EEG Foundation Models: A Critical Review of Current Progress and Future Directions

Gayal Kuruppu*
Department of Computer Science & Engineering
University of Minnesota, Twin Cities
Minneapolis, MN, USA
kurup016@umn.edu

Neeraj Wagh*
Department of Bioengineering
University of Illinois at Urbana-Champaign
Urbana, IL, USA
nwagh2@illinois.edu

Yogatheesan Varatharajah
Department of Computer Science & Engineering
University of Minnesota, Twin Cities
Minneapolis, MN, USA
yvaratha@umn.edu

Abstract

Electroencephalography (EEG) signals offer immense value in scientific and clinical investigations. In recent years, self-supervised EEG foundation models (EEG-FMs) have presented a viable path towards the robust and scalable extraction of EEG features. However, the real-world readiness of early EEG-FMs and the rubrics for long-term research progress remain unclear. This study conducts a critical review of ten early, first-generation EEG-FMs based on a) the representation of raw input data, b) self-supervised representation learning, and c) evaluation strategy. We synthesize key EEG-FM methodological trends, empirical findings, and remaining gaps. We find that EEG-FMs draw heavily from their counterparts in the language and vision domains for their model architecture and self-supervision. However, EEG-FM evaluations remain heterogeneous and largely limited, making it challenging to assess their practical off-the-shelf utility. In addition to adopting standardized and realistic evaluations, future efforts should demonstrate substantial scaling effects and make principled and trustworthy choices throughout the EEG-FM pipeline. We believe that the development of benchmarks, software tools, technical methodologies, and clinical/scientific applications in collaboration with domain experts may advance real-world adoption of EEG-FMs.

1 Introduction

Patterns of brain physiology embedded within electroencephalography (EEG) signals offer immense value to neuroscientists, biomedical engineers, and clinicians [1]. Despite decades of research into quantitative EEG features derived based on domain expertise, expert visual interpretation remains the gold standard for clinical EEG evaluation even to this day [2]. Nonetheless, the past decade has seen significant advances in deep learning-based approaches for extracting application-specific features from raw EEG data (EEG-DL) [3, 4]. However, EEG-DL models based on supervised learning met with limited success due to their reliance on costly annotations [5], which made them unscalable and prone to overfitting. This lack of robustness was further exacerbated by the variability of EEGs across sites, systems, subjects, and sessions, leading to an overall lack of trust in supervised EEG

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

encoders [6, 7]. These limitations emphasized a need for EEG-DL models that rely less on expert EEG labels and yield robust and trustworthy EEG features with high translational value.

The emerging paradigm of foundation models (FMs) [8], based on label-free self-supervised learning (SSL) and efficient transfer learning, is a promising solution for these data-related challenges. Similar to the mainstream vision [9–11] and language [12–14] FMs, EEG foundation models (EEG-FMs) are trained to identify salient EEG features from raw unlabeled EEG recordings by using various SSL pretext tasks. EEG-FMs learn to represent EEG data as compressed embeddings in a latent space by leveraging the intrinsic properties found in the raw data. Pretrained EEG-FMs can then be adapted for various downstream applications using only very small amounts of labeled data, thereby alleviating the burden of expert EEG annotations. As such, EEG-FMs hold promise as powerful, off-the-shelf feature extractors (or *encoders*) that can support scientific research, next-generation brain-computer interfaces, and augmented neurological decision support. Several first-generation EEG-FMs have been proposed over the last few years [15–24], whose cumulative count is shown in Figure 1.

Despite the growing interest, many questions remain unanswered regarding the design choices within EEG-FMs, the learned representations, the performance on various real-world applications, and the overall guarantees on robustness and trustworthiness. For example, the choice of EEG input representations, the architectural components, and the SSL pretext tasks can vary significantly between models, and the effects of those choices on the learned features are unclear. The complexity, quality, and flexibility of the representations learned by EEG-FMs and their relation to brain physiology have not been sufficiently studied. Furthermore, the performance and generalizability of these EEG-FMs beyond the common public datasets have not been adequately evaluated. These concerns call for a comprehensive review of the first-generation EEG-FMs focusing on the various architectural choices, pretraining approaches, evaluations, and trustworthiness aspects, to identify the rubrics for meaningful long-term progress in EEG-FM research and advance their translational value. The specific contributions of this review are as follows:

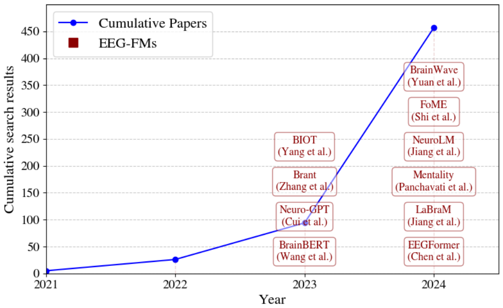


Figure 1: Cumulative EEG-FM search results from 2021 to September 30th, 2024.

