NEURAL SIGNALS GENERATE CLINICAL NOTES IN THE WILD

Jathurshan Pradeepkumar[1] **Xihao Piao**[2], **Zheng Chen**[2] **Jimeng Sun**[1]
[1]University of Illinois Urbana-Champaign [2]SANKEN, Osaka University
{jp65,jimeng}@illinois.edu,{park88,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 curate a large-scale clinical EEG dataset with $9{,}922$ reports paired with approximately $11{,}000$ hours of EEG recordings from $9{,}048$ patients. We therefore develop 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, including recording description, background activity, epileptiform abnormalities, events/seizures, and impressions. Experimental results show that, with patient history supervision, our method achieves 70%–95% average relative improvements in standard generation metrics (e.g., ROUGE-1 and METEOR) from 0.2–0.3 to 0.4–0.6. In the zero-shot setting without patient history, CELM attains generation scores in the range of 0.43–0.52, compared to baselines of 0.17–0.26. CELM integrates pretrained EEG foundation models with language models to enable scalable multimodal learning. We release our model 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 in hospitals for neurological diagnosis, such as epilepsy (Kotoge et al., 2025) and sleep health (Yang et al., 2023). However, writing clinical reports to summarize diagnostic findings over long-term EEG recordings remains a labor-intensive process. Typically, neurologists visually inspect EEG signals to assess brain activity and then compose clinical reports that document abnormal patterns (i.e., EEG phenotypes) and their clinical interpretations (Tatum IV et al., 2016). Moreover, this task requires substantial domain knowledge and extensive clinical experience, which in turn incur long-term training costs and require continual updates as new EEG phenotypes or waveforms are identified (Biswal et al., 2019).

To support this profound clinical practice, several automated clinical EEG report generation models based on advanced deep learning techniques have been proposed (Biswal et al., 2019; Lee et al., 2023; Yin & Shin, 2025). These methods aim to encode EEG recordings into latent representations that capture temporal structure and clinically relevant phenotypes, which are then decoded into clinical reports. However, existing methods remain limited in several aspects. ① Conceptually, some works formulate report generation as a phenotype classification or retrieval problem followed by text decoding. They do not model EEG-to-report generation as an end-to-end learning task, and the misalignment between classification and generation objectives does not guarantee optimal report generation. ② Methodologically, these methods operate on short EEG segments and fixed-context templates. They fail to model long-term temporal context and global diagnostic reasoning required for interpreting long-duration EEG data. Moreover, fixed-context templates prevent multi-granularity report generation beyond predefined categories. ③ Practically, these methods are developed as task-specific models, each tailored to a narrow reporting objective (e.g., impression-only reports) (Lee et al., 2023). In contrast, 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. We introduce CELM , which, to the best of our knowledge, is the first clinical EEG-to-language foundation model (ELM) capable of generating EEG reports at multiple scales. We train and evaluate CELM on $9{,}922$ clinical reports paired with approximately $11{,}000$ hours of EEG recordings from $9{,}048$ patients, spanning both clinical context–conditioned and zero-context settings, and covering key report sections including EEG description, background activity, epileptiform abnormalities, events/seizures, and impressions/interpretations.

Our motivation stems from recent advances in multimodal foundation models that enable text generation from non-text modalities (e.g., images and speech) (Bordes et al., 2024; Shen et al., 2025). Recent studies have begun to incorporate EEG signals into language-based modeling frameworks (Ndir et al., 2025; Gijsen & Ritter, 2025). However, they primarily leverage text supervision to enhance EEG classification and do not support EEG-to-text report generation.

To tackle aforementioned limitations, the contributions of this paper lie in: ❶ Conceptually, we formulate neurological clinical report generation as an *EEG-to-Language foundation modeling* problem. CELM enables end-to-end training and generation by directly translating raw EEG recordings into clinical text, without relying on intermediate phenotype classification or template-based pipelines. ❷ Methodologically, we introduce the first EEG-to-language multimodal framework, comprising 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 **Practically,** we introduce the first ELM benchmark construction pipeline based on the Harvard Electroencephalography Database (Sun et al., 2025; Zafar et al., 2025), which contains approximately 12k clinical reports paired with about 100k hours of EEG recordings from $10k$ patients. Rather than requiring access to the raw, unstructured 200TB corpus and associated metadata, our pipeline enables efficient filtering of EEG recordings matched to reports. We release all source code to support further evaluation and development.

Experimental results show that CELM consistently outperforms all baselines across all settings. With patient history, it achieves 70%–95% average relative improvements, with standard generation metrics (e.g., METEOR (Banerjee & Lavie, 2005) and ROUGE-1 (Lin, 2004)) from 0.2–0.3 to 0.4–0.6 across diverse cases. In the zero-context setting without patient history, CELM achieves generation scores in the range of 0.43–0.52, compared to baselines of 0.17–0.26. We provide multi-perspective case analyses to shed light on future ELM development.

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 (Herff et al., 2015; Wang & Ji, 2022). This line of work spans both invasive approaches using electrocorticography (ECoG) (Makin et al., 2020; Willett et al., 2023) and non-invasive methods based on scalp EEG (Wang & Ji, 2022; Duan et al., 2023; Zhou et al., 2024). However, these methods assume precise EEG–text alignment, whereas clinical EEG consists of heterogeneous events embedded in long, continuous recordings spanning hours to days (Sun et al., 2025; Zafar et al., 2025). 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 (Radford et al., 2021), a recent work (Ndir et al., 2025) 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 (Biswal et al., 2019), 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 (Lee et al., 2023), they rely on fixed segmentation and template-based generation (Yin & Shin, 2025). As a result, these methods fall short of end-to-end clinical report generation from long-duration

