for samples with clinical context. Both proposed variants consistently outperform all baselines across all evaluation metrics. Among them, `CELM` achieves better performance compared to the memory-efficient `CELM-SCC` variant across all metrics.

Table 8: Report generation performance comparison on samples with clinical context on S0001. We provide results for a complete set of evaluation metrics under two input settings: (i) *Unimodal + Text Only Input*, where language models generate reports solely from clinical context text, and (ii) *Unimodal + Text + EEG Features Input*, where EEG-derived features are additionally provided. The table compares multiple strong baselines, including general-purpose and medical LLMs, against our `CELM`. The best results are highlighted in orange, and the best baseline per category is highlighted in blue.

Method	BLEU-1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSUM	BERTScore	METEOR
	Unimodal + Text Only Input							
Gemma-3-1b-it	0.1897 ± 0.1340	0.0351 ± 0.0694	0.2112 ± 0.1408	0.0431 ± 0.0985	0.1356 ± 0.1129	0.1381 ± 0.1137	0.6461 ± 0.3014	0.1567 ± 0.1238
Gemma-3n-e2b-it	0.1967 ± 0.1592	0.0641 ± 0.1402	0.2410 ± 0.1722	0.0702 ± 0.1939	0.1657 ± 0.1826	0.1687 ± 0.1824	0.7848 ± 0.0765	0.1875 ± 0.1725
Gemma-3-4b-it	0.2372 ± 0.1508	0.0641 ± 0.1349	0.2857 ± 0.1590	0.0821 ± 0.1893	0.1893 ± 0.1749	0.1929 ± 0.1747	0.7941 ± 0.0922	0.2147 ± 0.1668
Gemma-3n-e4b-it	0.1793 ± 0.1685	0.0642 ± 0.1486	0.2416 ± 0.1778	0.0763 ± 0.2003	0.1659 ± 0.1892	0.1687 ± 0.1890	0.7847 ± 0.0747	0.1816 ± 0.1786
Medgemma-4b-it	0.1502 ± 0.1725	0.0485 ± 0.1383	0.2196 ± 0.1924	0.0654 ± 0.1954	0.1569 ± 0.1908	0.1593 ± 0.1910	0.6683 ± 0.2994	0.1529 ± 0.1835
Llama-3.2-1b-instruct	0.0337 ± 0.0667	0.0059 ± 0.0166	0.0919 ± 0.0744	0.0053 ± 0.0217	0.0615 ± 0.0500	0.0629 ± 0.0515	0.5580 ± 0.3284	0.0529 ± 0.0519
Llama-3.2-3b-instruct	0.1951 ± 0.1630	0.0625 ± 0.1487	0.2224 ± 0.1755	0.0669 ± 0.1860	0.1511 ± 0.1760	0.1550 ± 0.1759	0.7116 ± 0.2311	0.1843 ± 0.1864
Llama-3.1-8b-instruct	0.2128 ± 0.1702	0.0598 ± 0.1444	0.2511 ± 0.1872	0.0753 ± 0.1967	0.1689 ± 0.1889	0.1722 ± 0.1892	0.6994 ± 0.2605	0.1839 ± 0.1788
Meta-llama-3-8b-instruct	0.2249 ± 0.1602	0.0620 ± 0.1418	0.2673 ± 0.1731	0.0765 ± 0.1895	0.1758 ± 0.1785	0.1795 ± 0.1787	0.7392 ± 0.2055	0.1946 ± 0.1809
Qwen3-1.7b	0.1335 ± 0.1481	0.0213 ± 0.0450	0.1378 ± 0.1557	0.0285 ± 0.0704	0.0835 ± 0.1054	0.0855 ± 0.1075	0.3856 ± 0.3941	0.1102 ± 0.1316
Qwen3-4b-instruct-2507	0.2795 ± 0.1243	0.0593 ± 0.1100	0.3038 ± 0.1460	0.0763 ± 0.1672	0.1872 ± 0.1623	0.1916 ± 0.1622	0.7975 ± 0.0852	0.2418 ± 0.1440
	Unimodal + Text + EEG Features Input							
Gemma-3-1b-it	0.0068 ± 0.0442	0.0019 ± 0.0273	0.0086 ± 0.0524	0.0022 ± 0.0367	0.0063 ± 0.0449	0.0064 ± 0.0451	0.0311 ± 0.1512	0.0064 ± 0.0462
Gemma-3n-e2b-it	0.0083 ± 0.0504	0.0029 ± 0.0330	0.0116 ± 0.0605	0.0032 ± 0.0403	0.0080 ± 0.0500	0.0083 ± 0.0507	0.0415 ± 0.1756	0.0088 ± 0.0543
Gemma-3-4b-it	0.0085 ± 0.0532	0.0025 ± 0.0299	0.0093 ± 0.0594	0.0032 ± 0.0402	0.0062 ± 0.0466	0.0065 ± 0.0474	0.0247 ± 0.1384	0.0078 ± 0.0558
Gemma-3n-e4b-it	0.0101 ± 0.0348	0.0020 ± 0.0067	0.0397 ± 0.0606	0.0014 ± 0.0072	0.0259 ± 0.0395	0.0262 ± 0.0401	0.2498 ± 0.3570	0.0231 ± 0.0392
Medgemma-4b-it	0.1618 ± 0.1730	0.0488 ± 0.1400	0.2100 ± 0.1919	0.0611 ± 0.1870	0.1455 ± 0.1858	0.1482 ± 0.1862	0.6172 ± 0.3319	0.1529 ± 0.1832
Llama-3.2-1b-instruct	0.0310 ± 0.0615	0.0053 ± 0.0119	0.0820 ± 0.0727	0.0037 ± 0.0150	0.0547 ± 0.0485	0.0556 ± 0.0493	0.4978 ± 0.3562	0.0473 ± 0.0479
Llama-3.2-3b-instruct	0.1692 ± 0.1414	0.0528 ± 0.1147	0.1844 ± 0.1611	0.0543 ± 0.1575	0.1271 ± 0.1529	0.1342 ± 0.1536	0.6441 ± 0.2898	0.1751 ± 0.1871
Llama-3.1-8b-instruct	0.2040 ± 0.1450	0.0508 ± 0.1177	0.2151 ± 0.1693	0.0617 ± 0.1695	0.1469 ± 0.1690	0.1499 ± 0.1693	0.6930 ± 0.2289	0.1844 ± 0.1647
Qwen3-1.7b	0.2047 ± 0.1351	0.0331 ± 0.0611	0.2051 ± 0.1453	0.0422 ± 0.0971	0.1286 ± 0.1150	0.1317 ± 0.1163	0.6063 ± 0.3215	0.1683 ± 0.1304
Qwen3-4b-instruct-2507	0.1809 ± 0.0986	0.0288 ± 0.0555	0.2058 ± 0.1192	0.0402 ± 0.0969	0.1275 ± 0.1082	0.1303 ± 0.1087	0.7249 ± 0.1840	0.2067 ± 0.1079
	Multimodal + Linear Projector + CBraMod							
Gemma-3-4b-it	0.2201 ± 0.1349	0.0552 ± 0.1087	0.2688 ± 0.1448	0.0726 ± 0.1652	0.1803 ± 0.1558	0.1834 ± 0.1556	0.7904 ± 0.0798	0.2020 ± 0.1579
Medgemma-4b-it	0.0392 ± 0.1478	0.0260 ± 0.1212	0.0494 ± 0.1806	0.0356 ± 0.1668	0.0445 ± 0.1764	0.0447 ± 0.1765	0.0886 ± 0.2601	0.0441 ± 0.1719
Llama-3.2-3B-instruct	0.0312 ± 0.1055	0.0146 ± 0.0816	0.0568 ± 0.1311	0.0182 ± 0.1059	0.0446 ± 0.1193	0.0448 ± 0.1192	0.2402 ± 0.3583	0.0396 ± 0.1172
Llama-3.1-8B-instruct	0.0042 ± 0.0312	0.0017 ± 0.0190	0.0063 ± 0.0428	0.0029 ± 0.0284	0.0050 ± 0.0368	0.0051 ± 0.0370	0.0189 ± 0.1169	0.0050 ± 0.0369
Qwen3-4B-instruct-2507	0.2647 ± 0.0908	0.0420 ± 0.0643	0.2819 ± 0.1132	0.0595 ± 0.1139	0.1676 ± 0.1183	0.1720 ± 0.1187	0.7921 ± 0.0417	0.2444 ± 0.1093
CELM-SCC	**0.3383** ± 0.1936	**0.1345** ± 0.1860	**0.3843** ± 0.1876	**0.1651** ± 0.1960	**0.2699** ± 0.1867	**0.2732** ± 0.1865	**0.7949** ± 0.1783	**0.2889** ± 0.1866
CELM	**0.4823** ± 0.1920	**0.2831** ± 0.1952	**0.5565** ± 0.1683	**0.3458** ± 0.2036	**0.4687** ± 0.1857	**0.4702** ± 0.1852	**0.8697** ± 0.0901	**0.4734** ± 0.1941

D.2 Extended Report Generation Performance on S0002

Tables 9 and 10 present a comprehensive performance comparison of EEG report generation on the S0002 site under settings with and without patient clinical context (subset of S0002 samples without patient clinical context). Both proposed variants consistently outperform all baselines across all evaluation metrics. Among them, `CELM` achieves better performance compared to the memory-efficient `CELM-SCC` variant across all metrics.