1. We highlight key methodological trends and present empirical insights emerging from a systematic and comprehensive review of ten early, first-generation EEG-FMs.
2. We provide a data domain-centric critical analysis to identify outstanding research gaps that, if addressed, could increase the translational value and real-world impact of EEG-FMs.

2 Methods

We conducted a comprehensive search across various web platforms, including Google Scholar, arXiv, DBLP, IEEE Xplore, bioRxiv, and medRxiv, to identify relevant research in journals, conferences, workshops, and preprints. We limited our search query to **EEG "Foundation Model"**. We then removed duplicate instances and manually reviewed the title, abstract, and introduction sections to confirm relevant EEG-FMs by identifying phrases similar to *"We developed an EEG foundation model… "* and *"The proposed approach forms the basis for an EEG foundation model… "*. We note that our search starts from the year 2021 – the year the term *Foundation Model* was introduced [8] – and includes studies that were published or archived on or before September 30th, 2024. We identified nine EEG-FMs following this strategy, namely Neuro-GPT [16], Brant [17], BIOT [18], EEGFormer [19], LaBraM [20], Mentality [21], NeuroLM [22], FoME [23], and BrainWave [24]. We additionally included BrainBERT [15] in this set because of its common utilization as a baseline model in other EEG-FM evaluations [17, 19, 23, 24]. A detailed summary of these ten EEG-FMs is provided in the appendix (Table 1). We note that our review does not include FMs developed for polysomnography (PSG) data [25, 26], which are designed specifically for sleep analysis.

We compared the design and construction of EEG-FMs along three major axes: a) preparation and representation of input data, b) model architecture and self-supervised pretraining, and c) model evaluations. These considerations are illustrated in Figure 2.

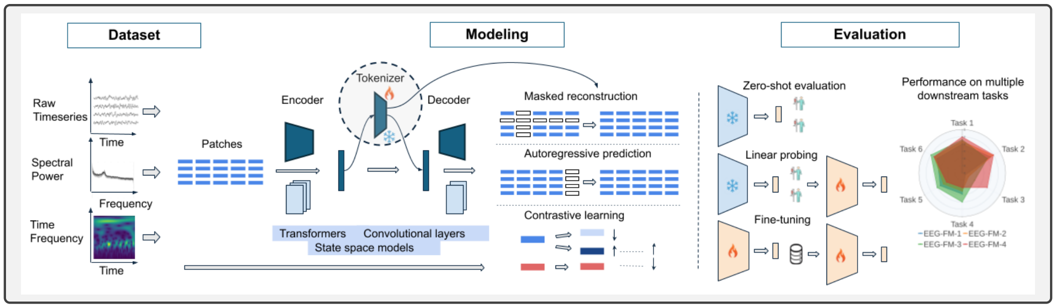


Figure 2: Comparative EEG-FM analysis along three major axes; input data, modeling, and evaluation.

3 Key Findings & Insights

General trends: EEG data were represented in one of three forms: raw time series, the magnitude power spectrum, and time-frequency representation. The FM architecture included convolutional blocks to learn low-level patterns and/or transformer blocks to learn higher-level relationships. The common SSL approaches to pretrain the FMs were either masked-reconstruction, auto-regressive modeling, or contrastive learning. The pretrained models were then adapted using expert labels and evaluated on various downstream tasks, including clinical (e.g., TUAB, TUEV [5]) and non-clinical tasks. Below we highlight the key insights gleaned from this review.

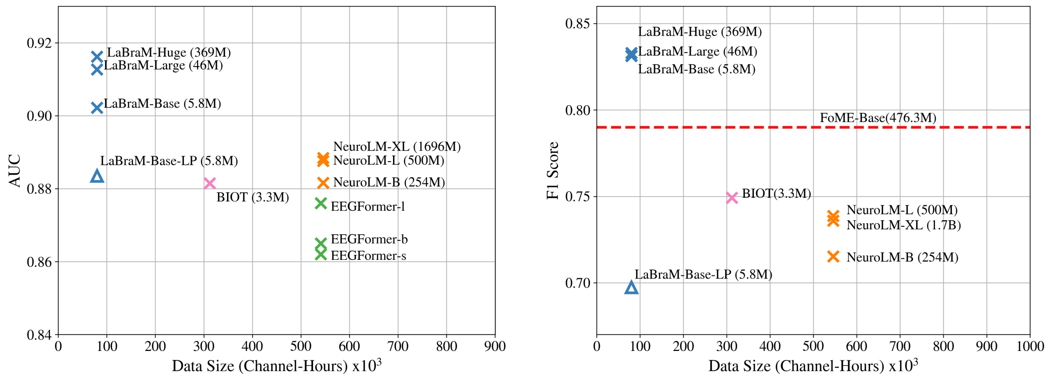
Diversity of pretraining data: Several EEG-FMs (LaBraM, NeuroLM, FoME, and BrainWave) leveraged a diverse set of EEG domains spanning clinical, sleep, and task-based BCI. LaBraM's lead in the TUAB and TUEV evaluations (Figure 3) may have emerged from higher diversity in pretraining data compared to EEGFormer and BIOT. The mixed use of scalp and intracranial EEG (iEEG), as done in FoME and BrainWave, can be considered another form of data diversity. Data ablations (scalp vs. iEEG) in BrainWave showed that joint pretraining (scalp + iEEG) boosted downstream task performance and transfer performance to unseen data types (electrocardiograms). Overall, the notion of data diversity may influence the performance, generalizability, and transferability of EEG-FMs.

