NEURAL INFORMATION PROCESSING SYSTEMS

Neural Signals Generate Clinical Notes in the Wild

Jathurshan Pradeepkumar[1] Zheng Chen[2] Jimeng Sun[1]

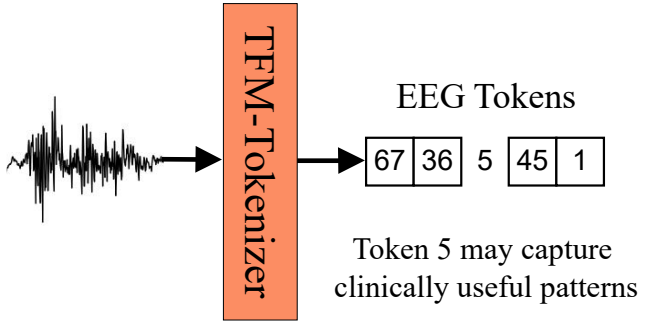
[1]University of Illinois Urbana-Champaign [2]Osaka University

Presenter: Jathurshan Pradeepkumar

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

SANKEN THE UNIVERSITY OF OSAKA

Problem



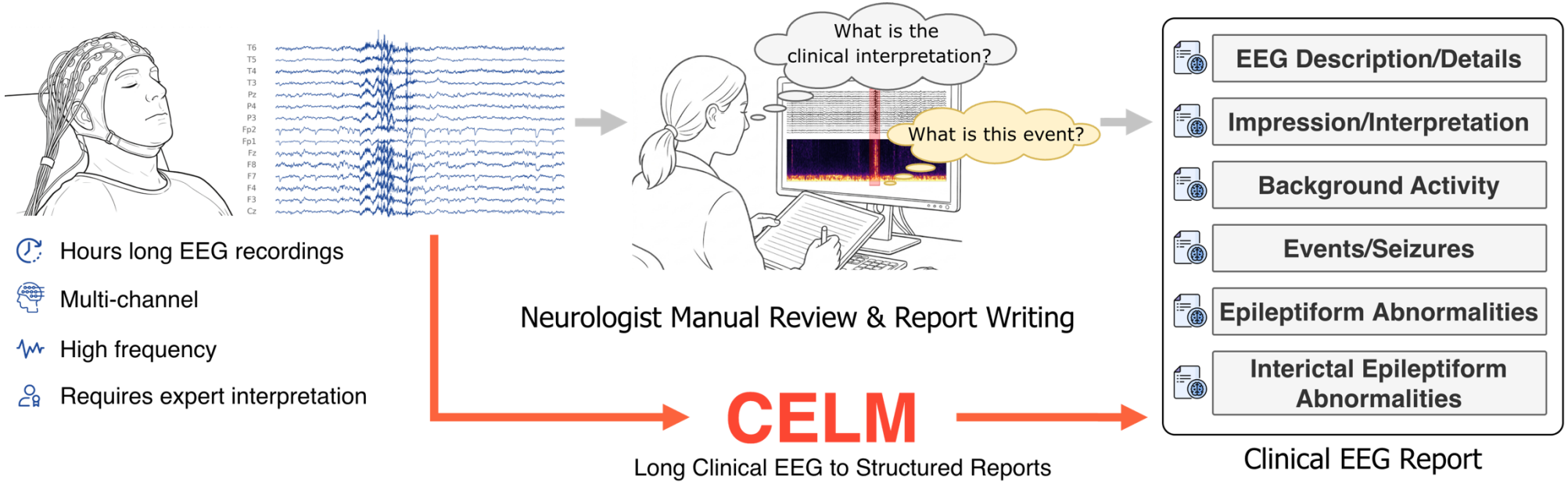
EEG Foundation Modeling

- Foundation models have revolutionized natural language processing.
- This success inspired a paradigm shift in EEG analysis toward the development of task-agnostic foundation models.