Table 9: Report generation performance comparison on samples with clinical context on S0002. We provide results for a complete set of evaluation metrics under two input settings: (i) *Unimodal + Text Only Input*, where language models generate reports solely from clinical context text, and (ii) *Unimodal + Text + EEG Features Input*, where EEG-derived features are additionally provided. The table compares multiple strong baselines, including general-purpose and medical LLMs, against our CELM. The best results are highlighted in orange, and the best baseline per category is highlighted in blue.

Method	BLEU-1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSUM	BERTScore	METEOR
	Unimodal + Text Only Input							
Gemma-3-1b-it	0.1509 ± 0.0898	0.0158 ± 0.0121	0.2330 ± 0.0876	0.0302 ± 0.0210	0.1233 ± 0.0467	0.1259 ± 0.0495	0.7058 ± 0.2423	0.1319 ± 0.0565
Gemma-3n-e2b-it	0.0925 ± 0.0769	0.0115 ± 0.0230	0.2318 ± 0.0600	0.0274 ± 0.0346	0.1224 ± 0.0377	0.1254 ± 0.0408	0.7683 ± 0.0560	0.1181 ± 0.0472
Gemma-3-4b-it	0.1538 ± 0.0990	0.0163 ± 0.0180	0.2775 ± 0.0558	0.0354 ± 0.0335	0.1407 ± 0.0336	0.1440 ± 0.0369	0.7851 ± 0.0730	0.1506 ± 0.0514
Gemma-3n-e4b-it	0.0539 ± 0.0664	0.0083 ± 0.0243	0.1998 ± 0.0649	0.0281 ± 0.0358	0.1152 ± 0.0392	0.1175 ± 0.0407	0.7635 ± 0.0525	0.0993 ± 0.0490
Medgemma-4b-it	0.0865 ± 0.0855	0.0109 ± 0.0194	0.2280 ± 0.0734	0.0341 ± 0.0341	0.1315 ± 0.0433	0.1345 ± 0.0458	0.7576 ± 0.1375	0.1198 ± 0.0542
Llama-3.2-1b-instruct	0.0340 ± 0.0700	0.0036 ± 0.0073	0.0954 ± 0.0869	0.0082 ± 0.0135	0.0581 ± 0.0483	0.0596 ± 0.0509	0.5139 ± 0.3501	0.0494 ± 0.0504
Llama-3.2-3b-instruct	0.1651 ± 0.1110	0.0186 ± 0.0256	0.2346 ± 0.0877	0.0321 ± 0.0351	0.1221 ± 0.0478	0.1271 ± 0.0517	0.7261 ± 0.1919	0.1430 ± 0.0689
Llama-3.1-8b-instruct	0.1629 ± 0.0970	0.0190 ± 0.0237	0.2655 ± 0.0766	0.0424 ± 0.0356	0.1368 ± 0.0444	0.1404 ± 0.0476	0.7502 ± 0.1531	0.1480 ± 0.0581
Meta-llama-3-8b-instruct	0.1441 ± 0.0847	0.0176 ± 0.0232	0.2588 ± 0.0586	0.0403 ± 0.0330	0.1354 ± 0.0374	0.1390 ± 0.0411	0.7702 ± 0.0817	0.1416 ± 0.0481
Qwen3-1.7b	0.0886 ± 0.1095	0.0097 ± 0.0124	0.1189 ± 0.1296	0.0142 ± 0.0191	0.0633 ± 0.0686	0.0652 ± 0.0716	0.3650 ± 0.3893	0.0735 ± 0.0821
Qwen3-4b-instruct-2507	0.1912 ± 0.0807	0.0218 ± 0.0197	0.2690 ± 0.0448	0.0398 ± 0.0298	0.1317 ± 0.0295	0.1350 ± 0.0324	0.7867 ± 0.0178	0.1622 ± 0.0434
	Unimodal + Text + EEG Features Input							
Gemma-3-1b-it	0.0957 ± 0.0925	0.0109 ± 0.0203	0.1755 ± 0.1157	0.0239 ± 0.0316	0.0980 ± 0.0653	0.1007 ± 0.0684	0.5785 ± 0.3384	0.0954 ± 0.0692
Gemma-3n-e2b-it	0.0750 ± 0.0771	0.0093 ± 0.0226	0.2190 ± 0.0840	0.0350 ± 0.0350	0.1256 ± 0.0516	0.1281 ± 0.0539	0.7175 ± 0.2153	0.1082 ± 0.0546
Gemma-3-4b-it	0.2089 ± 0.0895	0.0222 ± 0.0186	0.2886 ± 0.0458	0.0446 ± 0.0331	0.1451 ± 0.0315	0.1491 ± 0.0356	0.7883 ± 0.0418	0.1738 ± 0.0476
Gemma-3n-e4b-it	0.0169 ± 0.0486	0.0018 ± 0.0052	0.0357 ± 0.0903	0.0049 ± 0.0138	0.0194 ± 0.0490	0.0207 ± 0.0526	0.1073 ± 0.2678	0.0180 ± 0.0462
Medgemma-4b-it	0.0888 ± 0.0910	0.0110 ± 0.0195	0.1829 ± 0.1128	0.0248 ± 0.0349	0.1008 ± 0.0641	0.1029 ± 0.0660	0.5979 ± 0.3260	0.1000 ± 0.0702
Llama-3.2-1b-instruct	0.0312 ± 0.0668	0.0035 ± 0.0073	0.1016 ± 0.0773	0.0090 ± 0.0142	0.0620 ± 0.0420	0.0634 ± 0.0438	0.5903 ± 0.3062	0.0517 ± 0.0448
Llama-3.2-3b-instruct	0.1779 ± 0.1101	0.0207 ± 0.0224	0.2309 ± 0.0903	0.0350 ± 0.0349	0.1217 ± 0.0484	0.1364 ± 0.0575	0.7377 ± 0.1557	0.1540 ± 0.0743
Llama-3.1-8b-instruct	0.1960 ± 0.0959	0.0227 ± 0.0251	0.2256 ± 0.0904	0.0365 ± 0.0364	0.1198 ± 0.0542	0.1236 ± 0.0574	0.6895 ± 0.2171	0.1509 ± 0.0682
Qwen3-1.7b	0.1890 ± 0.1048	0.0185 ± 0.0118	0.2049 ± 0.1021	0.0236 ± 0.0179	0.1077 ± 0.0530	0.1110 ± 0.0568	0.6345 ± 0.2944	0.1413 ± 0.0724
Qwen3-4b-instruct-2507	0.2483 ± 0.0425	0.0217 ± 0.0098	0.2638 ± 0.0341	0.0334 ± 0.0179	0.1283 ± 0.0180	0.1322 ± 0.0250	0.7704 ± 0.0125	0.2017 ± 0.0285
	Multimodal + Linear Projector + CBraMod							
Gemma-3-4b-it	0.2683 ± 0.0751	0.0253 ± 0.0224	0.3084 ± 0.0491	0.0491 ± 0.0328	0.1500 ± 0.0331	0.1539 ± 0.0374	0.7879 ± 0.0729	0.2002 ± 0.0482
Medgemma-4b-it	0.0533 ± 0.0795	0.0058 ± 0.0084	0.1491 ± 0.1349	0.0202 ± 0.0215	0.0863 ± 0.0772	0.0878 ± 0.0788	0.4492 ± 0.3899	0.0729 ± 0.0701
Llama-3.2-3B-instruct	0.1529 ± 0.1010	0.0169 ± 0.0217	0.2425 ± 0.0837	0.0344 ± 0.0333	0.1327 ± 0.0486	0.1360 ± 0.0511	0.7333 ± 0.1900	0.1398 ± 0.0610
Llama-3.1-8B-instruct	0.0169 ± 0.0492	0.0018 ± 0.0051	0.0385 ± 0.0823	0.0041 ± 0.0114	0.0231 ± 0.0480	0.0238 ± 0.0500	0.1533 ± 0.3011	0.0209 ± 0.0458
Qwen3-4B-instruct-2507	0.2775 ± 0.0485	0.0283 ± 0.0121	0.2998 ± 0.0425	0.0507 ± 0.0240	0.1429 ± 0.0239	0.1473 ± 0.0306	0.7852 ± 0.0153	0.2143 ± 0.0349
CELM-SCC	**0.3767** ± 0.1557	**0.1671** ± 0.1054	**0.4487** ± 0.1283	**0.2149** ± 0.1109	**0.3232** ± 0.1263	**0.3268** ± 0.1258	**0.8229** ± 0.1341	**0.3232**± 0.1261
CELM	**0.5695** ± 0.1702	**0.4145** ± 0.1639	**0.6408** ± 0.1494	**0.4805** ± 0.1603	**0.5757** ± 0.1624	**0.5772** ± 0.1613	**0.8755** ± 0.1521	**0.5597** ± 0.1728

Table 10: Zero-context report generation performance on S0002. The best results are highlighted in orange, and the best baseline is highlighted in blue.