Multivariate time series EEG representation: All but four EEG-FMs (BrainBERT, BIOT, EEGFormer, and BrainWave) utilized the native multivariate time series representation of EEG, while two models (BrainBERT, BrainWave) utilized spectral representations. Two other models (Brant, FoME) combined time-series and spectral representations as input, while the remaining two models (BIOT, EEGFormer) exclusively adopted time-frequency representations. Subsequently, the respective EEG inputs were positionally encoded with their spatial and temporal order.

Temporal sequence modeling: Sequence-based transformer blocks were the primary workhorse of representation learning in most EEG-FMs, except Mentality, which was based on a Mamba-based architecture. A few models (NeuroLM, Neuro-GPT) utilized convolutions to capture low-level morphological features of time-domain EEG. However, the modeling of spatial EEG relationships was either ignored or limited to the positional encoding step. BrainWave, a notable exception in this trend, integrated a spatial attention mechanism. Supporting such emphasis on temporal modeling, the ablation experiments in Brant showed that their temporal encoder provided the largest contribution to downstream task performance, compared to the spatial encoder and frequency encoding. However, the temporal context length of EEG-FMs did not exceed 90 seconds (FoME), and as such, they may struggle to capture long-range EEG patterns, relationships, or dependencies.

Pretraining using masked reconstruction: Reconstruction of masked temporal EEG sequences was the predominant EEG-FM pretraining paradigm, albeit with varying strategies for masking the sequences of tokens/patches. Despite its origins in vision and language domains, this SSL paradigm seemingly holds merit in the EEG domain. Several studies (LaBraM, NeuroLM, EEGFormer) employed a learned discrete neural codebook to facilitate the pretraining process.

Limited and incomparable evaluations: Task performance after fine-tuning was the primary paradigm of EEG-FM evaluation. However, in half the reviewed studies (BrainBERT, Brant, Mentality, NeuroLM, and FoME), the downstream evaluation datasets were already utilized for FM pretraining, i.e., the evaluations were in-sample. Although several studies conducted external evaluations on unseen datasets, BrainWave is the only study that performed truly out-of-distribution (OOD) task evaluations, i.e., without fine-tuning the model on the evaluation set. Direct model rankings beyond the TUAB and TUEV tasks (Figure 3) are difficult to determine due to heterogeneous selections of downstream tasks across most EEG-FMs. Overall, the universality and robustness of EEG-FMs have not been convincingly demonstrated in most studies, with BrainWave being a notable exception.



(a) TUAB: normal vs. abnormal classification.

(b) TUEV: six-class event classification.

Figure 3: EEG-FM performances on TUAB and TUEV classifications. The performance of FoME in 3b is shown using a line since the pretraining data size was not available.

Scaling and task performance: The trends in Figures 3a and 3b suggest that scaling up pretraining data may not necessarily improve downstream task performance, even with significant model scaling (e.g., LaBraM vs. EEGFormer/NeuroLM). Notably, LaBraM demonstrated that effects of pretraining data scaling on TUAB and TUEV classifications are sharpest under $\sim$1000 hours of data and begin to plateau thereafter. Overall, the evidence for data scaling is weak, if any, based on the limited shared tasks and models evaluated thus far. Effects of model scaling are reported in the appendix (Figure 5).

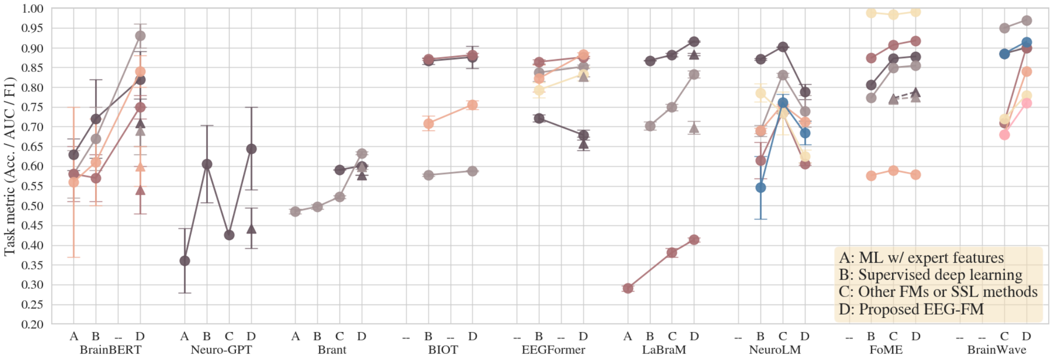


Figure 4: Impact of progressive learning paradigms (A → D) on downstream EEG classifications (◯ - fine-tuning, △ - linear probing). Tasks are not comparable across studies.