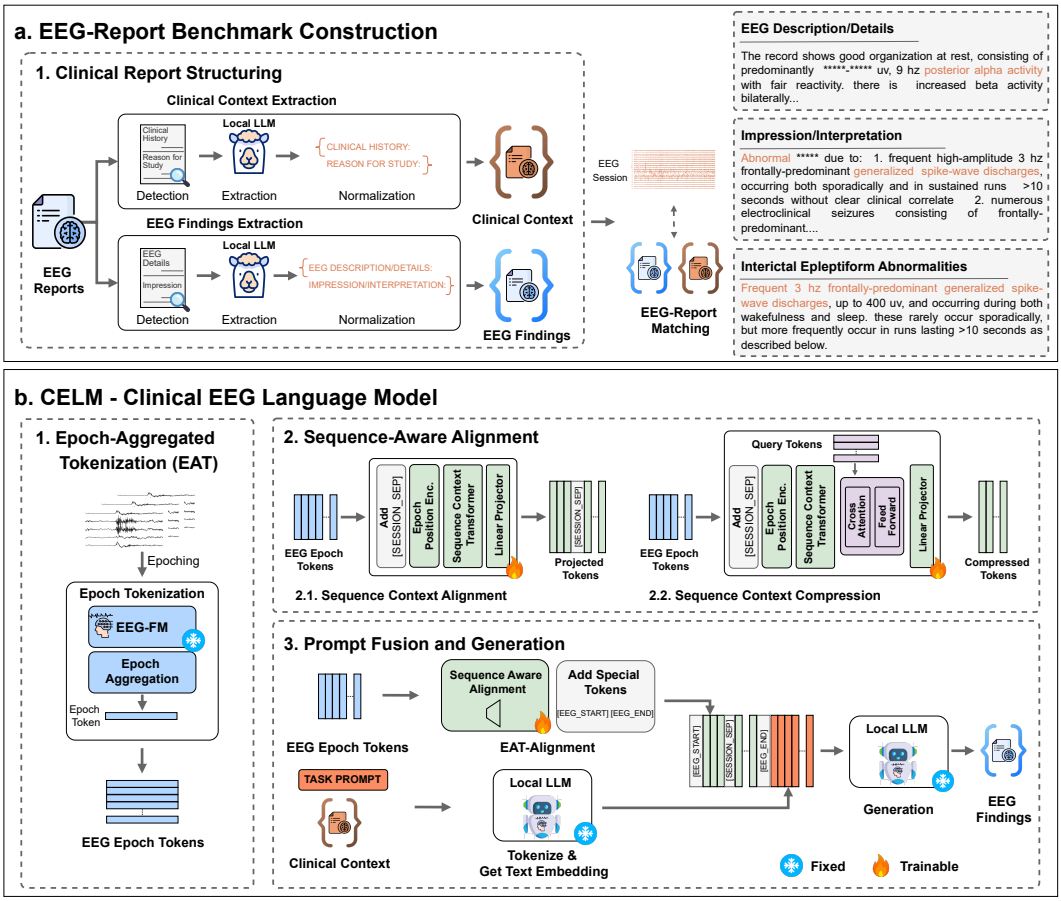

Figure 1: Overview of CELM. (a) EEG–Report benchmark construction pipeline, including clinical report structuring (Section 3.1), matching reports to EEG sessions, and examples of standardized report sections. (b) The proposed Clinical EEG Language Model (CELM) comprises (b.1) Epoch-Aggregated Tokenization, (b.2) Sequence-Aware Alignment, and (b.3) Prompt Fusion and Generation.

EEG recordings. We address the above challenges by introducing the first family of EEG–Language foundation models.

3 EEG-Report Benchmark Construction

Currently, no publicly available benchmark provides structured EEG–report pairs for training and evaluating EEG-to-Language Models (ELMs). In this work, we leverage two resources from the *Brain Data Science Platform*: (1) the Harvard Electroencephalography Database v4.1 (Zafar et al., 2025; Sun et al., 2025), which contains approximately $10k$ 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 1a). Specifically, Section 3.1 describes how neurology reports are extracted and organized based on whether their content is inferable from EEG data. Section 3.2 details the preprocessing and standardization of the matched EEG recordings.

3.1 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 observations and clinical interpretations that are directly inferable from EEG recordings. This distinction is critical for evaluating clinical ELMs, as sections within EEG Findings provide the appropriate ground truth for benchmarking a model's ability to generate clinical narratives from EEG signals.

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:* first, we curate a comprehensive list of candidate section headers under each category and detect their presence using string matching.
➢ *Section-wise Extraction:* second, 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. Local deployment is required, as data use agreements prohibit sending clinical reports to proprietary cloud-based LLMs.
➢ *Canonical Normalization:* finally, we standardize the extracted sections into canonical categories, including EEG description/details, impressions/interpretations, background activity, epileptiform abnormalities, and events/seizures.

3.2 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 recordings are 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. We then perform patient-level splitting into 60/20/20 training, validation, and test sets to ensure no patient overlap across splits, thereby preventing data leakage and producing a structured EEG–report benchmark.

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*. Unlike classification-oriented EEG models that learn discriminative representations for categorical labels, clinical ELMs must identify and aggregate clinically relevant information from 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 (Ndir et al., 2025). 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 $\sim 31.7\mathrm{M}+$ data points per recording, far exceeding the context limits of modern LLMs (Grattafiori et al., 2024; Team et al., 2025; Yang et al., 2025).
➢ *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 extreme temporal scale. These tokens are then projected into the language model 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 $\approx$ 158K tokens per session, still well beyond the context limits of most local LLMs (Grattafiori et al., 2024; Team et al., 2025; Yang et al., 2025). To overcome this limitation while harnessing pretrained EEG representations, we propose Epoch-Aggregated Tokenization (see Figure 1b.1).

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 mini 1s windows within each epoch using a pretrained EEG encoder. The resulting representations are then aggregated across mini-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 (Wang et al., 2024b), whereas LaBraM (Jiang et al., 2024) and TFM-Tokenizer (Pradeepkumar et al., 2025) uses `[CLS]` pooling. This tokenization compresses the tokens by up to $C \times T$ ($\sim 220\times$ compared to mini-window tokenization), enabling EEG recordings to fit within LLM context limits.