Method	BLEU-1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSUM	BERTScore	METEOR
	Unimodal + Text + EEG Features Input							
Gemma-3-4b-it	0.0544 ± 0.0904	0.0062 ± 0.0111	0.0907 ± 0.1259	0.0132 ± 0.0213	0.0476 ± 0.0658	0.0491 ± 0.0685	0.2800 ± 0.3788	0.0528 ± 0.0755
Medgemma-4b-it	0.0423 ± 0.0627	0.0052 ± 0.0074	0.1256 ± 0.1004	0.0142 ± 0.0171	0.0744 ± 0.0580	0.0754 ± 0.0588	0.5090 ± 0.3670	0.0668 ± 0.0563
Llama-3.2-3b-instruct	0.1375 ± 0.1060	0.0136 ± 0.0103	0.2035 ± 0.0874	0.0285 ± 0.0199	0.1105 ± 0.0441	0.1142 ± 0.0475	0.7042 ± 0.2133	0.1245 ± 0.0633
Qwen3-4b-instruct-2507	0.2211 ± 0.0834	0.0185 ± 0.0082	0.2307 ± 0.0838	0.0274 ± 0.0166	0.1152 ± 0.0421	0.1178 ± 0.0451	0.6919 ± 0.2319	0.1719 ± 0.0635
	Multimodal + Linear Projector + CBraMod							
Gemma-3-4b-it	0.2457 ± 0.0855	0.0231 ± 0.0101	0.2942 ± 0.0621	0.0458 ± 0.0178	0.1440 ± 0.0303	0.1475 ± 0.0343	0.7767 ± 0.1330	0.1901 ± 0.0508
Medgemma-4b-it	0.0343 ± 0.0454	0.0041 ± 0.0055	0.1785 ± 0.1146	0.0186 ± 0.0170	0.1035 ± 0.0645	0.1060 ± 0.0672	0.5798 ± 0.3438	0.0769 ± 0.0529
Llama-3.2-3B-instruct	0.0762 ± 0.0935	0.0081 ± 0.0099	0.1339 ± 0.1223	0.0164 ± 0.0197	0.0749 ± 0.0677	0.0769 ± 0.0702	0.4451 ± 0.3843	0.0748 ± 0.0708
Qwen3-4B-instruct-2507	0.2635 ± 0.0399	0.0278 ± 0.0081	0.2918 ± 0.0311	0.0479 ± 0.0168	0.1430 ± 0.0162	0.1468 ± 0.0238	0.7827 ± 0.0136	0.2100 ± 0.0332
CELM-SCC	**0.3097** ± **0.1511**	**0.1065** ± **0.0753**	**0.3864** ± **0.1199**	**0.1484** ± **0.0780**	**0.2499** ± **0.0957**	**0.2530** ± **0.0962**	**0.7940** ± **0.1658**	**0.2601** ± **0.1040**
CELM	**0.4976** ± 0.2062	**0.3313** ± 0.1761	**0.5625** ± 0.2024	**0.3931** ± 0.1840	**0.4880** ± 0.2013	**0.4896** ± 0.2008	**0.8113** ± 0.2517	**0.4854** ± 0.2075

D.3 Additional Results on Alignment Module Ablation

Comprehensive results of the alignment-module ablation study on the S0002 dataset, covering settings with clinical context, without clinical context, and all cases, are summarized in Table 11. We report results on three groups of S0002 test recordings: **(1) EEG Only**, the recordings whose reports contain no clinical-context section, **(2) With patient history**, the recordings with clinical context provided, and **(3) All**, the full test set of recordings. Across most evaluation settings, the SCA projector achieves significantly stronger performance than other alignment variants, highlighting the importance of modeling dependencies among EEG epoch tokens before projecting them into the LLM embedding space.

Table 11: Alignment module ablation on S0002 dataset. The best results are highlighted in orange

Projector	BLEU-1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSUM	BERTScore	METEOR
				EEG Only				
Linear Projector	0.2644 ± 0.0471	0.0247 ± 0.0063	0.2928 ± 0.0314	0.0447 ± 0.0141	0.1392 ± 0.0161	0.1392 ± 0.0161	0.7828 ± 0.0152	0.2072 ± 0.0382
Perceiver Projector	0.2636 ± 0.0696	0.0249 ± 0.0098	0.2897 ± 0.0344	0.0404 ± 0.0146	0.1377 ± 0.0165	0.1377 ± 0.0165	0.7888 ± 0.0162	0.1968 ± 0.0385
SCC Projector	0.2991 ± 0.1276	0.0912 ± 0.0590	0.3793 ± 0.0892	0.1378 ± 0.0666	0.2441 ± 0.0771	0.2441 ± 0.0771	**0.8192** ± 0.0299	0.2574 ± 0.0891
SCA Projector	**0.4652** ± 0.1884	**0.2666** ± 0.1354	**0.5248** ± 0.1869	**0.3271** ± 0.1494	**0.4339** ± 0.1767	**0.4339** ± 0.1767	0.7990 ± 0.2432	**0.4390** ± 0.1869
				With Patient History				
Linear Projector	0.2775 ± 0.0485	0.0283 ± 0.0121	0.2998 ± 0.0425	0.0507 ± 0.0240	0.1429 ± 0.0239	0.1473 ± 0.0306	0.7852 ± 0.0153	0.2143 ± 0.0349
Perceiver Projector	0.2672 ± 0.0556	0.0273 ± 0.0192	0.2879 ± 0.0378	0.0461 ± 0.0300	0.1384 ± 0.0281	0.1422 ± 0.0321	0.7878 ± 0.0159	0.1955 ± 0.0393
SCC Projector	0.3767 ± 0.1557	0.1671 ± 0.1054	0.4487 ± 0.1283	0.2149 ± 0.1109	0.3232 ± 0.1263	0.3268 ± 0.1258	0.8229 ± 0.1341	0.3232 ± 0.1261
SCA Projector	**0.5695** ± 0.1702	**0.4145** ± 0.1639	**0.6408** ± 0.1494	**0.4805** ± 0.1603	**0.5757** ± 0.1624	**0.5772** ± 0.1613	**0.8755** ± 0.1521	**0.5597** ± 0.1728
				All				
Linear Projector	0.2762 ± 0.0485	0.0279 ± 0.0117	0.2991 ± 0.0416	0.0501 ± 0.0233	0.1426 ± 0.0232	0.1465 ± 0.0296	0.7849 ± 0.0153	0.2135 ± 0.0353
Perceiver Projector	0.2669 ± 0.0571	0.0271 ± 0.0184	0.2881 ± 0.0375	0.0455 ± 0.0288	0.1384 ± 0.0271	0.1417 ± 0.0309	0.7879 ± 0.0160	0.1956 ± 0.0392
SCC Projector	0.3689 ± 0.1548	0.1594 ± 0.1042	0.4417 ± 0.1267	0.2071 ± 0.1098	0.3152 ± 0.1246	0.3185 ± 0.1243	0.8226 ± 0.1275	0.3166 ± 0.1245
SCA Projector	**0.5590** ± 0.1749	**0.3996** ± 0.1673	**0.6292** ± 0.1574	**0.4650** ± 0.1657	**0.5614** ± 0.1692	**0.5628** ± 0.1685	**0.8678** ± 0.1650	**0.5475** ± 0.1779

D.4 EEG Encoder Ablation

We conducted an ablation study to analyze report generation performance across different EEG encoders. Specifically, we compared CBraMod [26] and LaBraM [21], both EEG foundation models pretrained on large-scale EEG datasets. The results show that CBraMod consistently outperforms LaBraM, highlighting the critical role of high-quality EEG representations in effective clinical report generation.

Table 12: EEG encoder ablation study

EEG Encoder	BLEU-1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSUM	BERTScore	METEOR
Labram	0.4654 ± 0.2272	0.3188 ± 0.1904	0.5355 ± 0.2312	0.3809 ± 0.2027	0.4673 ± 0.2233	0.4689 ± 0.2231	0.7781 ± 0.2931	0.4600 ± 0.2259
Cbramod	**0.5590** ± 0.1749	**0.3996** ± 0.1673	**0.6292** ± 0.1574	**0.4650** ± 0.1657	**0.5614** ± 0.1692	**0.5628** ± 0.1685	**0.8678** ± 0.1650	**0.5475** ± 0.1779

D.5 Performance Analysis by Report Section

Clinical EEG reports comprise multiple sections, each serving a distinct purpose: EEG description/details provides a detailed narrative of observed waveforms and patterns, impression/interpretation summarizes the clinical significance of findings, background activity characterizes baseline rhythms, and events/seizures documents seizure episodes. To evaluate whether `CELM` can reliably generate these diverse sections, we conduct a section-wise analysis on S0001, which exhibits richer section diversity compared to S0002 (where reports predominantly contain only EEG description/details; see Figure 7 in Appendix). Figure 11 presents ROUGE-1 scores, comparing `CELM` against the best-performing unimodal baseline (text + EEG features) from each LLM family. `CELM` achieves the highest performance in 6 out of 7 report sections. Performance degrades in the interictal epileptiform abnormalities section, highlighting a key limitation and a challenge for ELMs in modeling rare and clinically complex events, an important direction for future work. Detailed results are provided in Table 13, which reports the complete section-wise performance across all models and evaluation metrics. These results further demonstrate the consistency of CELM across diverse components of clinical reports.