Tokenization Challenge

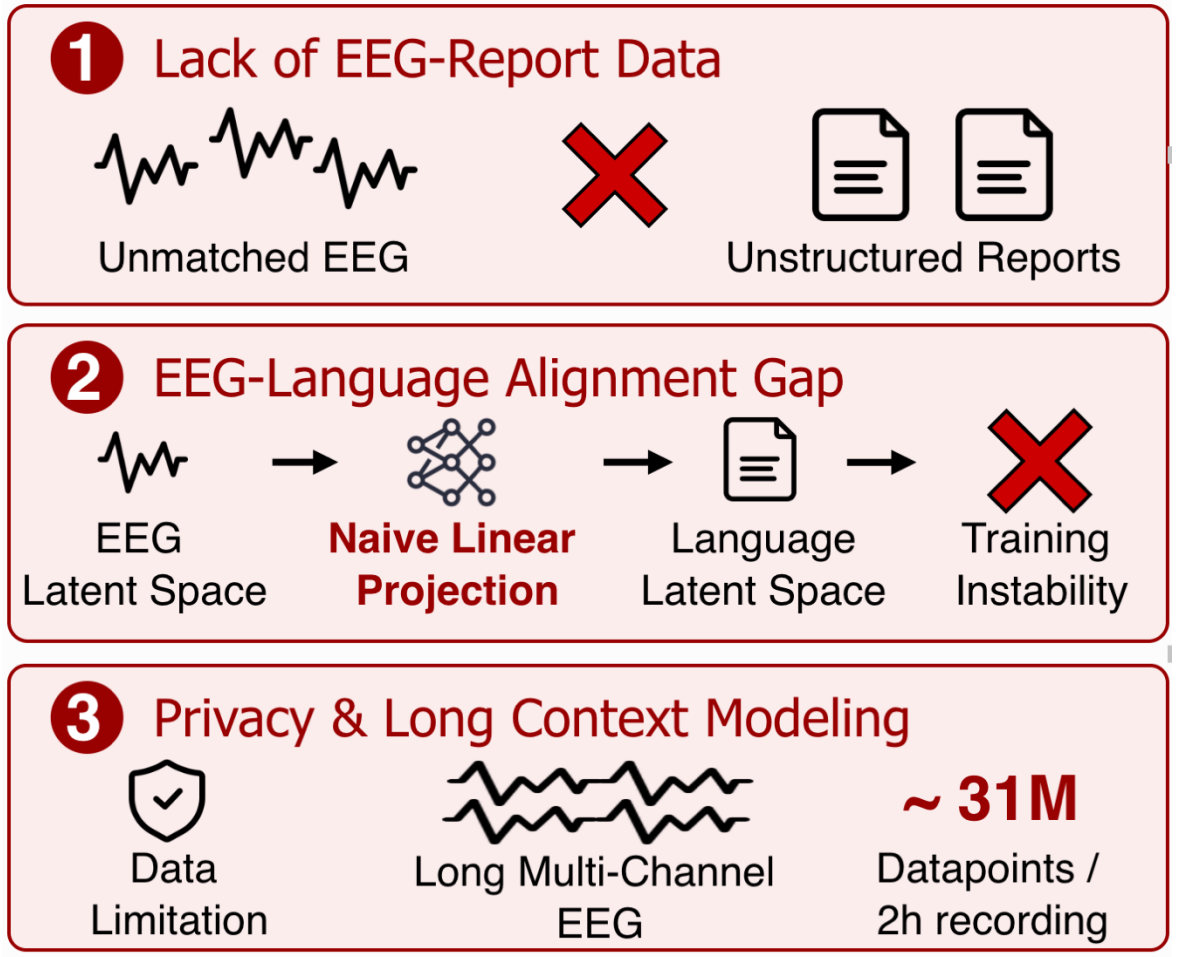
- Tokenization, a core component in NLP, transforms raw text into meaningful tokens, which reduces data complexity and introduces a helpful inductive bias in foundation models.
- Despite progress in EEG FMs, an important open problem remains:
  - **How to design an effective tokenization method for EEG signals?**
- Existing EEG foundation models tokenize signals by directly segmenting continuous EEGs into short-duration → This fails to capture reusable, statistically grounded tokens from the signal.

Problem



- EEG captures brain activity with millisecond-level resolution and is widely used in neurological diagnosis.
- Interpreting long EEG recordings and writing neurology reports is a time-consuming and labor-intensive task, that requires substantial clinical expertise.
- Existing EEG FMs are limited to classifying short EEG segments and mapping to fixed templates, limiting end-to-end, long-context, and flexible clinical report generation.

Challenges

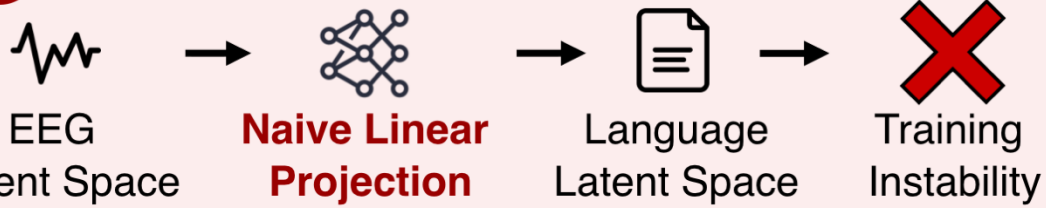

1 Lack of EEG-Report Data
Unmatched EEG
Unstructured Reports
2 EEG-Language Alignment Gap
EEG Latent Space
Naive Linear Projection
Language Latent Space
Training Instability
3 Privacy & Long Context Modeling
Data Limitation
Long Multi-Channel EEG
~ 31M
Datapoints / 2h recording