Advance over other feature paradigms: In Figure 4, we compare progressive feature extraction paradigms (expert features with classical ML, data-driven supervised DL features, self-supervised DL features, and proposed EEG-FMs) on various tasks. In a majority of tasks, fine-tuned EEG-FMs ('D') provided at least some improvement, if not drastic, over previous DL baselines and FMs ('B' and 'C', respectively). Linear probing results reported in EEG-FMs, however, were relatively worse in comparison. Fine-tuned EEG-FMs showed substantial improvements over classical ML models with expert features ('A'), although such assessments were presented only in four studies.

Performance of general-purpose time series models: Interestingly, we observed that general time series foundation models, such as TimesNet [27], performed reasonably well on several EEG tasks post-fine-tuning, and sometimes outperformed EEG-FMs (e.g., sleep-stage classification in FoME). Additionally, experiments in BrainWave show that, a time series FM – MOMENT [28] – outperformed EEG-FMs in specific tasks, such as seizure detection. Experiments in Brant show that general time series architectures, such as PatchTST [29] and CoST [30], perform relatively better in some tasks, such as short/long term signal forecasting and imputation, respectively, than some EEG-specific architectures. These findings highlight that certain EEG tasks could be tackled using general time series architectures or FMs without any EEG-specific inductive biases.

4 Research Gaps

Our review revealed several gaps in the existing literature that, if addressed, could significantly increase the real-world value of EEG-FMs. We summarize those gaps below.

Long temporal context and spatial modeling: Slow variations can exist in multi-day intracranial EEGs or multi-hour sleep-related EEG recordings [31]. However, current EEG-FMs can only process patterns within sequences of 90 seconds or less. There is a need for solutions that expand the effective context length of EEG-FMs. Moreover, the explicit modeling of spatial or inter-channel relationships and their contributions relative to temporal modeling requires further investigation.

Quality of pretraining strategy: EEG-FM linear probing evaluations reported by some studies have performed significantly worse than fine-tuned versions and other baselines in several instances (see Figure 4). However, some gains can be seen when FMs are fine-tuned on task data compared to fully-supervised training on the same data. This contrasting observation casts doubt on the inherent quality of the representations learned via self-supervision in EEG-FMs. Further investigations into the effects of SSL pretraining on downstream evaluations are needed to fully understand the extent of transferability achieved through SSL. Beyond the SSL strategy itself, the effects of data diversity, data volume, and model scale on the quality of EEG-FM pretraining remain unknown.

Data and model scaling: The scaling up of data and models is a defining principle of foundation modeling. However, empirical performance benefits of scaling in EEG-FMs have been either weak, limited, or inconclusive. Investigations that scale up data diversity, data volume (channel-hours), model size (trainable parameters), and evaluate on an expansive set of downstream tasks are lacking in current EEG-FMs, particularly at sufficiently large scales where effects are clear and discernible.

Practically relevant evaluations and metrics: Current fine-tuning evaluations are limited in their capacity to assess the real-world practical utility of EEG-FMs. There is a need for evaluation schemes and task metrics that capture the reality of EEG research and clinical use. Zero-shot evaluations are required for off-the-shelf EEG-FM usage for novel, unseen tasks. Few-shot or low-label performance can assess how efficiently EEG-FMs can leverage EEG labels that are typically expensive and laborious to collect. Out-of-distribution performance, especially without fine-tuning, on a known task can help understand EEG-FM robustness to the idiosyncratic cross-subject and cross-site variability of EEG. Finally, application-specific metrics, such as false positives per hour for seizure detection, and comparisons with expert EEG feature baselines can further clarify the real-world utility of EEG-FMs.

Standardized and challenging benchmarks: The tasks utilized by the EEG-FMs for evaluation were heterogeneous. The lack of shared evaluation tasks across multiple EEG-FMs makes it challenging to understand the state-of-the-art, and highlights a need to identify a common core set of evaluations for future EEG-FM evaluations. This set must cover multiple task types and include both classifications and regressions, with sparse (one label for the entire EEG recording) and dense (one label for each EEG segment) labels. Finally, the tasks must be challenging for previous generations of EEG-DL models with ample room for improvement, unlike TUAB, where performance may have already saturated (85-87% accuracy) with traditional approaches [32].

Trustworthy modeling: None of the reviewed EEG-FMs focused on model explainability or interpretability, which remain key requirements in scientific pursuits and for data-driven modeling in high-risk and expert-centric domains such as medicine. Studies that demystify the EEG-FM black box are needed to gain insight into the knowledge learned by EEG-FMs (EEG patterns, dependencies, relationships) through pretraining and the practical robustness of their decision-making process for downstream applications. Connections to known patterns of brain physiology or pathology may be necessary to make EEG-FMs trustworthy in the eyes of expert users.