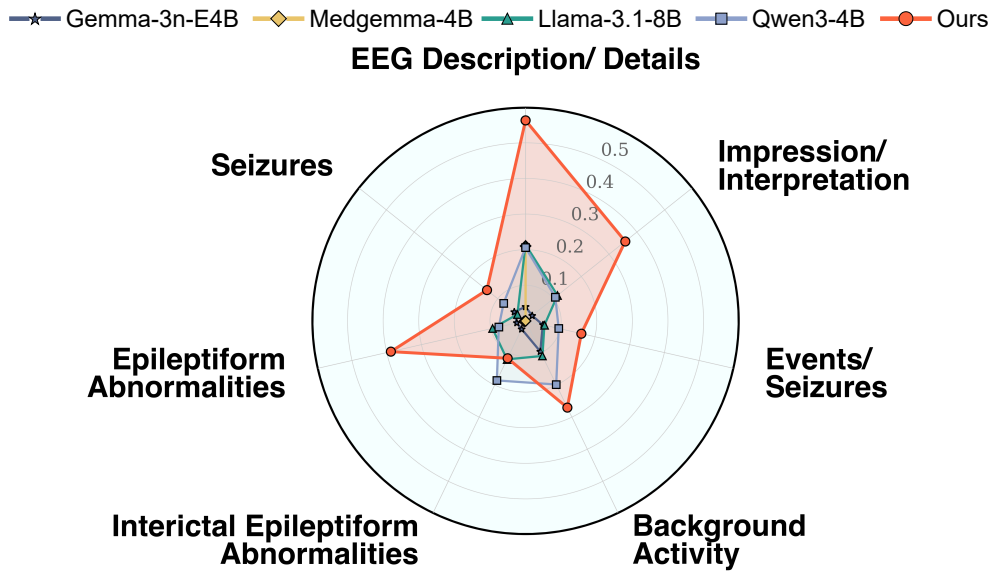

Figure 11: Section-wise comparison of report generation performance between CELM and the best-performing baselines from different LLM families.

Table 13: Performance by report section on S0001 dataset

Report Section	Method	BLEU-1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSUM	METEOR
EEG Description/Details	gemma-3n-E4B-it	0.0095 ± 0.0344	0.0017 ± 0.0057	0.0396 ± 0.0604	0.0012 ± 0.0057	0.0256 ± 0.0389	0.0259 ± 0.0395	0.0228 ± 0.0385
	medgemma-4b-it	0.1624 ± 0.1729	0.0489 ± 0.1401	0.2108 ± 0.1917	0.0612 ± 0.1872	0.1460 ± 0.1858	0.1487 ± 0.1861	0.1535 ± 0.1831
	Llama-3.1-8B-Instruct	0.2045 ± 0.1452	0.0501 ± 0.1179	0.2143 ± 0.1691	0.0603 ± 0.1694	0.1454 ± 0.1687	0.1485 ± 0.1690	0.1838 ± 0.1645
	Qwen3-4B-Instruct-2507	0.1818 ± 0.0988	0.0285 ± 0.0555	0.2059 ± 0.1190	0.0393 ± 0.0969	0.1271 ± 0.1082	0.1300 ± 0.1087	0.2069 ± 0.1080
	CELM-SCC	0.3381 ± 0.1900	0.1291 ± 0.1832	0.3826 ± 0.1854	0.1601 ± 0.1935	0.2650 ± 0.1841	0.2687 ± 0.1840	0.2875 ± 0.1833
	CELM	0.4907 ± 0.1834	0.2841 ± 0.1911	0.5631 ± 0.1607	0.3483 ± 0.1987	0.4727 ± 0.1813	0.4745 ± 0.1809	0.4798 ± 0.1873
Epileptiform Abnormalities	gemma-3n-E4B-it	0.0212 ± 0.0508	0.0081 ± 0.0201	0.0256 ± 0.0642	0.0048 ± 0.0181	0.0212 ± 0.0529	0.0217 ± 0.0539	0.0193 ± 0.0563
	medgemma-4b-it	0.0015 ± 0.0117	0.0008 ± 0.0064	0.0043 ± 0.0330	0.0034 ± 0.0265	0.0043 ± 0.0330	0.0043 ± 0.0330	0.0030 ± 0.0234
	Llama-3.1-8B-Instruct	0.0786 ± 0.1185	0.0343 ± 0.0596	0.0954 ± 0.1437	0.0422 ± 0.0860	0.0858 ± 0.1337	0.0863 ± 0.1347	0.0992 ± 0.1523
	Qwen3-4B-Instruct-2507	0.0574 ± 0.0919	0.0150 ± 0.0240	0.0772 ± 0.1201	0.0273 ± 0.0473	0.0568 ± 0.0808	0.0578 ± 0.0830	0.0759 ± 0.0942
	CELM-SCC	0.2870 ± 0.3051	0.1862 ± 0.2173	0.3392 ± 0.2614	0.2104 ± 0.2319	0.3233 ± 0.2597	0.3221 ± 0.2598	0.2717 ± 0.2560
	CELM	0.3172 ± 0.3472	0.2061 ± 0.2641	0.3881 ± 0.2967	0.2401 ± 0.2796	0.3600 ± 0.2902	0.3607 ± 0.2897	0.3222 ± 0.3107
Background Activity	gemma-3n-E4B-it	0.0440 ± 0.0909	0.0058 ± 0.0113	0.0954 ± 0.1133	0.0071 ± 0.0175	0.0584 ± 0.0679	0.0584 ± 0.0679	0.0534 ± 0.0685
	medgemma-4b-it	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
	Llama-3.1-8B-Instruct	0.0318 ± 0.0706	0.0047 ± 0.0102	0.1093 ± 0.1004	0.0102 ± 0.0181	0.0759 ± 0.0690	0.0759 ± 0.0690	0.0443 ± 0.0433
	Qwen3-4B-Instruct-2507	0.2172 ± 0.1394	0.0302 ± 0.0202	0.1984 ± 0.1289	0.0396 ± 0.0306	0.1158 ± 0.0752	0.1158 ± 0.0752	0.1492 ± 0.0969
	CELM-SCC	0.1494 ± 0.1142	0.0266 ± 0.0281	0.2442 ± 0.0629	0.0441 ± 0.0252	0.1670 ± 0.0521	0.1670 ± 0.0521	0.1355 ± 0.0524
	CELM	0.1825 ± 0.1225	0.0332 ± 0.0244	0.2703 ± 0.0536	0.0701 ± 0.0476	0.1935 ± 0.0530	0.1935 ± 0.0530	0.1786 ± 0.0655
Interictal Epileptiform Abnormalities	gemma-3n-E4B-it	0.0207 ± 0.0414	0.0096 ± 0.0192	0.0255 ± 0.0318	0.0075 ± 0.0150	0.0255 ± 0.0318	0.0255 ± 0.0318	0.0159 ± 0.0182
	medgemma-4b-it	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
	Llama-3.1-8B-Instruct	0.0891 ± 0.0899	0.0125 ± 0.0126	0.1200 ± 0.1008	0.0228 ± 0.0221	0.0672 ± 0.0550	0.0708 ± 0.0581	0.0863 ± 0.0781
	Qwen3-4B-Instruct-2507	0.1594 ± 0.0755	0.0251 ± 0.0092	0.1856 ± 0.1088	0.0334 ± 0.0218	0.0985 ± 0.0502	0.1048 ± 0.0539	0.1253 ± 0.0472
	CELM-SCC	0.0100 ± 0.0133	0.0027 ± 0.0039	0.0739 ± 0.0444	0.0228 ± 0.0140	0.0635 ± 0.0361	0.0666 ± 0.0338	0.0457 ± 0.0217
	CELM	0.0765 ± 0.0815	0.0135 ± 0.0129	0.1166 ± 0.1240	0.0236 ± 0.0232	0.0787 ± 0.0779	0.0787 ± 0.0779	0.0650 ± 0.0559
Events/Seizures	gemma-3n-E4B-it	0.0316 ± 0.0632	0.0086 ± 0.0172	0.0497 ± 0.0696	0.0012 ± 0.0016	0.0286 ± 0.0341	0.0291 ± 0.0341	0.0364 ± 0.0610
	medgemma-4b-it	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
	Llama-3.1-8B-Instruct	0.0448 ± 0.0551	0.0225 ± 0.0302	0.0544 ± 0.0495	0.0124 ± 0.0211	0.0528 ± 0.0497	0.0528 ± 0.0497	0.0212 ± 0.0201
	Qwen3-4B-Instruct-2507	0.0511 ± 0.0465	0.0167 ± 0.0138	0.0955 ± 0.0374	0.0040 ± 0.0071	0.0749 ± 0.0315	0.0758 ± 0.0300	0.0677 ± 0.0648
	CELM-SCC	0.0481 ± 0.0674	0.0257 ± 0.0353	0.1617 ± 0.1375	0.0004 ± 0.0009	0.1571 ± 0.1414	0.1578 ± 0.1407	0.0290 ± 0.0309
	CELM	0.0110 ± 0.0220	0.0090 ± 0.0181	0.1606 ± 0.1248	0.0024 ± 0.0033	0.1512 ± 0.1322	0.1531 ± 0.1305	0.0288 ± 0.0119
Seizures	gemma-3n-E4B-it	0.0168 ± 0.0352	0.0032 ± 0.0070	0.0405 ± 0.0588	0.0005 ± 0.0013	0.0256 ± 0.0343	0.0300 ± 0.0409	0.0187 ± 0.0273
	medgemma-4b-it	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
	Llama-3.1-8B-Instruct	0.0273 ± 0.0489	0.0169 ± 0.0328	0.0294 ± 0.0432	0.0010 ± 0.0026	0.0279 ± 0.0427	0.0294 ± 0.0432	0.0286 ± 0.0542
	Qwen3-4B-Instruct-2507	0.0388 ± 0.0558	0.0121 ± 0.0180	0.0790 ± 0.1025	0.0085 ± 0.0149	0.0502 ± 0.0552	0.0544 ± 0.0606	0.0464 ± 0.0543
	CELM-SCC	0.0631 ± 0.1319	0.0551 ± 0.1180	0.0773 ± 0.1057	0.0000 ± 0.0000	0.0725 ± 0.1075	0.0735 ± 0.1072	0.0336 ± 0.0502
	CELM	0.0272 ± 0.0489	0.0253 ± 0.0471	0.1390 ± 0.2061	0.0005 ± 0.0012	0.1285 ± 0.2088	0.1329 ± 0.2081	0.0419 ± 0.0545
Impression/Interpretation	gemma-3n-E4B-it	0.0172 ± 0.0725	0.0108 ± 0.0622	0.0237 ± 0.0949	0.0125 ± 0.0876	0.0207 ± 0.0923	0.0207 ± 0.0923	0.0193 ± 0.0990
	medgemma-4b-it	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
	Llama-3.1-8B-Instruct	0.0809 ± 0.1339	0.0294 ± 0.0774	0.1148 ± 0.1863	0.0489 ± 0.1334	0.0883 ± 0.1654	0.0895 ± 0.1663	0.0876 ± 0.1499
	Qwen3-4B-Instruct-2507	0.0885 ± 0.1173	0.0201 ± 0.0328	0.1069 ± 0.1399	0.0308 ± 0.0570	0.0675 ± 0.0910	0.0693 ± 0.0945	0.0996 ± 0.1172
	CELM-SCC	0.2194 ± 0.2899	0.1465 ± 0.2993	0.2687 ± 0.2946	0.1531 ± 0.3116	0.2190 ± 0.2929	0.2215 ± 0.2929	0.2206 ± 0.2996
	CELM	0.2493 ± 0.3481	0.2103 ± 0.3515	0.3581 ± 0.3381	0.2473 ± 0.3793	0.3358 ± 0.3434	0.3372 ± 0.3428	0.2829 ± 0.3428