4.3 SEQUENCE-AWARE ALIGNMENT

To enable EEG-conditioned clinical report generation and leverage LLMs' generative capabilities, EEG tokens must be aligned with the text embedding space of LLMs. In the vision-language setting, models such as LLaVA achieve this via a simple linear projection (Liu et al., 2023). However, this paradigm does not generalize well to EEG (Section 5.3) due to its inherent sequential nature, unlike static images. EEG signals exhibit long-range temporal dynamics across epochs that must be preserved during alignment. As a result, we face two key challenges: modeling long temporal dependencies and scalability for very long sequences. These challenges motivate the design of two alignment strategies in the proposed Sequence-Aware Alignment module (see Figure 1b.2): (1) Sequence Context Alignment, which preserves the full temporal sequence, and (2) Sequence Context Compression, which compresses the sequence into a fixed-length latent representation for memory-efficient alignment.

Sequence Context Alignment (SCA). Given a sequence of epoch tokens $\mathbf{E}_{\text{eeg}}$, we augment the sequence with learnable position encodings and `[SESSION_SEP]` tokens to differentiate EEG sessions. A lightweight linear-attention transformer (Wang et al., 2020; Katharopoulos et al., 2020) 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). In an attempt to tackle the memory constraints and improve scalability for long EEG sequences, we explore a compression strategy inspired by Perceiver architectures (Jaegle et al., 2021b;a). 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 1b.3). 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 to the LLM is formed by concatenation,

Table 1: Report generation performance on samples with clinical context across S0001 and S0002. We evaluate under two settings: (i) *Unimodal + Text Only Input*, and (ii) *Unimodal + Text + EEG Features Input*. Best results in orange, and best baseline in blue. Relative improvements are reported.

Method	S0001			S0002		
	BLEU-1	ROUGE-1	METEOR	BLEU-1	ROUGE-1	METEOR
	Unimodal + Text Only Input					
Gemma-3-1b-it	0.1897 ± 0.1340	0.2112 ± 0.1408	0.1567 ± 0.1238	0.1509 ± 0.0898	0.2330 ± 0.0876	0.1319 ± 0.0565
Gemma-3n-e2b-it	0.1967 ± 0.1592	0.2410 ± 0.1722	0.1875 ± 0.1725	0.0925 ± 0.0769	0.2318 ± 0.0600	0.1181 ± 0.0472
Gemma-3-4b-it	0.2372 ± 0.1508	0.2857 ± 0.1590	0.2147 ± 0.1668	0.1538 ± 0.0990	0.2775 ± 0.0558	0.1506 ± 0.0514
Gemma-3n-e4b-it	0.1793 ± 0.1685	0.2416 ± 0.1778	0.1816 ± 0.1786	0.0539 ± 0.0664	0.1998 ± 0.0649	0.0993 ± 0.0490
Medgemma-4b-it	0.1502 ± 0.1725	0.2196 ± 0.1924	0.1529 ± 0.1835	0.0865 ± 0.0855	0.2280 ± 0.0734	0.1198 ± 0.0542
Llama-3.2-1b-instruct	0.0337 ± 0.0667	0.0919 ± 0.0744	0.0529 ± 0.0519	0.0340 ± 0.0700	0.0954 ± 0.0869	0.0494 ± 0.0504
Llama-3.2-3b-instruct	0.1951 ± 0.1630	0.2224 ± 0.1755	0.1843 ± 0.1864	0.1651 ± 0.1110	0.2346 ± 0.0877	0.1430 ± 0.0689
Llama-3.1-8b-instruct	0.2128 ± 0.1702	0.2511 ± 0.1872	0.1839 ± 0.1788	0.1629 ± 0.0970	0.2655 ± 0.0766	0.1480 ± 0.0581
Meta-llama-3-8b-instruct	0.2249 ± 0.1602	0.2673 ± 0.1731	0.1946 ± 0.1809	0.1441 ± 0.0847	0.2588 ± 0.0586	0.1416 ± 0.0481
Qwen3-1.7b	0.1335 ± 0.1481	0.1378 ± 0.1557	0.1102 ± 0.1316	0.0886 ± 0.1095	0.1189 ± 0.1296	0.0735 ± 0.0821
Qwen3-4b-instruct-2507	0.2795 ± 0.1243	0.3038 ± 0.1460	0.2418 ± 0.1440	0.1912 ± 0.0807	0.2690 ± 0.0448	0.1622 ± 0.0434
	Unimodal + Text + EEG Features Input					
Gemma-3-1b-it	0.0068 ± 0.0442	0.0086 ± 0.0524	0.0064 ± 0.0462	0.0957 ± 0.0925	0.1755 ± 0.1157	0.0954 ± 0.0692
Gemma-3n-e2b-it	0.0083 ± 0.0504	0.0116 ± 0.0605	0.0088 ± 0.0543	0.0750 ± 0.0771	0.2190 ± 0.0840	0.1082 ± 0.0546
Gemma-3-4b-it	0.0085 ± 0.0532	0.0093 ± 0.0594	0.0078 ± 0.0558	0.2089 ± 0.0895	0.2886 ± 0.0458	0.1738 ± 0.0476
Gemma-3n-e4b-it	0.0101 ± 0.0348	0.0397 ± 0.0606	0.0231 ± 0.0392	0.0169 ± 0.0486	0.0357 ± 0.0903	0.0180 ± 0.0462
Medgemma-4b-it	0.1618 ± 0.1730	0.2100 ± 0.1919	0.1529 ± 0.1832	0.0888 ± 0.0910	0.1829 ± 0.1128	0.1000 ± 0.0702
Llama-3.2-1b-instruct	0.0310 ± 0.0615	0.0820 ± 0.0727	0.0473 ± 0.0479	0.0312 ± 0.0668	0.1016 ± 0.0773	0.0517 ± 0.0448
Llama-3.2-3b-instruct	0.1692 ± 0.1414	0.1844 ± 0.1611	0.1751 ± 0.1871	0.1779 ± 0.1101	0.2309 ± 0.0903	0.1540 ± 0.0743
Llama-3.1-8b-instruct	0.2040 ± 0.1450	0.2151 ± 0.1693	0.1844 ± 0.1647	0.1960 ± 0.0959	0.2256 ± 0.0904	0.1509 ± 0.0682
Qwen3-1.7b	0.2047 ± 0.1351	0.2051 ± 0.1453	0.1683 ± 0.1304	0.1890 ± 0.1048	0.2049 ± 0.1021	0.1413 ± 0.0724
Qwen3-4b-instruct-2507	0.1809 ± 0.0986	0.2058 ± 0.1192	0.2067 ± 0.1079	0.2483 ± 0.0425	0.2638 ± 0.0341	0.2017 ± 0.0285
CELM - SCC	**0.3383** ± 0.1936	**0.3843** ± 0.1876	**0.2889** ± 0.1866	**0.3767** ± 0.1557	**0.4487** ± 0.1283	**0.3232** ± 0.1261
CELM	**0.4823** ± 0.1920	**0.5565** ± 0.1683	**0.4734** ± 0.1941	**0.5695** ± 0.1702	**0.6408** ± 0.1494	**0.5597** ± 0.1728
Improvement	+72.56%	+83.18%	+95.78%	+129.36%	+122.04%	+177.49%