Our Contributions

1 Lack of EEG-Report Data



2 EEG-Language Alignment Gap



3 Privacy & Long Context Modeling



Benchmark Construction

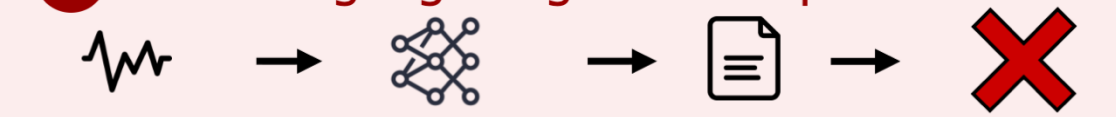


Our Contributions

1 Lack of EEG-Report Data



2 EEG-Language Alignment Gap



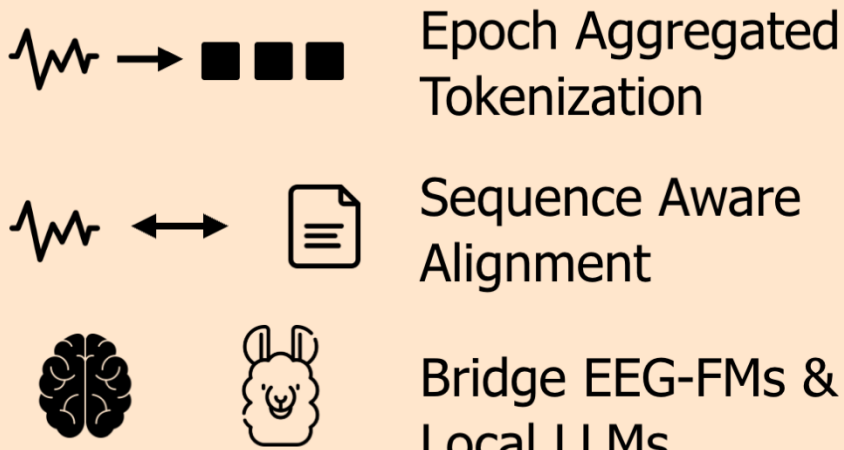
3 Privacy & Long Context Modeling



Benchmark Construction

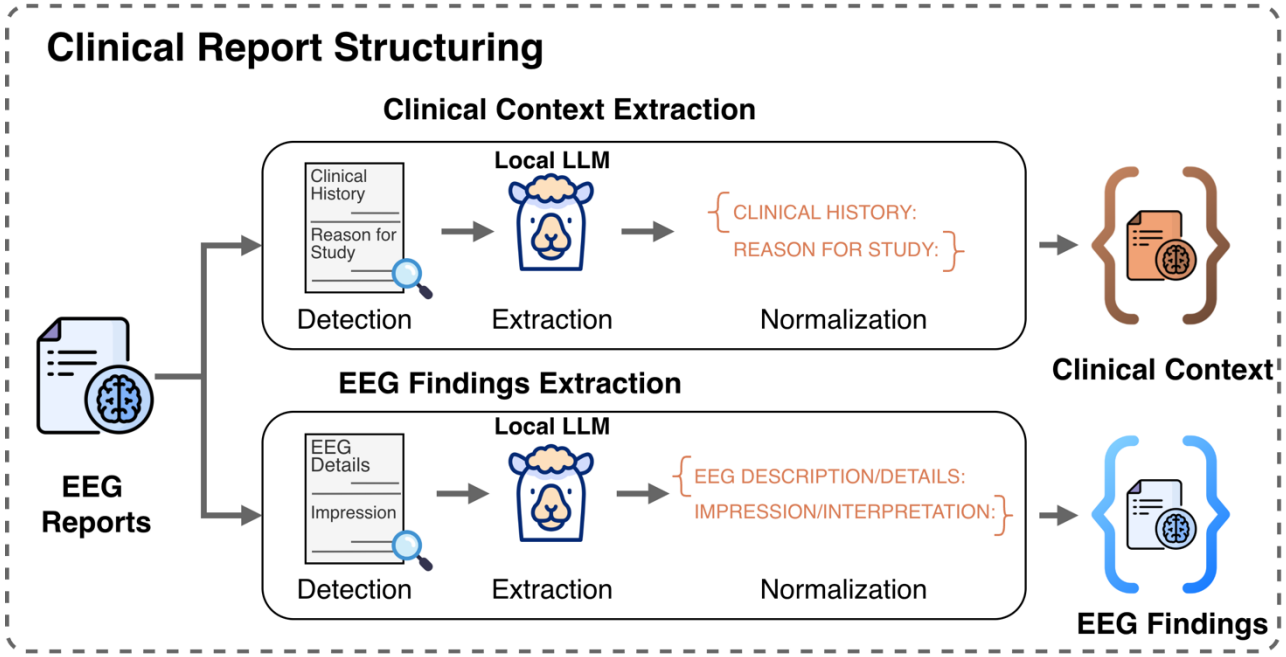


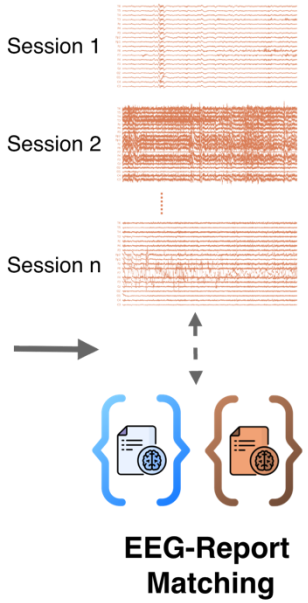
Clinical EEG Language Model



EEG–Report Benchmark Construction Pipeline







EEG Description/Details

The record shows good organization at rest, consisting of predominantly *****-***** uv, 9 hz posterior alpha activity with fair reactivity. there is increased beta activity bilaterally...