D.6 Cross-Site Generalization

To assess whether CELM generalizes beyond the reporting conventions and patient distributions of the site it was trained on, we train CELM on S0002 and evaluate it on the S0001 test set without any fine-tuning on S0001. We compare it with the in-site model trained and tested on S0001 (Table 14). Lexical metrics decrease substantially in the cross-site setting (ROUGE-1: 0.5565 $\rightarrow$ 0.3585), whereas BERTScore decreases by only 7.8% (0.8697 $\rightarrow$ 0.8015). This gap between lexical and semantic metrics is consistent with the two sites differing in reporting style as S0001 includes more diverse visit types (routine EEG and epilepsy monitoring unit) and more multi-section report structures that the S0002-trained model has not seen. The cross-site model therefore retains much of the semantic content of the reports but does not reproduce S0001's reporting conventions.

Table 14: Cross-site generalization of CELM on S0001. We compare a model trained and tested on S0001 (in-site) with a model trained on S0002 and tested on S0001 (cross-site). We report mean $\pm$ standard deviation.

Setting	BLEU-1	ROUGE-1	METEOR	BERTScore
Train and test on S0001	**0.4823** $\pm$ 0.1920	**0.5565** $\pm$ 0.1683	**0.4734** $\pm$ 0.1941	**0.8697** $\pm$ 0.0901
Train on S0002, test on S0001	0.3411 $\pm$ 0.1646	0.3585 $\pm$ 0.1673	0.3215 $\pm$ 0.1712	0.8015 $\pm$ 0.1584

D.7 Score Distribution

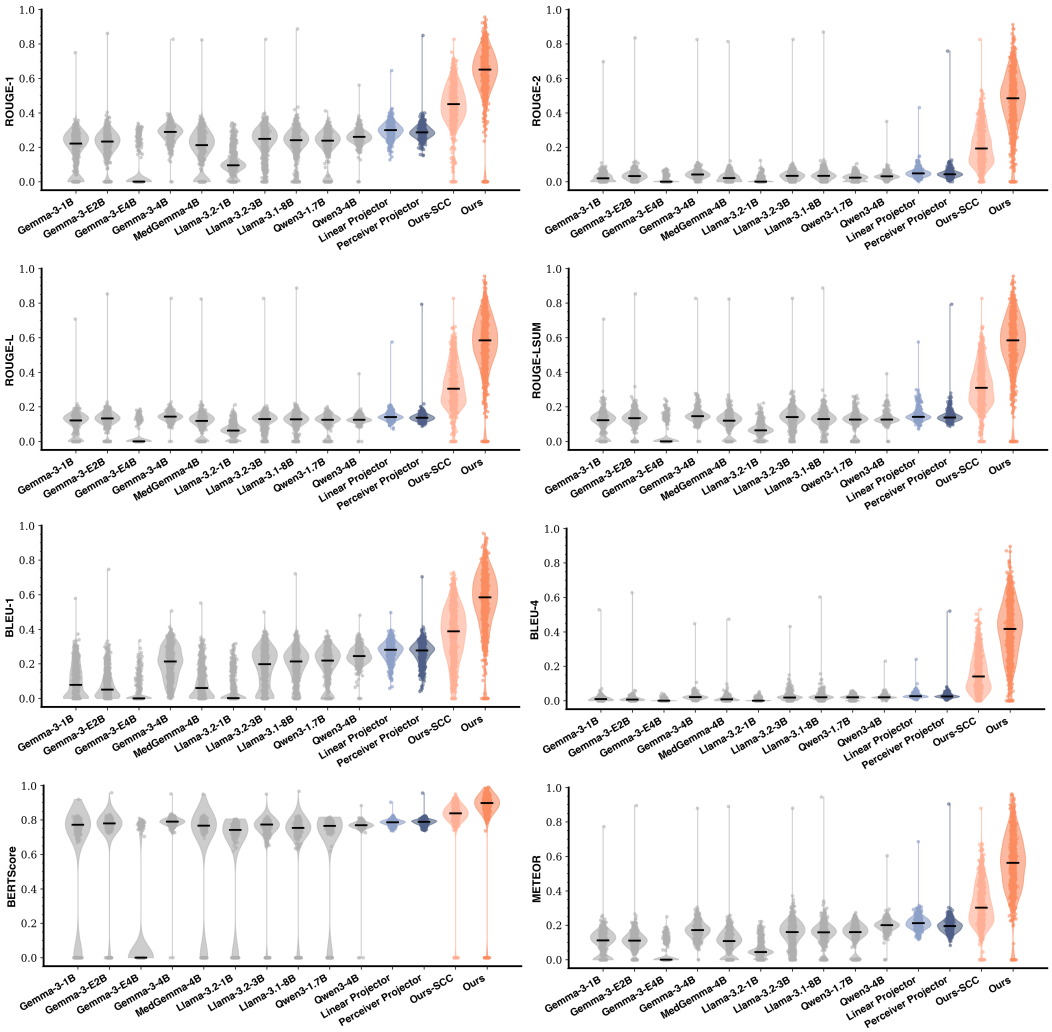

Figure 12: Distribution of various metrics for overall report generation in the S0002 dataset.