$\mathbf{H}_{\text{in}} = [\texttt{[EEG\_START]}; \mathbf{H}_{\text{eeg}}; \texttt{[EEG\_END]}; \mathbf{H}_{\text{text}}]$ which is processed autoregressively to generate a structured EEG report text. We employ instruction-tuned Qwen3-4B (Yang et al., 2025) as our base LLM, CBraMod (Wang et al., 2024b) as our EEG encoder and perform supervised fine-tuning using the next-token prediction objective: $\mathcal{L} = -\sum_{t=1}^{T} \log P_\theta(y_t \mid H_{\text{input}}, y_{<t})$ 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 (Zafar et al., 2025; Sun et al., 2025): 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. In our work, we restrict the data to reports 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, long-term monitoring), 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 B.

Baselines. As no existing EEG multimodal models with generative capabilities are available for EEG report generation, we establish baselines by augmenting local LLMs with EEG-derived inputs. We evaluate two settings: (1) Unimodal + text-only input, where the LLMs receive only clinical context as input without EEG data, providing a lower bound that quantifies hallucination and measures performance when the model does not attend to neural signals, and (2) Unimodal + text and EEG features input, where we augment the clinical context with channel-wise spectral band power features (delta, theta, alpha, beta, gamma) extracted from each EEG session, addressing context length limitation that prevents providing raw EEG arrays directly. We evaluate LLMs from LLaMA-3 (Grattafiori et al., 2024), Gemma-3 (Team et al., 2025), and Qwen-3 (Yang et al.,

2025) families, along with MedGemma (Sellergren et al., 2025) as a medical baseline. Additional implementation details are provided in Appendix C. We report smoothed BLEU-1 (Papineni et al., 2002), ROUGE-1 (Lin, 2004), and METEOR (Banerjee & Lavie, 2005) scores to evaluate report generation performance. Comprehensive results and metrics are provided in Appendix E.

5.1 Report Generation Performance Comparison

Table 1 compares CELM against unimodal baselines using either text-only or text with handcrafted EEG features as input, evaluated on test samples containing patient history. We report results for two CELM variants: **CELM-SCC**, which uses sequence context compression to compress and produce a fixed number of projected EEG tokens, and **CELM** , which uses non-compressed EEG representations via sequence context alignment. When EEG features are provided, most baseline LLMs show modest improvements over text-only inputs on S0002 ($0.1912 \rightarrow 0.2483$), except LLaMA-3.1-8B, which fails due to its 8K token context limit. However, on S0001, performance degrades across baselines due to greater report complexity introduced by multiple sections.

Both CELM variants substantially outperform all baselines, with CELM achieving approximately two-fold improvement over the strongest baseline (e.g., ROUGE score on S0002, $0.2886 \rightarrow 0.6408$). Although CELM-SCC outperforms the baselines, its gap relative to the non-compressed model (0.4487 vs 0.6408) highlights the need for efficient compression strategies that preserve clinically relevant information from long EEG sequences within LLM context constraints.

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 as input. This experiment also serves as a test of potential language-model bias, in which reports could be generated by inferring plausible clinical context while ignoring EEG inputs. We evaluate this task on a subset of the S0002 dataset containing reports without patient clinical context, and the results are summarized in Table 2. CELM significantly outperforms all baselines across all metrics, demonstrating that it effectively extracts and utilizes clinically relevant information from EEG recordings.

Table 2: Zero-context report generation performance on S0002. Best results in orange, and the best baseline in blue. Relative improvements are also reported.

Method	BLEU-1	ROUGE-1	METEOR
	Unimodal + Text + EEG Features Input		
Gemma-3-1b-it	0.1019 ± 0.0782	0.1935 ± 0.0953	0.1052 ± 0.0556
Gemma-3n-e2b-it	0.0717 ± 0.0710	0.2203 ± 0.0581	0.1116 ± 0.0411
Gemma-3-4b-it	0.1699 ± 0.0885	0.2557 ± 0.0484	0.1514 ± 0.0412
Gemma-3n-e4b-it	0.0058 ± 0.0304	0.0127 ± 0.0509	0.0067 ± 0.0281
Medgemma-4b-it	0.0559 ± 0.0660	0.1591 ± 0.0959	0.0820 ± 0.0520
Llama-3.2-1b-instruct	0.0265 ± 0.0627	0.1041 ± 0.0723	0.0530 ± 0.0396
Llama-3.2-3b-instruct	0.1261 ± 0.1028	0.1995 ± 0.0806	0.1214 ± 0.0617
Llama-3.1-8b-instruct	0.1556 ± 0.0817	0.2090 ± 0.0737	0.1298 ± 0.0506
Qwen3-1.7b	0.1772 ± 0.0999	0.1819 ± 0.0938	0.1304 ± 0.0688
Qwen3-4b-instruct-2507	0.2260 ± 0.0592	0.2315 ± 0.0501	0.1790 ± 0.0420
CELM - SCC	$\mathbf{0.2991} \pm 0.1276$	$\mathbf{0.3793} \pm 0.0892$	$\mathbf{0.2574} \pm 0.0891$
CELM	$\mathbf{0.4652} \pm 0.1884$	$\mathbf{0.5248} \pm 0.1869$	$\mathbf{0.4390} \pm 0.1869$
Improvement	+105.84%	+105.24%	+145.25%

5.3 Alignment Module Ablation

The alignment module is a critical component of EEG-language models. In our initial experiments, we found that naive linear projection into the LLM embedding space is insufficient for the EEG domain. In this section, we present an ablation study comparing different alignment strategies. We evaluate two baseline alignment modules: (1) Linear projector, which directly maps EEG embeddings to the LLM embedding space using a single linear layer, and (2) Perceiver projector, which employs learnable query tokens that attend to EEG sequences via cross-attention followed by feedforward and projection. We compare these against our proposed SCA and SCC projectors, described in Section 4.3. Results are summarized in Figure 2.