5 Conclusion & Future Directions

The promise of foundation models lies in effective and robust feature learning, feature re-usability, and label efficiency. The first-generation EEG-FMs and those released more recently continue to make moderate strides in realizing this promise for the EEG domain. However, future EEG-FMs must prioritize substantial scaling efforts, principled and trustworthy self-supervised representation learning, and practically relevant evaluations. In addition to technical modeling, future research should also pursue the collaborative development of meaningful EEG benchmarks, novel clinical/non-clinical applications, and evaluation schemes that can measurably track the real-world readiness and impact of EEG-FMs. Encouragingly, recent EEG-FM efforts have emphasized data curation and preprocessing strategies [33], integration of disease- or task-specific constraints [33, 34], interpretability of learned EEG-FM latent codes [35], and the development of standardized, reproducible, and realistic benchmarks [36, 37]. The practical value of these efforts can be further enhanced by introducing model cards [38] that inform users of the functional design, strengths, and weaknesses of each EEG-FM. With sustained efforts in these directions, EEG-FMs are poised to advance scientific research, brain-computer interfaces, and clinical decision support systems.

6 Acknowledgments

We would like to thank Mario Serrafero and Saeid Cheshmi for fruitful discussions on foundation modeling.

7 Funding

This study was supported in part by the Mayo Clinic & Illinois Alliance Fellowship for Technology-based Healthcare Research, the Edward Heiken Interdisciplinary Health Sciences Institute Fund, and NSF grants IIS-2105233, IIS-2344731, and IIS-2337909.

References

[1] Faisal Mushtaq, Dominik Welke, Anne Gallagher, Yuri G. Pavlov, Layla Kouara, Jorge Bosch-Bayard, Jasper J. F. van den Bosch, Mahnaz Arvaneh, Amy R. Bland, Maximilien Chaumon, Cornelius Borck, Xun He, Steven J. Luck, Maro G. Machizawa, Cyril Pernet, Aina Puce, Sidney J. Segalowitz, Christine Rogers, Muhammad Awais, Claudio Babiloni, Neil W. Bailey, Sylvain Baillet, Robert C. A. Bendall, Daniel Brady, Maria L. Bringas-Vega, Niko A. Busch, Ana Calzada-Reyes, Armand Chatard, Peter E. Clayson, Michael X. Cohen, Jonathan Cole, Martin Constant, Alexandra Corneyllie, Damien Coyle, Damian Cruse, Ioannis Delis, Arnaud Delorme, Damien Fair, Tiago H. Falk, Matthias Gamer, Giorgio Ganis, Kilian Gloy, Samantha Gregory, Cameron D. Hassall, Katherine E. Hiley, Richard B. Ivry, Karim Jerbi, Michael Jenkins, Jakob Kaiser, Andreas Keil, Robert T. Knight, Silvia Kochen, Boris Kotchoubey, Olave E. Krigolson, Nicolas Langer, Heinrich R. Liesefeld, Sarah Lippé, Raquel E. London, Annmarie MacNamara, Scott Makeig, Welber Marinovic, Eduardo Martínez-Montes, Aleya A. Marzuki, Ryan K. Mathew, Christoph Michel, José d R. Millán, Mark Mon-Williams, Lilia Morales-Chacón, Richard Naar, Gustav Nilsonne, Guiomar Niso, Erika Nyhus, Robert Oostenveld, Katharina Paul, Walter Paulus, Daniela M. Pfabigan, Gilles Pourtois, Stefan Rampp, Manuel Rausch, Kay Robbins, Paolo M. Rossini, Manuela Ruzzoli, Barbara Schmidt, Magdalena Senderecka, Narayanan Srinivasan, Yannik Stegmann, Paul M. Thompson, Mitchell Valdes-Sosa, Melle J. W. van der Molen, Domenica Veniero, Edelyn Verona, Bradley Voytek, Dezhong Yao, Alan C. Evans, and Pedro Valdes-Sosa. One hundred years of EEG for brain and behaviour research. *Nature Human Behaviour*, 8(8):1437–1443, August 2024.

[2] Jordana Borges Camargo Diniz, Laís Silva Santana, Marianna Leite, João Lucas Silva Santana, Sarah Isabela Magalhães Costa, Luiz Henrique Martins Castro, and João Paulo Mota Telles. Advancing epilepsy diagnosis: A meta-analysis of artificial intelligence approaches for interictal epileptiform discharge detection. *Seizure: European Journal of Epilepsy*, 122:80–86, 2024.

[3] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: A systematic review. *Journal of Neural Engineering*, 16(5):051001, August 2019.

[4] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (EEG) classification tasks: A review. *Journal of Neural Engineering*, 16(3):031001, April 2019.

[5] Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016.

[6] Neeraj Wagh, Jionghao Wei, Samarth Rawal, Brent M Berry, and Yogatheesan Varatharajah. Evaluating latent space robustness and uncertainty of eeg-ml models under realistic distribution shifts. *Advances in Neural Information Processing Systems*, 35:21142–21156, 2022.

[7] Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical eeg signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021.