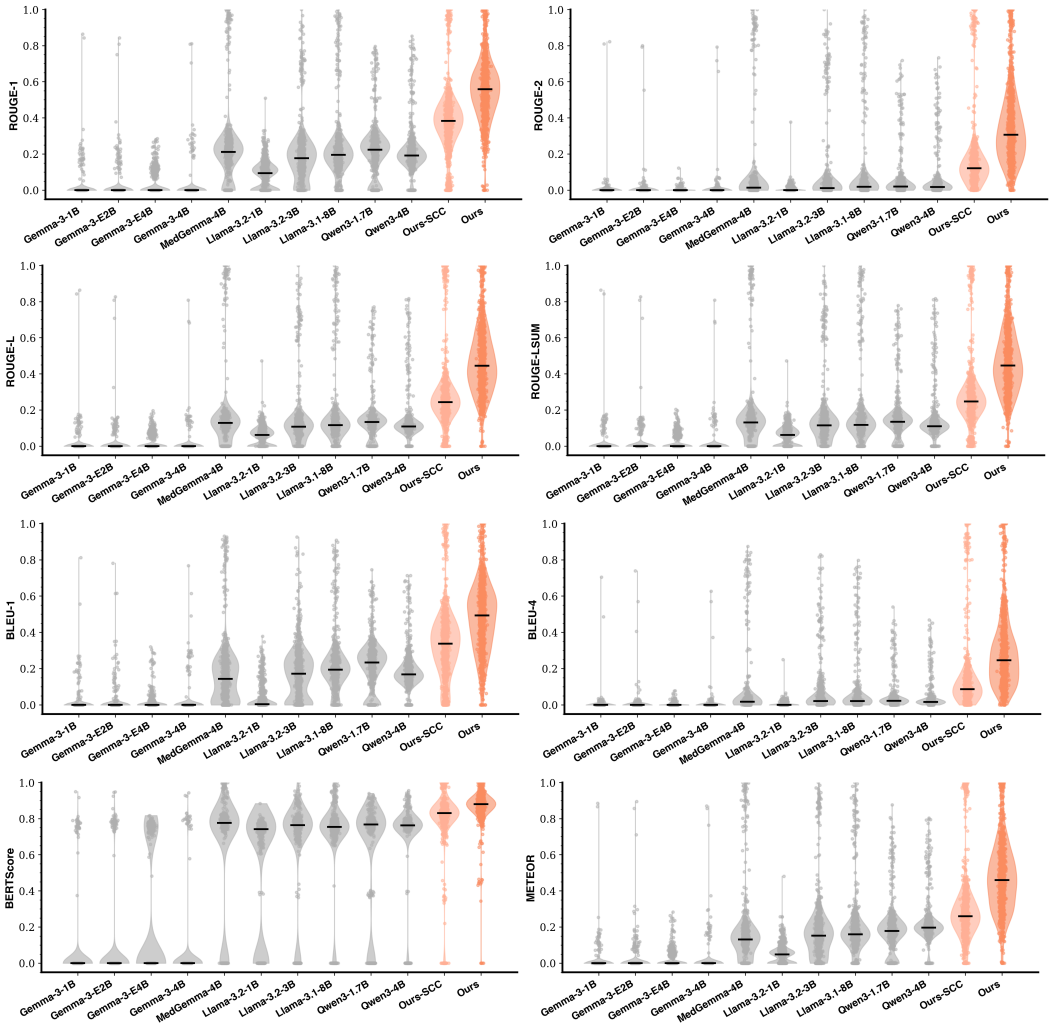

Figure 13: Distribution of various metrics for overall report generation in the S0001 dataset.

D.8 SCA vs. SCC by Recording Duration

To study how the performance of SCA and SCC projectors vary with recording length, we split each test set into duration quartiles and compare SCA and SCC within each quartile (Table 15). SCA outperforms SCC in every quartile on both sites (Wilcoxon signed-rank test, $p < 0.001$ within each quartile). On S0001, SCC degrades more with duration than SCA as ROUGE-1 falls from 0.4031 to 0.3479 from Q1 to Q4 (-14%), compared with 0.5522 to 0.5173 for SCA (-6%). This attributes to the fixed set of query tokens becoming an information bottleneck as recordings grow longer. On S0002, where recordings are comparatively shorter, both models improve with duration, and the SCA-SCC gap widens as SCA benefits more from longer inputs.

Table 15: Report generation performance of SCA and SCC by recording-duration quartile.

Quartile	Dur. (min)	Range (min)	Proj.	ROUGE-1	METEOR	BERTScore
			S0001			
Q1	54.4	0.7–65.2	SCA	**0.5522**	**0.4553**	**0.8709**
			SCC	0.4031	0.2895	0.8112
Q2	71.5	65.3–77.8	SCA	**0.5940**	**0.5089**	**0.8824**
			SCC	0.4041	0.3081	0.8013
Q3	86.4	78.0–94.2	SCA	**0.5618**	**0.4828**	**0.8758**
			SCC	0.3816	0.2899	0.7939
Q4	103.4	94.3–161.5	SCA	**0.5173**	**0.4461**	**0.8494**
			SCC	0.3479	0.2680	0.7729
			S0002			
Q1	37.1	24.5–46.3	SCA	**0.5521**	**0.4845**	0.8030
			SCC	0.3861	0.2401	**0.8048**
Q2	49.9	46.5–52.3	SCA	**0.6603**	**0.5635**	**0.8962**
			SCC	0.4645	0.3414	0.8232
Q3	55.4	52.5–58.2	SCA	**0.6761**	**0.5996**	**0.8991**
			SCC	0.4801	0.3676	0.8372
Q4	66.4	58.3–128.0	SCA	**0.6757**	**0.5917**	**0.9044**
			SCC	0.4646	0.3444	0.8266

D.9 Benchmark Construction Quality Validation

To assess the reliability of our benchmark construction pipeline, we conducted a systematic manual validation of section extraction and canonical normalization using 50 randomly sampled clinical EEG reports from both sites. **(1) Section extraction quality.** We manually compared 150 extracted sections against their corresponding original report text using a four-point scale: 3 (faithful extraction), 2 (minor boundary errors), 1 (major extraction errors), and 0 (failed or empty extraction). The mean score was 2.9 out of 3.0, demonstrating high extraction fidelity using our detect-then-extract pipeline with a local LLM. **(2) Canonical normalization accuracy.** We evaluated whether extracted EEG section headers were correctly mapped to their canonical categories. Across 80 evaluated sections spanning all seven canonical categories, the normalization procedure achieved 100% accuracy. These results support the reliability of the section extraction and normalization procedures used in constructing our benchmark.

D.10 Qualitative Analysis and Case Studies

Figure 14 presents qualitative comparisons on S0002 between the best-performing unimodal baseline (text with EEG features), the linear projector variant, and `CELM`, alongside ground truth reports. Figure 14a shows the EEG description/details section, while Figure 14b shows the impression/interpretation section. Across these examples, `CELM` produces reports that are more closely aligned with the ground truth, correctly identifying clinically relevant findings such as moderate bilateral beta activity, intermittent drowsiness, and bilateral slowing in the theta and delta ranges. In contrast,

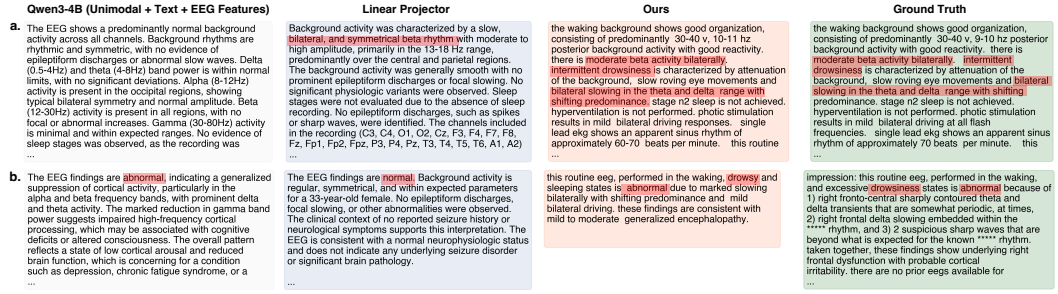

Figure 14: Examples of generated reports on S0002. Comparisons between the unimodal baseline (text + EEG features), the linear-projector alignment variant, CELM, and the ground-truth reports. (a) EEG description/details; (b) Impression/interpretation.

the linear projector and unimodal baselines fail to capture these. In the impression section, CELM identifies the recording as abnormal, whereas the linear projector variant incorrectly predicts it as normal. These results demonstrate the promise of ELMs in learning to generate clinical reports from unstructured notes paired with long EEG recordings. At the same time, they underscore the need for more rigorous and standardized benchmarking to assess these models. Additional examples spanning both sites, diverse report sections, and varying performance levels are provided below to further illustrate model strengths and failure modes. In Figure 15, 16, 17, and 18, we compare generated reports from the unimodal baseline, CELM-SCC, and CELM against ground-truth reports across multiple EEG report sections and datasets.