Impression/Interpretation

Abnormal ***** due to: 1. frequent high-amplitude 3 hz frontally-predominant generalized spike-wave discharges, occurring both sporadically and in sustained runs >10 seconds without clear clinical correlate 2. numerous electroclinical seizures consisting of frontally-predominant....

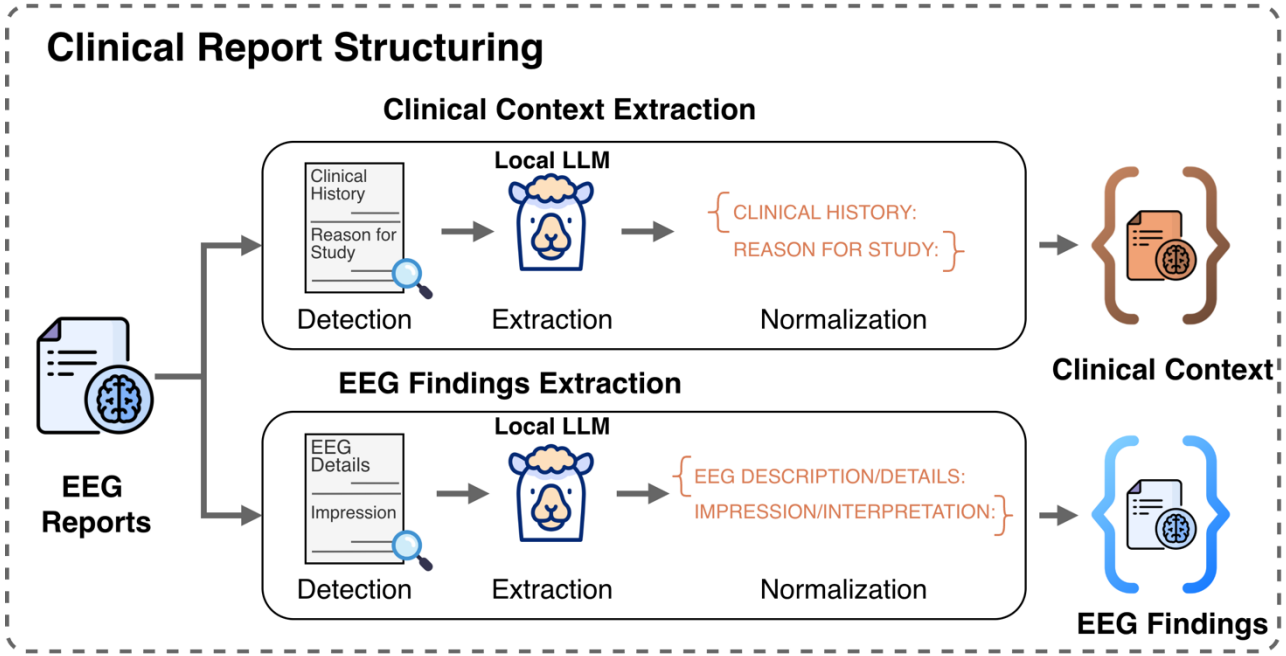
Interictal Epileptiform Abnormalities

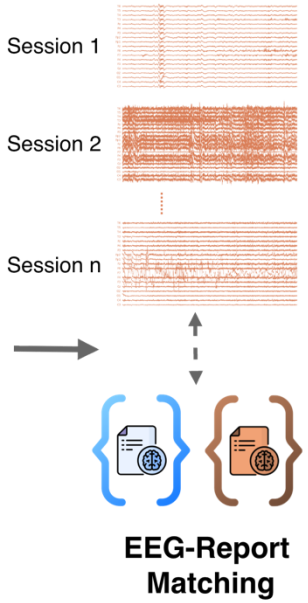
Frequent 3 hz frontally-predominant generalized spike-wave discharges, up to 400 uv, and occurring during both wakefulness and sleep. these rarely occur sporadically, but more frequently occur in runs lasting >10 seconds as described below.