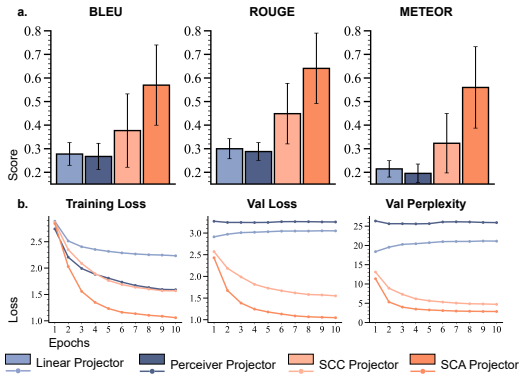

Figure 2: (a) Report generation performance of different projector variants. (b) Training dynamics of each variant.

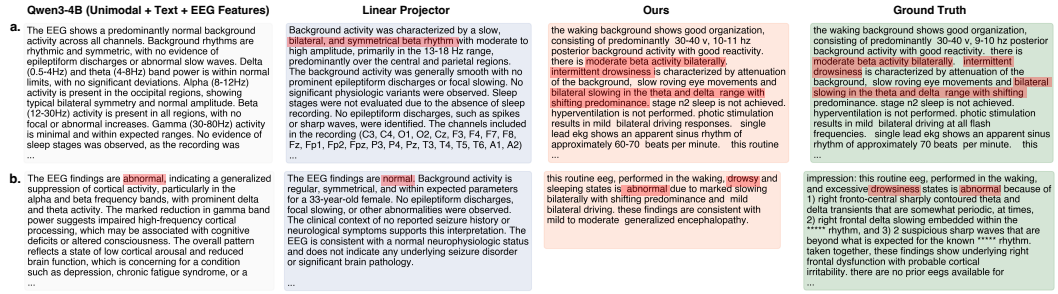

Figure 3: Examples of generated reports on S0002. Comparisons between the unimodal baseline (text + EEG features), the linear-projector alignment variant, CELM, and the ground-truth reports. (a) EEG description/details; (b) Impression/interpretation.

As shown in Figure 2a, the SCA projector consistently achieves the best performance across all evaluation metrics. Analysis of training dynamics in Figure 2b further highlights the importance of modeling inter-epoch temporal dependencies prior to alignment. While the Linear and Perceiver projectors' training losses decrease, their validation losses and perplexities increase, indicating overfitting. In contrast, both SCA and SCC exhibit more stable convergence, with lower validation loss and perplexity over time, and SCA converges faster across all variants. Notably, SCA achieves lower validation perplexity than other variants, indicating substantially higher confidence in token generation. Overall, this study confirms that modeling temporal dependencies between EEG epochs before projection is essential for effective EEG-to-language alignment.

5.4 Qualitative Analysis

Figure 3 presents qualitative comparisons on S0002 between the best-performing unimodal baseline (text with EEG features), the linear projector variant, and CELM , against the ground truth report. Figure 3a shows the EEG description/details section, while Figure 3b shows the impression/interpretation section. Across these examples, CELM produces reports that are more closely aligned with the ground truth, correctly identifying findings such as moderate bilateral beta activity, intermittent drowsiness, and bilateral slowing in the theta and delta ranges, while 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 for learning to generate clinical reports from unstructured notes and long EEG recordings. At the same time, they underscore the need for more rigorous, standardized benchmarking to evaluate these models.

6 Discussion and Implication

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. ☞*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 3 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.

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.

ACKNOWLEDGEMENT

This work was supported by NSF awards SCH-2205289, SCH-2014438, IIS-2034479, and JSPS KAKENHI Grant-in-Aid for Scientific Research Number JP24K20778. This project has been funded by the Jump ARCHES endowment through the Health Care Engineering Systems Center. We also wish to acknowledge the creators of the Harvard Electroencephalography Database for their valuable contribution in making this dataset publicly accessible.

REFERENCES

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*, pp. 65–72, 2005.

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*, pp. 513–531. PMLR, 2019.

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.

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.

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.

Sam Gijsen and Kerstin Ritter. EEG-language pretraining for highly label-efficient clinical phenotyping. In *Forty-second International Conference on Machine Learning*, 2025.

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.

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.

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*, 2021a.

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*, pp. 4651–4664. PMLR, 2021b.

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.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020.

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.

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.

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

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

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

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.

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.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

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

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*, pp. 8748–8763. PmLR, 2021.

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.

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.

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.

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.

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.

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, 2024a.

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*, 2024b.

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

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, pp. 5350–5358, 2022.

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.

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.

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.

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

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.

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.

APPENDIX

Contents

A MORE RELATED WORKS

EEG Foundation Models. Recent years have witnessed rapid advances in EEG foundation models. BENDR (Kostas et al., 2021), BIOT (Yang et al., 2023), LaBraM (Jiang et al., 2024), TFM-Tokenizer (Pradeepkumar et al., 2025), EEGPT (Wang et al., 2024a), LUNA (Döner et al., 2025), REVE (Ouahidi et al., 2025), and CBraMod (Wang et al., 2024b) 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.

B DETAILS OF EEG-REPORT BENCHMARK

This section provides detailed information on the EEG–Report benchmark constructed in our study. Table 3 summarizes key statistics, including the number of EEG–report pairs, patient counts, and total EEG recording duration for each site. Figure 4 further illustrates dataset distributions across EEG session duration, patient demographics (age and gender), and EEG report sections.

Table 3: Statistics of the EEG–Report benchmark.