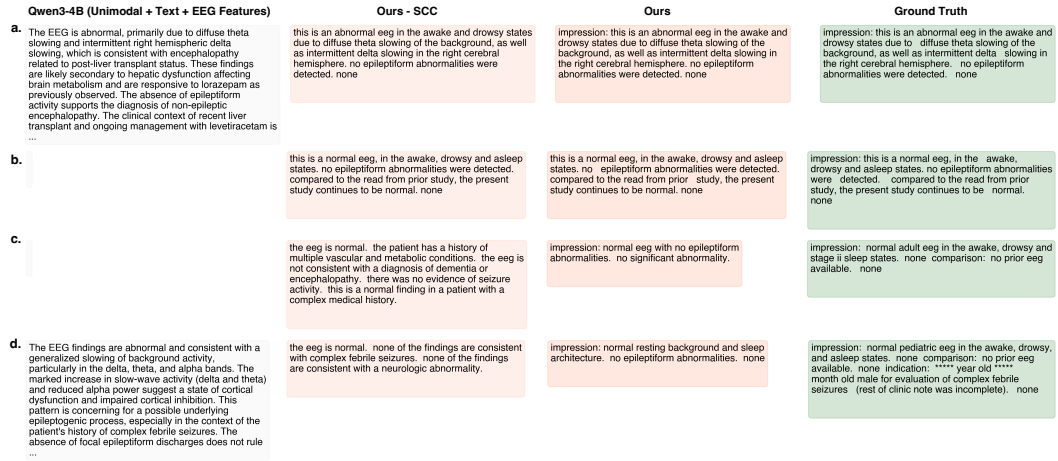

Figure 15: Impression/interpretation generation examples on S0001. Examples are ordered from (a) to (d) by decreasing ROUGE-1 score. We compare outputs from the unimodal baseline (text + EEG features), CELM-SCC, CELM, and the ground-truth reports.

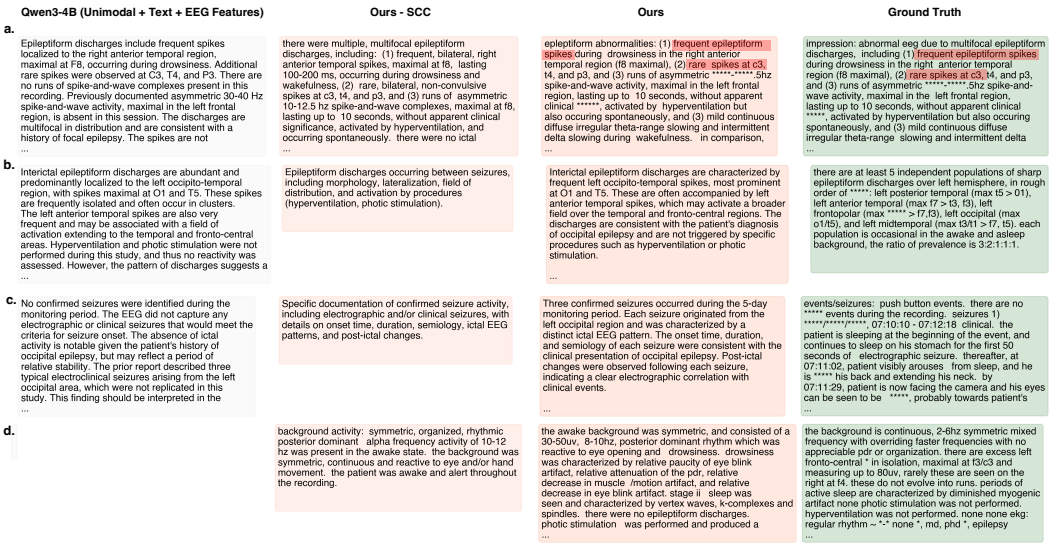

Figure 16: Section-wise generation examples on S0001. (a) Epileptiform abnormalities, (b) Interictal epileptiform abnormalities, (c) Seizures, and (d) Background activity. We compare outputs from the unimodal baseline (text + EEG features), CELM-SCC, CELM, and the ground-truth reports.

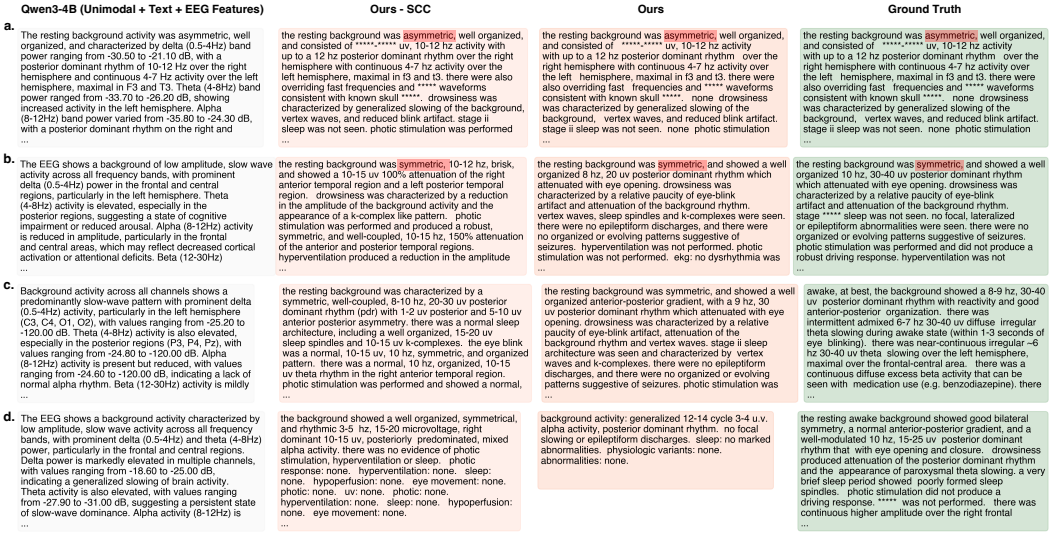

Figure 17: EEG description/details generation examples on S0001. Examples are ordered from (a) to (d) by decreasing ROUGE-1 score. We compare outputs from the unimodal baseline (text + EEG features), CELM-SCC, CELM, and the ground-truth reports.

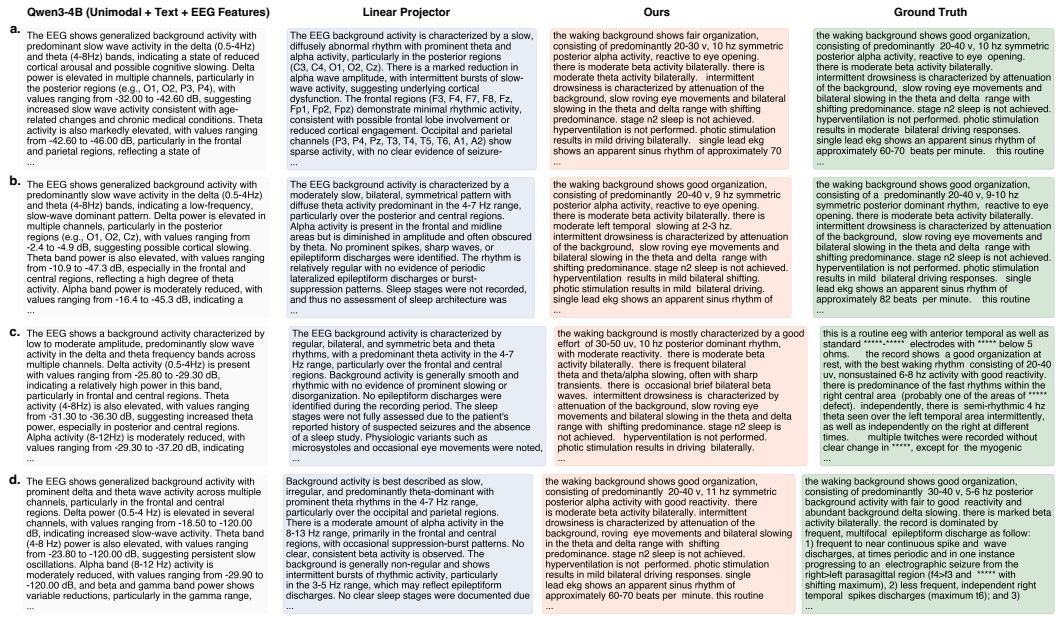

Figure 18: EEG description/details generation examples on S0002. Examples are ordered from (a) to (d) by decreasing ROUGE-1 score. We compare outputs from the unimodal baseline (text + EEG features), CELM-SCC, CELM, and the ground-truth reports.

E Human Expert Evaluation Protocol

We describe the evaluation protocol, rating criteria, statistical analysis, and ethical oversight for the human expert evaluation presented in Section 5.4.

Evaluators. Six experts participated in the evaluation, comprising clinicians and BME researchers with expertise in clinical EEG interpretation, signal analysis, and neurophysiology. Evaluators were recruited from the authors' academic and clinical research networks and participated voluntarily without compensation. All evaluators completed the BDSP access requirements for the Harvard EEG Database, including credentialed access, CITI training, and a Data Use Agreement.

Sample Selection and Evaluation Procedure. We randomly selected 20 cases from the held-out test sets across two sites. For each case, evaluators reviewed the reference clinical report alongside three generated reports: CELM, the best *Unimodal + Text + EEG Features Input* baseline (Qwen3-4b-instruct-2507), and the best *Multimodal + Linear Projector + CBraMod* baseline (Qwen3-4b-instruct-2507). This was a blinded study where model identities were concealed. All evaluators received standardized written instructions. For each case, the ground-truth clinical report was presented alongside three anonymized generated reports to minimize evaluation bias. Evaluators reviewed only de-identified report text, without access to raw EEG recordings or patient-identifiable information. Evaluators completed three tasks:

1. **Report rating:** Independently assess each generated report on the five dimensions defined in Table 16, using the specified five-point scales.
2. **Preference ranking:** Rank the three generated reports according to overall preference.
3. **Optional feedback:** Provide qualitative comments or suggestions for improving the generated reports.