Chenxi Sun, et al. Harvard electroencephalography database: A comprehensive clinical electroencephalographic resource from four boston hospitals. Epilepsia, 2025.

EEG–Report Benchmark Construction Pipeline







EEG Description/Details

The record shows good organization at rest, consisting of predominantly *****-***** uv, 9 hz posterior alpha activity with fair reactivity. there is increased beta activity bilaterally...

Impression/Interpretation

Abnormal ***** due to: 1. frequent high-amplitude 3 hz frontally-predominant generalized spike-wave discharges, occurring both sporadically and in sustained runs >10 seconds without clear clinical correlate 2. numerous electroclinical seizures consisting of frontally-predominant....

Interictal Epileptiform Abnormalities

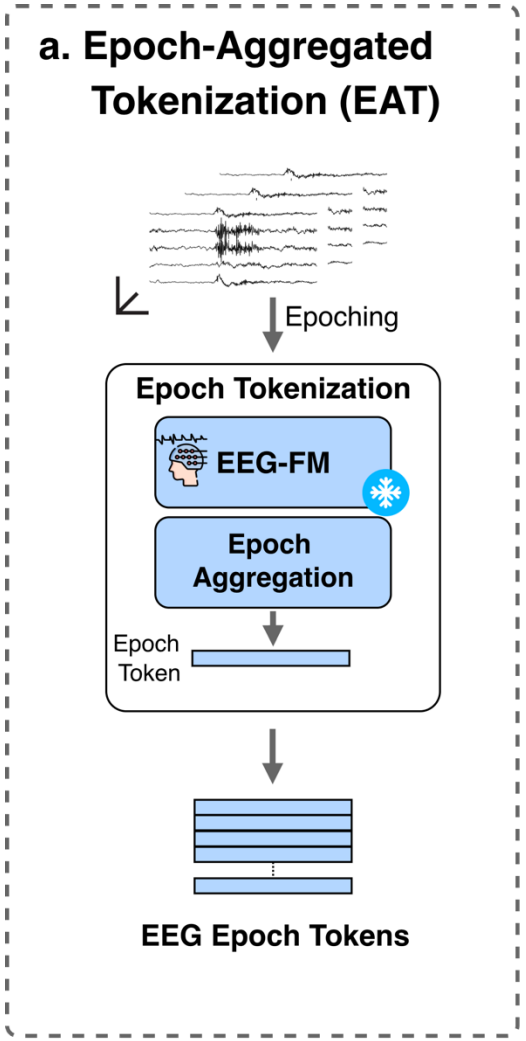
Frequent 3 hz frontally-predominant generalized spike-wave discharges, up to 400 uv, and occurring during both wakefulness and sleep. these rarely occur sporadically, but more frequently occur in runs lasting >10 seconds as described below.

Table 4: Statistics of the EEG–Report benchmark.

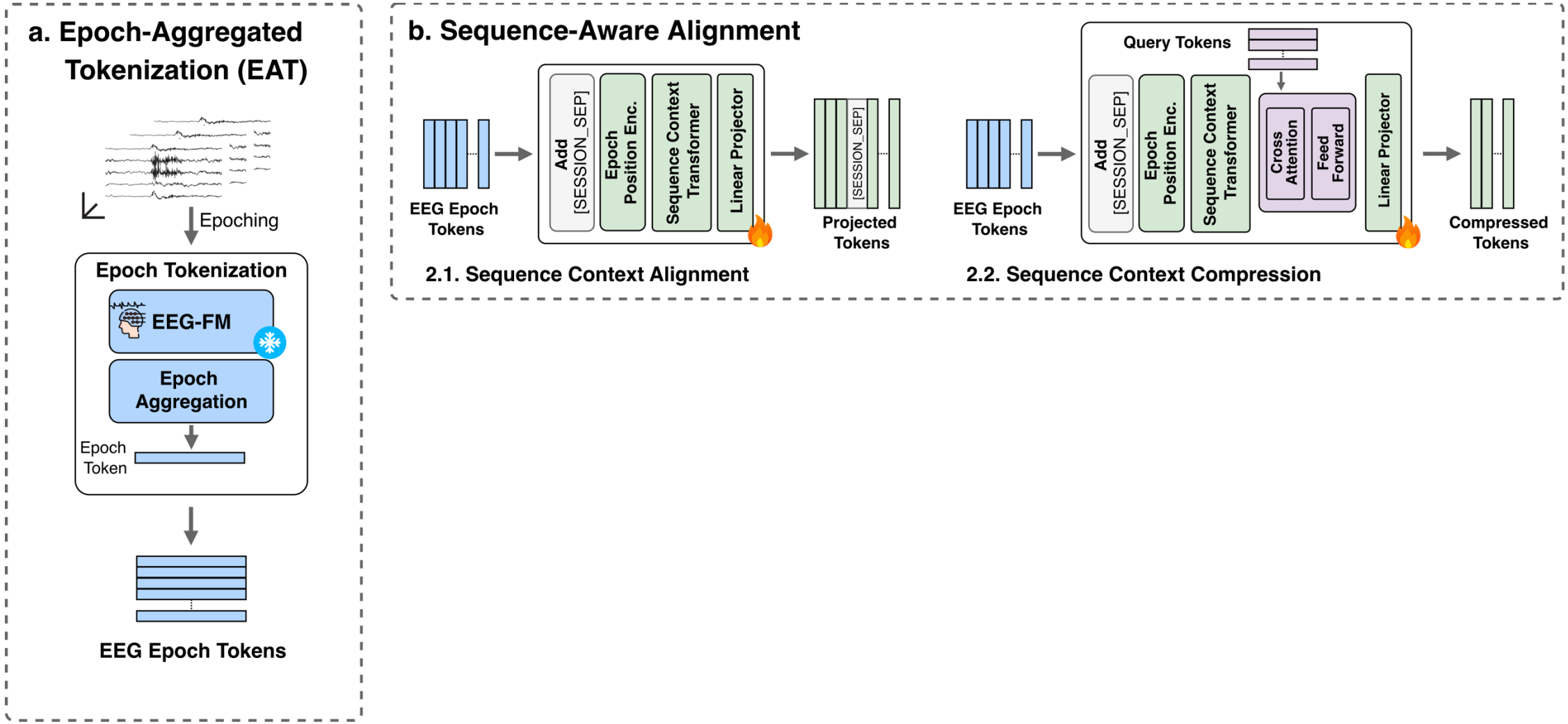
Site	# of Paired EEG & Reports	# of Patients	Total Duration (hrs)
S0001	5049	4669	6639
S0002	4873	4379	4367