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

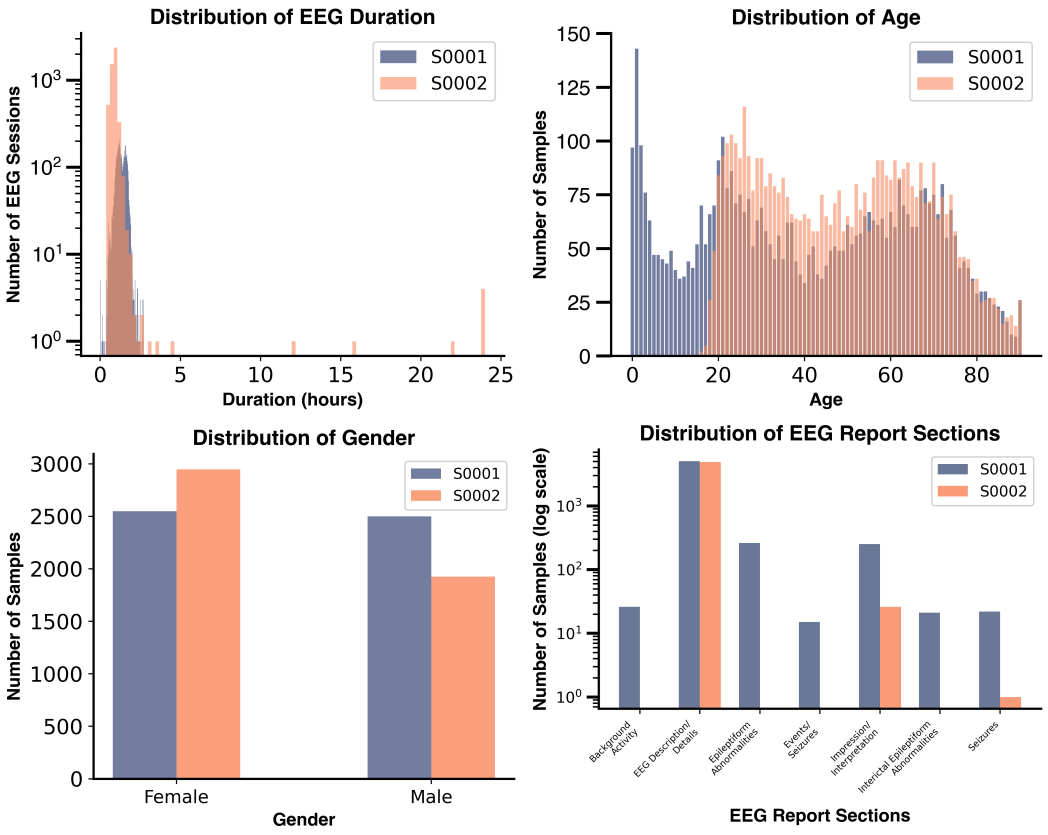

Figure 4: Dataset statistics of the filtered and constructed EEG-Report Benchmark used in our study

C Additional Experiment Details

C.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 6).

C.2 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 smoothed

BLEU-1, ROUGE-1, and METEOR scores. In Appendix E, we additionally report smoothed BLEU-4, ROUGE-2, ROUGE-L, and ROUGE-LSUM. We further include per-sample score distributions for all metrics in Appendix **??**.

C.3 Baselines and Prompts

In this section, we provide additional details on the baseline models and the prompts used in our experiments. Table 4 summarizes all LLMs considered in our study along with their respective context lengths[1], further underscoring the practical challenges of directly conditioning LLMs on long-duration EEG recordings. Figures 5 and 6 present 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.

Table 4: 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

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

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 5: Prompt for unimodal + text only baselines.

Unimodal + Text + EEG Features Prompt

EEG-DERIVED STATISTICS

```
{"eeg_session_0":{"delta (0.5-4Hz) band power (dB)": [[-22.90, -21.60, -22.20, ...]],
                  "theta (4-8Hz) band power (dB)": [[-25.30, -24.80, -24.50, ...]],
                  "alpha (8-12Hz) band power (dB)": [[-27.50, -26.60, -27.30, ...]],
                  "beta (12-30Hz) band power (dB)": [[-30.60, -30.20, -30.70, ...]],
                  "gamma (30-80Hz) band power (dB)": [[-36.30, -36.10, -36.30, ...]],
```

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

D 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, data loaders, and pretrained weights at `https://anonymous.4open.science/r/CELM-3AF4`. Additionally, we provide the detailed implementation details of our approach in this section and the prompts used for the baselines in Appendix C. 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.

D.1 MODEL DETAILS AND HYPERPARAMETERS

We use CBraMod (Wang et al., 2024b) 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 5, we provide the hyperparameters of the alignment modules used in our study.

Table 5: 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

D.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 6.

Table 6: 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

D.3 PROMPTS

Figure 7 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 7: Prompt for CELM.

E Additional Experiment Results

E.1 Extended Report Generation Performance on S0001

Table 7 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 7: 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
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

E.2 Extended Report Generation Performance on S0002

Tables 8 and 9 present a comprehensive performance comparison of EEG report generation on the S0002 site under settings with and 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, except for BERTScore in the zero-context setting.