Each evaluator assessed all 20 cases, yielding 60 individual report assessments per evaluator and 360 ratings per dimension across the six evaluators.

Rating Dimensions and Scale. Evaluators assessed each generated report against the corresponding ground-truth clinical report across five dimensions: clinical correctness, findings coverage, finding accuracy, real-world readiness, and completeness. Each dimension was rated on a five-point Likert-type scale, with dimension-specific anchors provided in Table 16. Evaluators additionally ranked the three generated reports by overall preference and could optionally provide qualitative feedback.

Table 16: Expert evaluation dimensions and rating criteria. All assessments were made with reference to the ground-truth clinical report.

Dimension	Evaluation criterion and rating scale
Clinical Correctness	Correctness of clinical content relative to the ground-truth report. 5: Very good; 4: Good; 3: Acceptable; 2: Poor; 1: Very poor.
Findings Coverage	Extent to which key clinical findings in the ground-truth report are identified. 5: All found; 4: Most found; 3: Some found; 2: Few found; 1: Missed.
Finding Accuracy	Accuracy of finding attributes, including lateralization, frequency, and morphology. 5: All correct; 4: Minor errors; 3: Some errors; 2: Many errors; 1: Inaccurate.
Real-World Readiness	Suitability of the generated report for real-world clinical use. 5: Very good; 4: Good; 3: Acceptable; 2: Poor; 1: Very poor.
Completeness	Extent to which the generated report covers important information. 5: Very good; 4: Good; 3: Acceptable; 2: Poor; 1: Very poor.

F Model Card

We document CELM following the model-card framework from [41].

Model Details	
Developers	Jathurshan Pradeepkumar, Zheng Chen, and Jimeng Sun (University of Illinois Urbana-Champaign; SANKEN, Osaka University).
Model date	April 2026.
Model version	v1.0. Two variants: CELM (Sequence Context Alignment, SCA) and CELM-SCC (Sequence Context Compression). Separate models for S0001 and S0002
Model type	EEG-to-language model for structured clinical report generation. A frozen pretrained EEG encoder (CBraMod) produces one token per 10-s epoch (Epoch-Aggregated Tokenization). A trainable sequence-aware alignment module maps these tokens into the embedding space of a frozen instruction-tuned LLM (Qwen3-4B-Instruct-2507). In SCA this module is a linear-attention transformer followed by a linear projection and in SCC it is Perceiver-style cross-attention with learnable queries. The LLM generates report text conditioned on the EEG tokens, a task prompt, and optional clinical context.
Inputs and outputs	Input: scalp EEG resampled to 200 Hz, 22-channel 10-20 montage in a fixed channel order, up to $\sim$3 hours long. Optional clinical context (reason for study, clinical history). Output: free-text report sections (EEG description/details, impression/interpretation, background activity, epileptiform abnormalities, events/-seizures).
Training	Supervised fine-tuning with a next-token prediction objective. Only the alignment module is updated and the EEG encoder and LLM stay frozen. Hyperparameters and compute are in Appendix C. No fairness constraints were applied during training.
Paper	This paper. Code: `https://github.com/Jathurshan0330/CELM`.
License	Code: MIT License. Model checkpoints are not publicly released and access will be gated and subject to the BDSP Credentialed Health Data License 1.5.0.
Contact	`jp65@illinois.edu`, `chenz@sanken.osaka-u.ac.jp`, or GitHub issues.
Intended Use	
Primary intended uses	Research on EEG–language modeling and on automated drafting of clinical EEG reports, including benchmarking new methods on the EEG–Report benchmark.
Primary intended users	Machine learning and clinical neurophysiology researchers with credentialed access to the Harvard EEG Database through BDSP.
Out-of-scope uses	Any clinical deployment, autonomous clinical reporting, diagnostic or treatment decisions, triage or prioritization of patient studies. use on patient data without review by a qualified neurologist, and any use that violates the BDSP data use agreement, which restricts use to research.
Metrics	
Performance measures	Lexical: BLEU-1/4, ROUGE-1/2/L/Lsum, METEOR. Semantic: BERTScore. LLM-as-judge (Qwen2.5-72B-Instruct-AWQ, run locally): clinical correctness, findings coverage, finding accuracy. Expert ratings (1–5) on five clinical dimensions, plus preference rankings. Manual accuracy on abnormality prediction, lateralization, spectral dynamics, and epileptiform abnormalities. Lexical metrics are reported for comparability but do not measure clinical correctness.
Decision thresholds	Not applicable: CELM generates text and has no classification threshold. Its normal/abnormal judgment appears only in the generated text.
Variation approaches	Mean $\pm$ standard deviation across test samples; paired Wilcoxon signed-rank tests against baselines and perturbation controls; inter-rater ICC(2,k), adjacent agreement, and Kendall's W for expert ratings.
Evaluation Data	
Datasets	Held-out test splits (20%, patient-wise) of the EEG–Report benchmark built from the Harvard EEG Database v4.1 and the linked EHR repository on BDSP: S0001 (Massachusetts General Hospital) and S0002 (Brigham and Women's Hospital).
Motivation	The largest available source of paired clinical EEG and reports. The two sites differ in visit types and report structure, which allows a within- and cross-site evaluation.

Preprocessing	EEG: 0.1–75 Hz band-pass, 60 Hz notch, resampling to 200 Hz, 10-s epochs, 22-channel 10–20 montage; single-session reports only; S0001 recordings over 10,000 s excluded. Reports: automated structuring (section detection, extraction with a local Llama-3-8B-Instruct, canonical normalization; Section 3). A manual audit gave a mean extraction score of 2.9/3 over 150 sections and 100% correct normalization.
Training Data	
	Training (60%) and validation (20%) patient-wise splits of the same two sites, with the same preprocessing. In total, 9,922 reports, $\sim$11,000 hours of EEG, and 9,048 patients (5,049 reports from 4,669 patients at S0001; 4,873 from 4,379 at S0002). Duration, age, gender, and report-section distributions are in Table 4 and Figure 7. No clinical data or derived text are redistributed.
Ethical Considerations	
	The training data are sensitive clinical records. They were de-identified by the data providers, are accessed under a credentialed data use agreement, and were processed only with locally hosted models. Because model weights could in principle memorize report text, checkpoints are not publicly released. Generated reports can omit or invent findings, used without oversight, they could cause misdiagnosis or encourage automation bias.
Caveats and Recommendations	
	(1) CELM often classifies mildly abnormal recordings as normal, and it is weaker on interictal epileptiform abnormalities and other rare findings. Treat a "normal" impression with caution. (2) CELM is sensitive to channel order and inputs must follow the same 22-channel montage and ordering used in training. (3) Recordings longer than $\sim$3 hours are not supported, CELM-SCC handles longer inputs at lower quality. (4) Intended only for research purposes

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims and contributions mentioned in the abstract and introduction accurately reflect the paper's contribution and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations and implications of our work are discussed under Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All the necessary details to reproduce our work are provided in the main text, along with extended details in the Appendix. Additionally, we provide our code for the proposed model and benchmark construction pipeline at `https://github.com/Jathurshan0330/CELM`. The data utilized for this work is publicly available at `https://bdsp.io/content/harvard-eeg-db/4.1/`, which can be accessed after completing the necessary steps to obtain credentialed access.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide our code for the proposed model and benchmark construction pipeline at `https://github.com/Jathurshan0330/CELM` for reproducibility. The data utilized for this work is publicly available at `https://bdsp.io/content/harvard-eeg-db/4.1/`, which can be accessed after completing the necessary steps to obtain credentialed access.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Provided in the main text and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report standard deviations and error bars where relevant, and assess statistical significance with paired Wilcoxon signed-rank tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Provided in Appendix

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, it follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Included in Section 6

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We release only code; no clinical data or derived reports are redistributed. Model checkpoints are not publicly released, and access will be gated under the BDSP Credentialed Health Data License 1.5.0 in the future. We provide a model card (Appendix F) documenting intended use, out-of-scope uses, and known failure modes, and we state that `CELM` is for research use only (Section 6). All data processing used locally hosted LLMs, as required by the data use agreement.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators of the dataset used in this study are cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release the benchmark construction code and the model code at `https://github.com/Jathurshan0330/CELM` under the MIT License, with documentation and a model card (Appendix F).

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The paper does not include crowdsourcing or human-subjects research on participants. It does include an expert evaluation of model outputs (Section 5.4). Appendix E gives the full task instructions, survey design, evaluator background and recruitment. Participation was voluntary and uncompensated.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: No new data were collected from patients. We perform a retrospective analysis and modeling of a de-identified, credentialed-access database under its data use agreement (Ethics Statement).

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The proposed method utilizes local LLMs in the proposed framework. Additionally, LLMs were used to refine and correct grammar in the writing.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.