Chenxi Sun, et al. Harvard electroencephalography database: A comprehensive clinical electroencephalographic resource from four boston hospitals. Epilepsia, 2025.

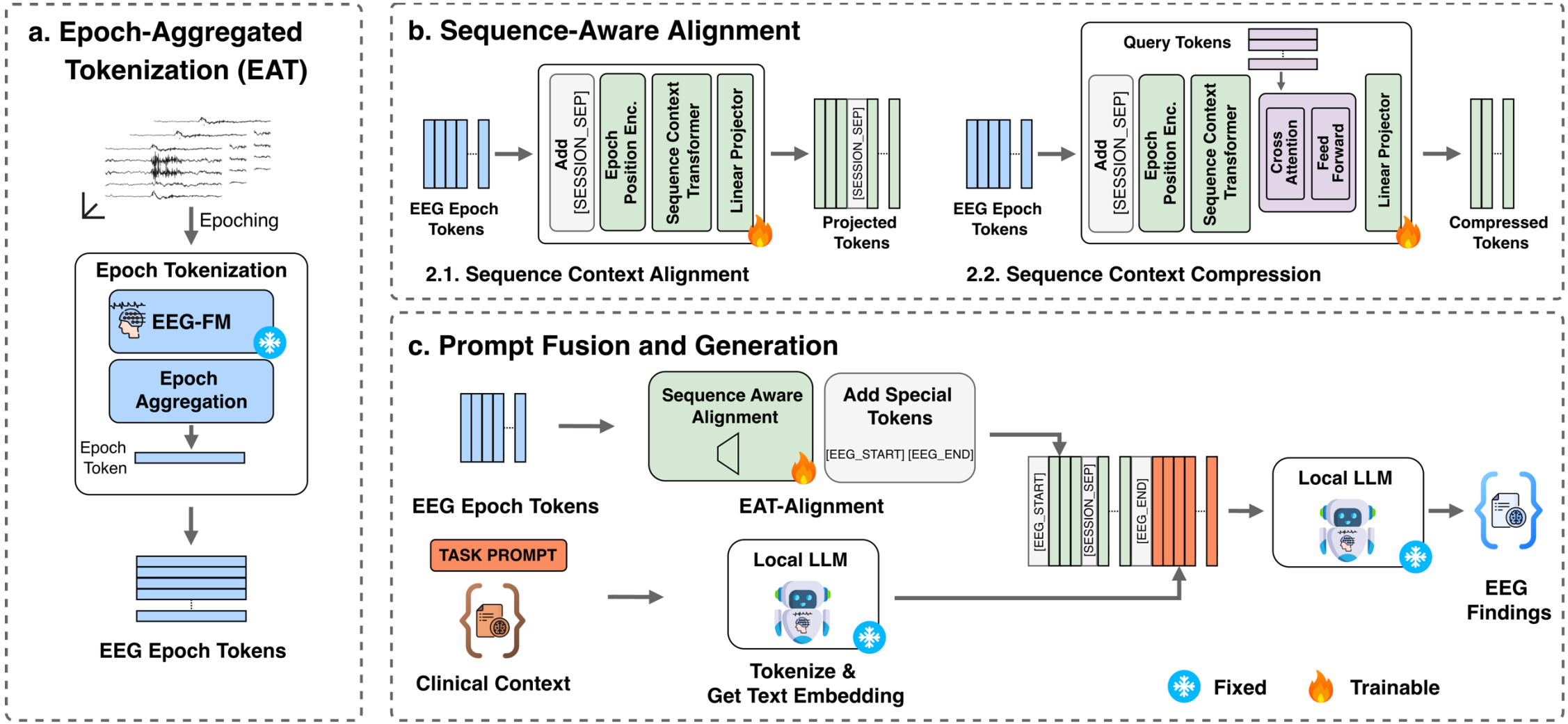