[8] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[9] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

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

[12] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[13] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[14] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[15] Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brainbert: Self-supervised representation learning for intracranial recordings. *arXiv preprint arXiv:2302.14367*, 2023.

[16] Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A Joshi, and Richard M Leahy. Neuro-gpt: Towards a foundation model for eeg. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

[17] Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36, 2024.

[19] Yuqi Chen, Kan Ren, Kaitao Song, Yansen Wang, Yifan Wang, Dongsheng Li, and Lili Qiu. Eegformer: Towards transferable and interpretable large-scale eeg foundation model. *arXiv preprint arXiv:2401.10278*, 2024.

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

[21] Saarang Panchavati and William Speier. Mentality. In *ICLR 2024 Workshop on Learning from Time Series For Health*, 2024.

[22] Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: A Universal Multi-task Foundation Model for Bridging the Gap between Language and EEG Signals, August 2024. arXiv:2409.00101 [cs, eess].

[23] Enze Shi, Kui Zhao, Qilong Yuan, Jiaqi Wang, Huawen Hu, Sigang Yu, and Shu Zhang. FoME: A Foundation Model for EEG using Adaptive Temporal-Lateral Attention Scaling, September 2024. arXiv:2409.12454 [cs, eess].

[24] Zhizhang Yuan, Fanqi Shen, Meng Li, Yuguo Yu, Chenhao Tan, and Yang Yang. Brainwave: A brain signal foundation model for clinical applications, 2024.

[25] Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore IV, Gauri Ganjoo, Emmanuel Mignot, and James Y Zou. Sleepfm: Multi-modal representation learning for sleep across ecg, eeg and respiratory signals. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

[26] Hans van Gorp, Merel M van Gilst, Pedro Fonseca, Fokke B van Meulen, Johannes P van Dijk, Sebastiaan Overeem, and Ruud JG van Sloun. A generative foundation model for five-class sleep staging with arbitrary sensor input. *arXiv preprint arXiv:2408.15253*, 2024.

[27] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

[28] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

[29] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

[30] Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. Cost: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. *arXiv preprint arXiv:2202.01575*, 2022.

[31] Filip Mivalt, Vaclav Kremen, Vladimir Sladky, Jie Cui, Nicholas M Gregg, Irena Balzekas, Victoria Marks, Erik K St Louis, Paul Croarkin, Brian Nils Lundstrom, et al. Impedance rhythms in human limbic system. *Journal of Neuroscience*, 43(39):6653–6666, 2023.

[32] Ann-Kathrin Kiessner, Robin T Schirrmeister, Joschka Boedecker, and Tonio Ball. Reaching the ceiling? empirical scaling behaviour for deep eeg pathology classification. *Computers in Biology and Medicine*, 178:108681, 2024.

[33] Dingkun Liu, Zhu Chen, Jingwei Luo, Shijie Lian, and Dongrui Wu. Mirepnet: A pipeline and foundation model for eeg-based motor imagery classification, 2025.

[34] Yihe Wang, Nan Huang, Nadia Mammone, Marco Cecchi, and Xiang Zhang. Lead: Large foundation model for eeg-based alzheimer's disease detection. *arXiv preprint arXiv:2502.01678*, 2025.

[35] Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. Codebrain: Bridging decoupled tokenizer and multi-scale architecture for eeg foundation model, 2025.

[36] Jiamin Wu, Zichen Ren, Junyu Wang, Pengyu Zhu, Yonghao Song, Mianxin Liu, Qihao Zheng, Lei Bai, Wanli Ouyang, and Chunfeng Song. Adabrain-bench: Benchmarking brain foundation models for brain-computer interface applications, 2025.

[37] Wei Xiong, Jiangtong Li, Jie Li, and Kun Zhu. Eeg-fm-bench: A comprehensive benchmark for the systematic evaluation of eeg foundation models. *arXiv preprint arXiv:2508.17742*, 2025.

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

[39] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[40] Dennis Gabor. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-part III: radio and communication engineering*, 93(26):429–441, 1946.

[41] Vasile V Moca, Harald Bârzan, Adriana Nagy-Dăbâcan, and Raul C Mureșan. Time-frequency super-resolution with superlets. *Nature communications*, 12(1):337, 2021.

[42] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[43] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[44] John C Mosher, Richard M Leahy, and Paul S Lewis. Eeg and meg: forward solutions for inverse methods. *IEEE Transactions on biomedical engineering*, 46(3):245–259, 2002.

[45] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[46] Karan Goel, Albert Gu, Chris Donahue, and Christopher Ré. It's raw! audio generation with state-space models. In *International conference on machine learning*, pages 7616–7633. PMLR, 2022.