Table 8: 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
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 9: 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-3-1b-it	0.1019 ± 0.0782	0.0105 ± 0.0082	0.1935 ± 0.0953	0.0185 ± 0.0148	0.1046 ± 0.0509	0.1046 ± 0.0509	0.6441 ± 0.2920	0.1052 ± 0.0556
Gemma-3n-e2b-it	0.0717 ± 0.0710	0.0089 ± 0.0090	0.2203 ± 0.0581	0.0331 ± 0.0175	0.1341 ± 0.0331	0.1346 ± 0.0332	0.7716 ± 0.0826	0.1116 ± 0.0411
Gemma-3-4b-it	0.1699 ± 0.0885	0.0189 ± 0.0122	0.2557 ± 0.0484	0.0382 ± 0.0203	0.1341 ± 0.0271	0.1341 ± 0.0271	0.7875 ± 0.0203	0.1514 ± 0.0412
Gemma-3n-e4b-it	0.0058 ± 0.0304	0.0007 ± 0.0037	0.0127 ± 0.0509	0.0021 ± 0.0094	0.0072 ± 0.0288	0.0072 ± 0.0288	0.0471 ± 0.1844	0.0067 ± 0.0281
Medgemma-4b-it	0.0559 ± 0.0660	0.0069 ± 0.0080	0.1591 ± 0.0959	0.0208 ± 0.0191	0.0917 ± 0.0535	0.0917 ± 0.0535	0.6073 ± 0.3121	0.0820 ± 0.0520
Llama-3.2-1b-instruct	0.0265 ± 0.0627	0.0033 ± 0.0074	0.1041 ± 0.0723	0.0109 ± 0.0116	0.0632 ± 0.0361	0.0632 ± 0.0361	0.6324 ± 0.2729	0.0530 ± 0.0396
Llama-3.2-3b-instruct	0.1261 ± 0.1028	0.0125 ± 0.0094	0.1995 ± 0.0806	0.0304 ± 0.0188	0.1114 ± 0.0425	0.1141 ± 0.0453	0.7328 ± 0.1553	0.1214 ± 0.0617
Llama-3.1-8b-instruct	0.1556 ± 0.0817	0.0160 ± 0.0088	0.2090 ± 0.0737	0.0305 ± 0.0170	0.1142 ± 0.0416	0.1149 ± 0.0417	0.7010 ± 0.1985	0.1298 ± 0.0506
Qwen3-1.7b	0.1772 ± 0.0999	0.0164 ± 0.0104	0.1819 ± 0.0938	0.0187 ± 0.0149	0.0982 ± 0.0498	0.0991 ± 0.0502	0.6265 ± 0.3013	0.1304 ± 0.0688
Qwen3-4b-instruct-2507	0.2260 ± 0.0592	0.0183 ± 0.0071	0.2315 ± 0.0501	0.0241 ± 0.0139	0.1162 ± 0.0254	0.1162 ± 0.0254	0.7418 ± 0.1337	0.1790 ± 0.0420
CELM-SCC	**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
CELM	**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

E.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 10. Across most evaluation settings, the SCT 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 10: 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
SCT Perceiver 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
SCT 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
SCT Perceiver 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
SCT 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
SCT Perceiver 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
SCT 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

E.4 EEG Encoder Ablation

We conducted an ablation study to analyze report generation performance across different EEG encoders. Specifically, we compared CBraMod (Wang et al., 2024b) and LaBraM (Jiang et al., 2024), 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 11: 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

E.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 4 in Appendix). Figure 8 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 12.

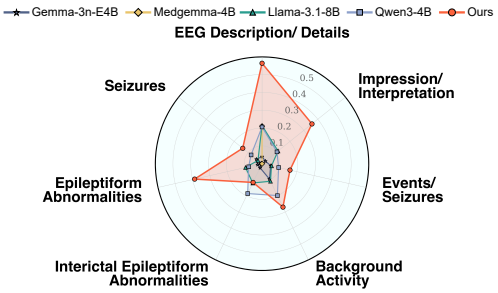

Figure 8: Section-wise comparison of report generation performance between CELM and the best-performing baselines from different LLM families.

Table 12: 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

Table 12 reports the complete section-wise performance across all models and evaluation metrics. These results extend the main text by providing detailed comparisons for each report section and further show the consistency of CELM across diverse clinical report components.

E.6 Qualitative Analysis and Case Studies

This subsection presents qualitative analyses and representative case studies to complement the quantitative evaluation. In Figure 9, 10, 11, and 12, we compare generated reports from the unimodal baseline, CELM-SCC, and CELM against ground-truth reports across multiple EEG report sections and datasets.

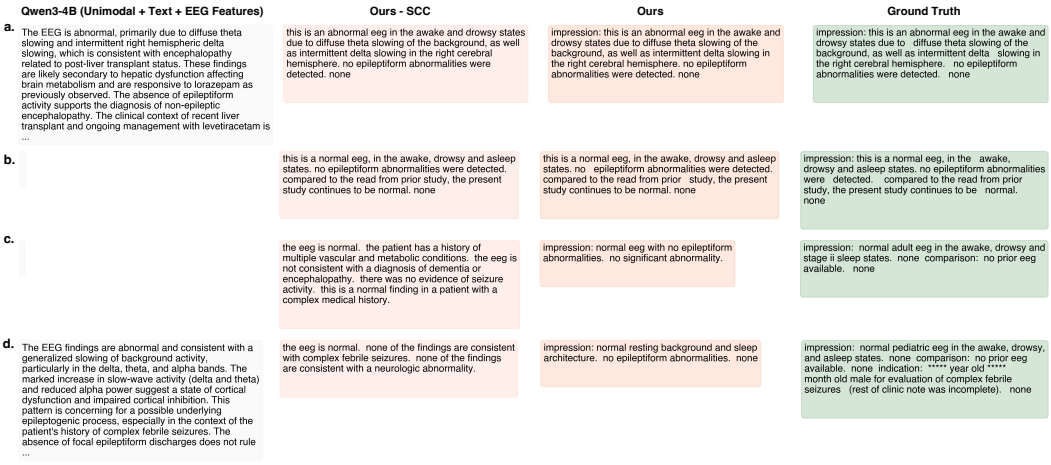

Figure 9: Impression/interpretation generation examples on S0001. Examples are ordered from (a) to (d) by decreasing ROUGE-1 score. We compare outputs from the unimodal baseline (text + EEG features), CELM-SCC, CELM, and the ground-truth reports.