CELM: Clinical EEG Language Model



CELM: Clinical EEG Language Model



CELM: Clinical EEG Language Model

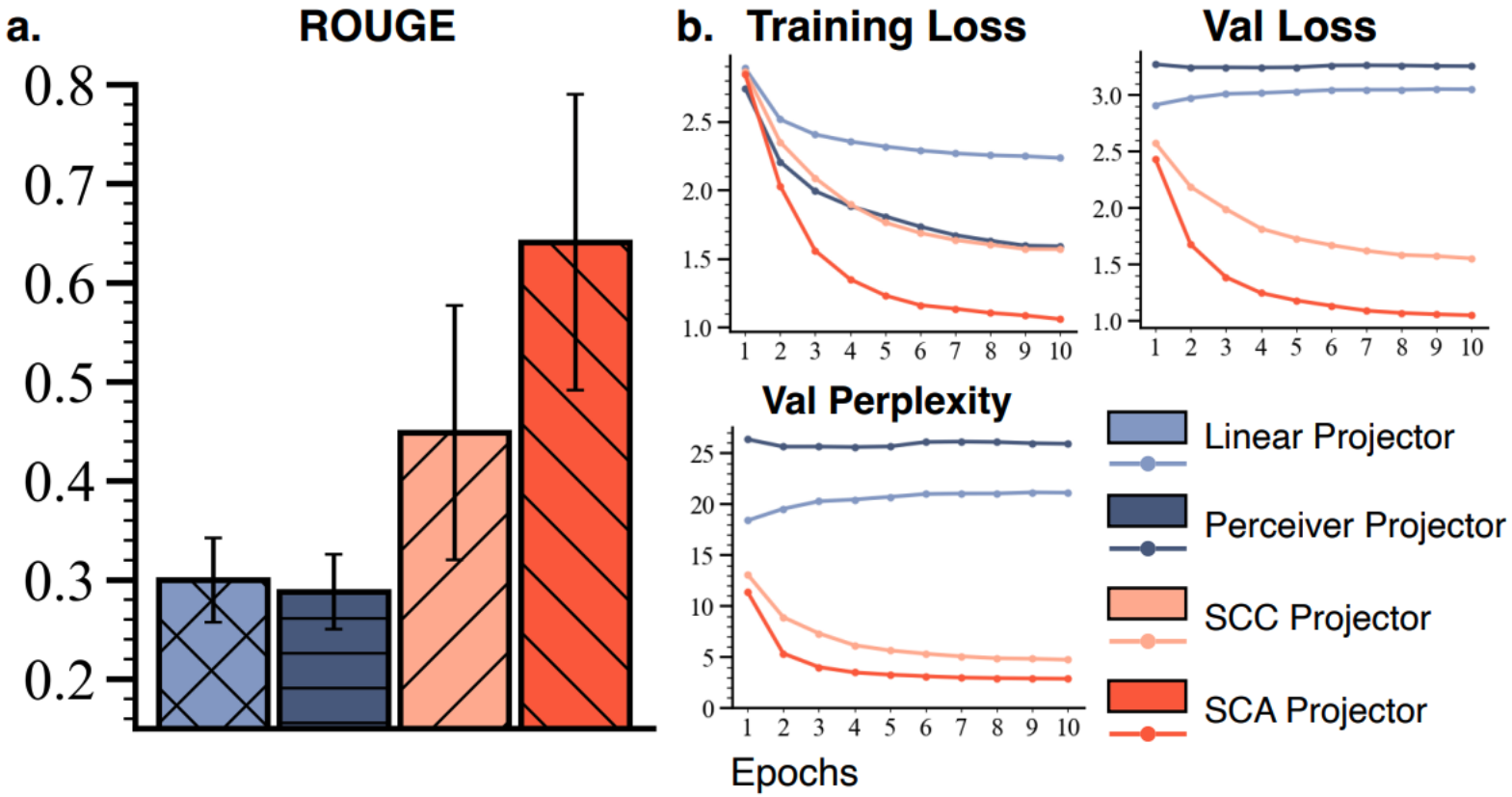


Report Generation Performance

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.