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[48] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

A Technical Appendices and Supplementary Material

A.1 EEG-FM Summaries

Table 1 summarizes all ten EEG-FMs identified in our search. Although all the EEG-FMs share many commonalities, each FM is unique in its own right. In order to highlight the building blocks of each FM and their unique strengths, below we summarize each FM considering several factors, such as the amount of training data (in *channel-hours*, calculated as the total recording duration multiplied by the number of EEG channels), model size (in terms of the number of trainable parameters), the types of EEG data (scalp EEG and/or iEEG), the way inputs are configured (raw time series, power spectra, or time-frequency representation), architectural components (convolutional and/or transformer blocks), the SSL tasks used for pretraining (masked reconstruction, auto-regressive modeling, and/or contrastive learning), and the evaluations performed.

Table 1: Brief model summaries, including training data size, input configurations, data types, architectural components, and SSL tasks. Hyperlinks point to code and model weights, if available.

Model	**Training Data** (channel-hours)	**Number of Parameters**	**Input Configuration**	**Data Type**	**Architectural Components**	**SSL Tasks**
BrainBERT [link]	4.5k	43.18M	Single-channel spectrogram data	intracranial EEG	Transformer encoder and shallow decoder with two linear layers	Masked reconstruction
Neuro-GPT [link]	541k	79.53M	Fixed multi-channel time series data	Scalp EEG	Encoder with both convolution and transformer layers and GPT-2 as the decoder	Masked reconstruction (causally masked latent embeddings)
Brant [link]	281k	68M, 104M, 249M and 506M	Variable multi-channel time series data	intracranial EEG	Two transformer encoders for time and space and a linear decoder	Masked reconstruction
BIOT [link]	312k	3.3M	Variable multi-channel spectral data	Scalp EEG	Linear transformer, encoder-only architecture	Contrastive learning
EEGFormer	541k	N/A	Multi-channel spectral data	Scalp EEG	A transformer encoder and a shallow transformer decoder	Codebook-based reconstruction
LaBraM [link]	80k	5.8M, 46M and 369M	Fixed multi-channel time series data	Scalp EEG	Convolutional temporal encoder and transformer encoder layers and a linear decoder. A separate decoder for tokenization.	Masked reconstruction (token-level)
Mentality	N/A	N/A	Fixed multi-channel time series data	Scalp EEG	Convolutional layers and Mamba blocks in both encoder and decoder	Masked reconstruction
NeuroLM [link]	546k	250M, 500M and 1.7B	Variable multi-channel time series data	Scalp EEG	Vector quantization for tokenization with convolutional temporal encoder and transformer spatial encoder	Autoregressive reconstruction (token-level)
FoME	N/A	476M and 745M	Variable multi-channel time series data	Scalp EEG and intracranial EEG	Temporal and Spatial transformer encoder and a linear decoder	Masked signal reconstruction
BrainWave	878k	N/A	Variable multi-channel spectrogram data	Scalp EEG and intracranial EEG	Transformer encoder with channel attention and a lightweight decoder	Masked reconstruction (whole spectrogram)

BrainBERT [15]: As the first released EEG-FM, BrainBERT is relatively smaller than others, with 43.18M parameters and was trained using a modest dataset of 4.5k channel-hours of iEEG data. The inputs were represented as channel-wise spectrograms and a BERT[39]-type model was trained to predict masked patches for different types of spectrograms such as Short-Time Fourier Transform [40](STFT) and Superlets [41]. The model comprised a transformer encoder with a shallow decoder with two linear layers. Evaluations showed generalizability to unseen subjects and unseen electrode locations; however, the test data were from the same distribution as the training data. The evaluations also showed that performing linear probing on BrainBERT embeddings was as good as training supervised deep neural networks (DNN) from scratch for most of the evaluation tasks, which are focused on predicting brain-evoked responses for watching movies. Additionally, their evaluations showed that fine-tuning BrainBERT embeddings can reach DNN performance from scratch with as little as 15% of the training data for one of the tasks. A task-agnostic intrinsic dimensionality [42]

(ID)-based analysis showed that the BrainBERT embeddings of different brain regions had different IDs, compared to a relatively constant distribution across electrodes in randomly initialized weights.

Neuro-GPT [16]: This mid-size EEG-FM with 79.53M parameters was trained entirely using scalp EEG data from the full TUH corpus, including 541k channel-hours of clinical scalp EEG data. The model takes raw time series of the 22 EEG channels in the standard 10-20 layout as input and learns EEG representations using a combination of convolution and transformer layers. Those representations are then used as input to a GPT-2 [43] decoder which autoregressively predict the masked latents. It is noteworthy that the decoder has more parameters than the encoder in this setup compared which is not a common practice in other EEG-FMs. This model was evaluated only on EEG data from a BCI motor imagery task with four classes with a different channel configuration than the training data. However, the downstream data were transformed to the original input configuration using an inverse-forward approach[44]. The results showed that fine-tuning or linear probing the pretrained model performed better than models trained from scratch, including some EEG-specific fully-supervised DL approaches.