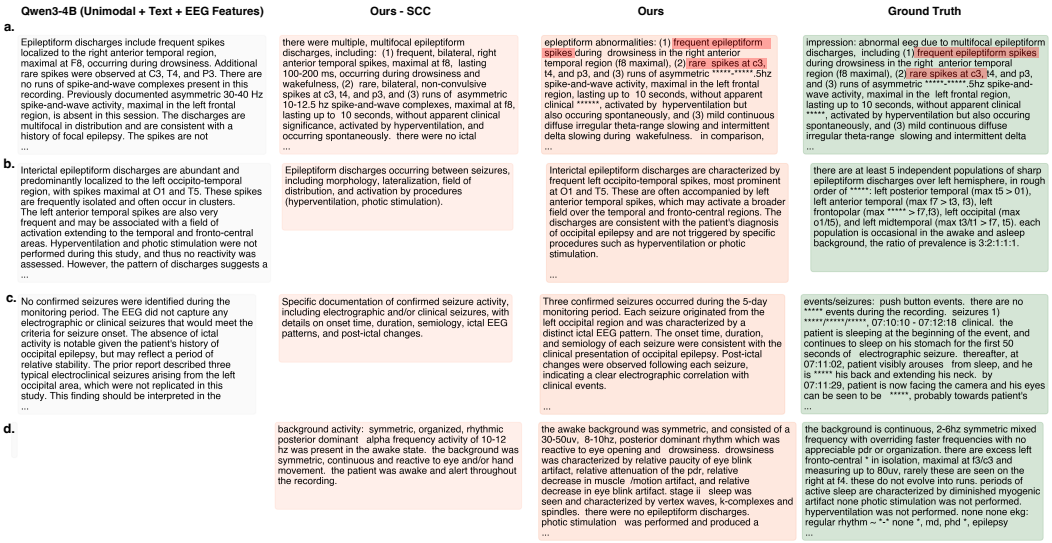

Figure 10: Section-wise generation examples on S0001. (a) Epileptiform abnormalities, (b) Interictal epileptiform abnormalities, (c) Seizures, and (d) Background activity. We compare outputs from the unimodal baseline (text + EEG features), CELM-SCC, CELM, and the ground-truth reports.

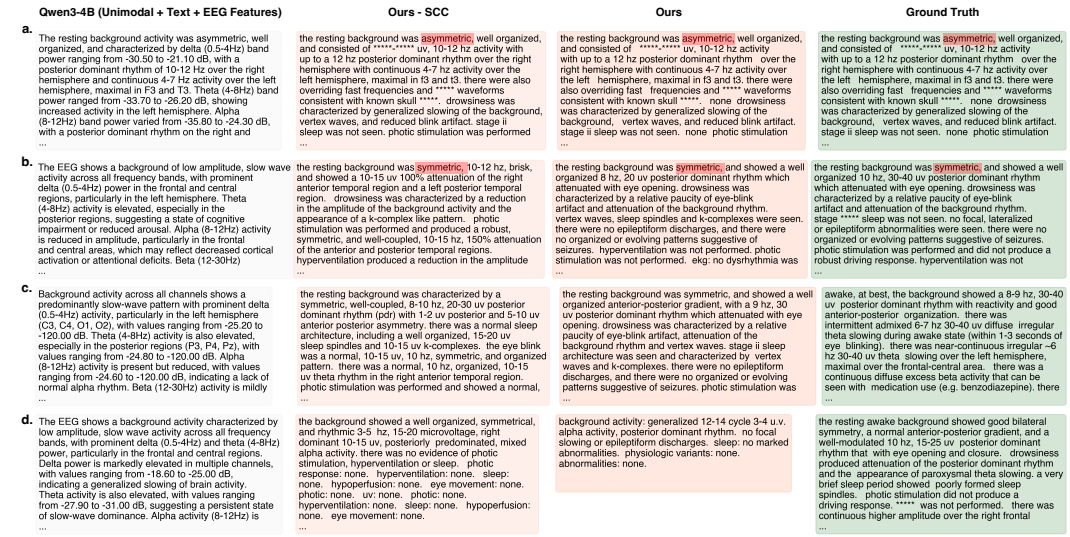

Figure 11: EEG description/details generation examples on S0001. Examples are ordered from (a) to (d) by decreasing ROUGE-1 score. We compare outputs from the unimodal baseline (text + EEG features), CELM-SCC, CELM, and the ground-truth reports.

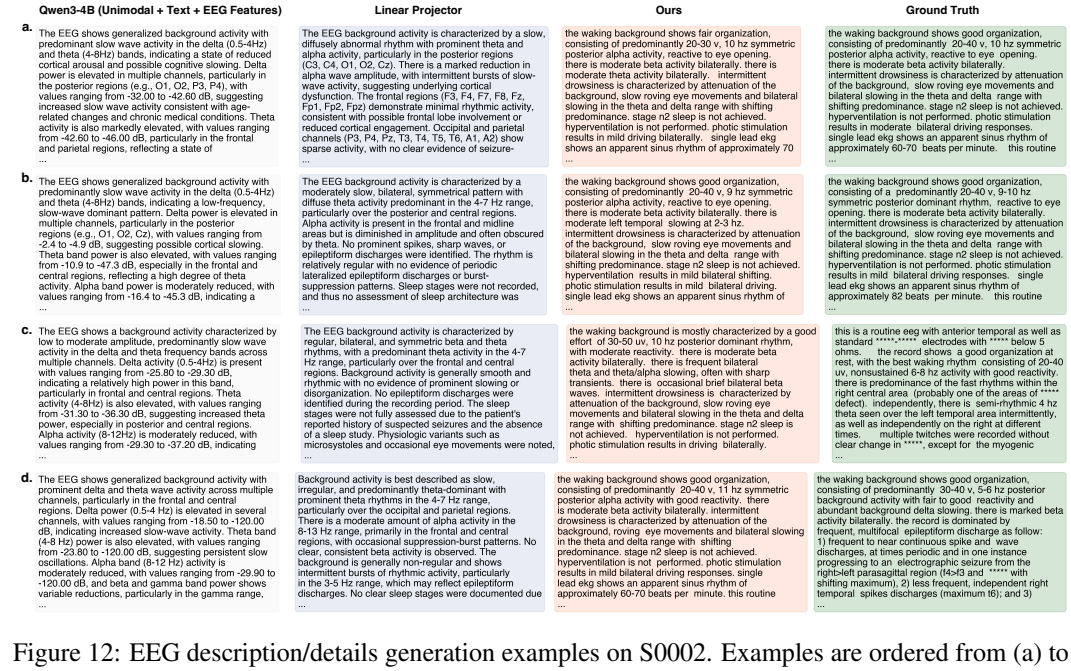

Figure 12: EEG description/details generation examples on S0002. Examples are ordered from (a) to (d) by decreasing ROUGE-1 score. We compare outputs from the unimodal baseline (text + EEG features), CELM-SCC, CELM, and the ground-truth reports.