Method	S0001			S0002		
	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%

Alignment Module Ablation



Are Generated Reports Grounded in EEG Signals?

Zero-Context Report Generation

Method	S0001		S0002	
	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%

EEG Grounding Control Experiments

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%

Qualitative Analysis and Expert Evaluation

Manual qualitative analysis

Human Expert Study and LLM-as-a-judge results

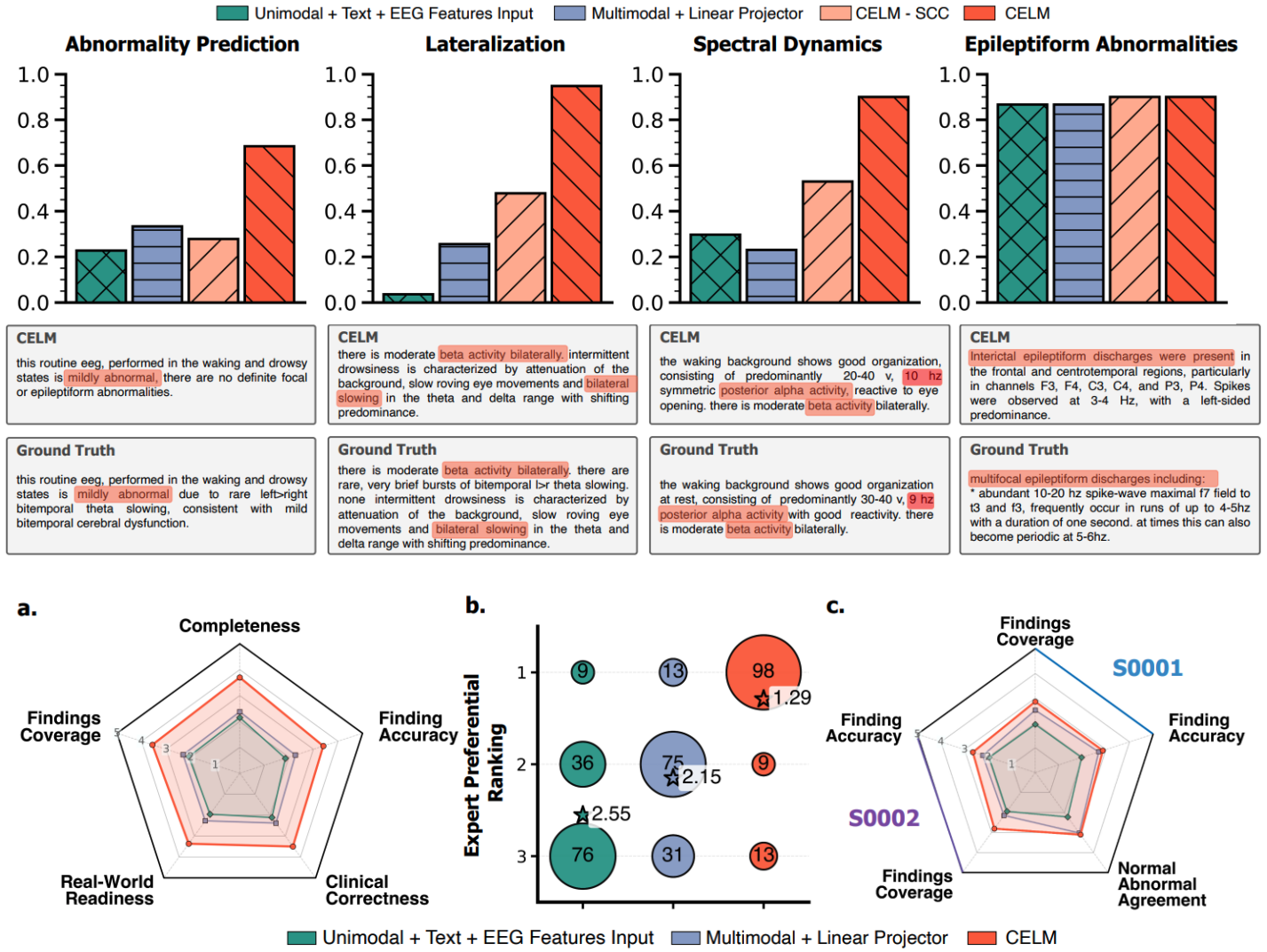


Figure 5: (a) Human expert evaluation study results, (b) Human expert preferential ranking for each report. (c) LLMs-as-a-judge results

NEURAL INFORMATION PROCESSING SYSTEMS

Neural Signals Generate Clinical Notes in the Wild

Paper

Code

Project Page

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

SANKEN THE UNIVERSITY OF OSAKA

https://www.jathurshanpradeepkumar.com/CELM/