Brant [17]: Brant is a relatively larger FM with 500M parameters and was trained using 281k channel-hours of iEEG data. However, its pretraining data were limited to a single dataset comprising 9 subjects. This model also takes raw EEG time series as input and is able to take inputs with different channel configurations. The model consisted of a temporal encoder that learns long-term temporal dependencies, a spatial encoder that learns spatial correlations, and a simple linear decoder. The spatial encoder used in this model to capture spatial relationships is a novel contribution compared to previous EEG-FMs. This model also has three scaled-down versions with of 68M, 104M, and 249M parameters, respectively, that are trained on the same data. These models were evaluated on short/long term signal forecasting, frequency phase forecasting, imputation, and seizure detection tasks.

BIOT [18]: This is the smallest of the ten EEG-FMs considered in this review, with only 3.3M parameters, and was pretrained using a contrastive learning objective. The BIOT model introduced a novel approach to take input data with variable lengths and a variable number of channels; it tokenizes each channel into fixed-length segments representing frequency energy vectors, organizes them into “sentences", and uses channel and position embeddings to preserve spatio-temporal information. This model also utilized linear transformers to reduce training time. The model was then evaluated on multiple clinical tasks, such as seizure detection and seizure type classsification, in which the model showed superior performance even without pretraining, and showed even better performance with pretraining, compared to previous fully-supervised DL models.

EEGFormer [19]: This model was also pretrained on the whole TUH corpus, which includes approximately 541k channel-hours of clinical scalp EEG data. It included a transformer encoder and a shallow transformer decoder and it was pretrained using a masked-reconstruction objective. The encoder generated latents of input EEG patches are used to train a vector quantizer to match neural codes generated by a neural codebook. The decoder then reconstructs the original EEG patches using these neural codes. Evaluations included several downstream tasks derived from the TUH corpus and an out-of-distribution (OOD) evaluation on a neonatal seizure detection task. Additional experiments also included interpretability analyses using using the learned codebook.

LaBraM [20]: This model utilized the most diverse, yet a small dataset for pretraining, including 80k channel-hours of scalp EEG data – a subset of the TUH corpus. It was developed in three different scales, with 5.8M, 46M to 369M parameters, respectively. The model consisted of two parts, the neural tokenizer and the neural transformer (LaBraM pretraining model). Both took temporally and spatially patched raw time series data as input and then passed through convolutional temporal encoder of which the outputs are then concatenated with temporal and spatial embeddings and then passed through a set of transformer blocks. The neural tokenizer is trained to reconstruct the amplitude and phase of the patch through a separate decoder, during which the codebook is trained. The neural transformer is trained from scratch, and similar to the neural transformer, except to predict the neural codes inferred using the frozen neural tokenizer, by masked prediction. Although the evaluations demonstrated performance gains various subsets of the TUH corpus, it is unclear whether these results generalize to OOD data.

Mentality [21]: This model aims to capture the complex spatio-temporal dynamics of EEG signals using a Mamba [45]-based state-space model. The architecture of Mentality drew inspiration from other models such as SaShiMi [46], U-Net [47], and EEGNet [48] with the inclusion of Mamba blocks. However, the model was trained and evaluated exclusively on the TUSZ dataset. Furthermore, the unavailability of code or pretrained models limits reproducibility and further evaluations, and makes it inaccessible for broader use as a foundation model.

NeuroLM [22]: This model was inspired by a previous EEG-FM, LaBraM. However, NeuroLM was trained on 7x more data and is one of the largest EEG-FMs with 1.7B parameters, along with 250M and 500M parameter versions. NeuroLM utilized a text-aligned neural tokernizer, which is trained using temporal and frequency reconstruction along with a text/EEG domain classifier that is trained adversarially. The neural tokenizer is similar to that of LaBraM, except for the text alignment component and also included temporal reconstruction in addition to frequency reconstruction. However, despite this novel contribution, the evaluations indicated that the model loses some performance on downstream tasks compared to LaBraM and other state-of-the-art models.

FoME [23]: This model included two versions with 476M and 745M parameters, respectively. The size of the training data was not provided in the manuscript. The model takes masked time series patches and their power spectral densities as input, which are transformed by a temporal encoder. The outputs of the temporal encoder are then reorganized by channels and given as inputs to a spatial encoder to reconstruct masked time patches. FoME was evaluated on multiple downstream tasks, including classification, forecasting, and imputation; however, the evaluations were performed on in-distribution data, limiting any broader conclusions.

BrainWave [24]: This model was pretrained using a large dataset of size 878k channel-hours, including both scalp EEG and iEEG. It includes a transformer encoder and a channel attention module that transform EEG spectrograms into latent representations, which are then decoded by a lightweight decoder. This encoder-decoder architecture was trained using a masked-reconstruction objective. BrainWave is one of the few EEG-FMs trained and evaluated on both scalp EEG and iEEG signals, demonstrating the benefits of joint pretraining over unimodal approaches. The model has been extensively evaluated under different settings, such as cross-subject, cross-hospital, cross-subtype and few-shot classification, showcasing the generalizability and robustness across various clinical tasks.

A.2 Model scaling and downstream task performance

In Figure 5, we analyze the impact of model scaling on task performance using studies with at least two model variants. Each line plot indicates a specific downstream task, and the x-axis shows model variants, which were typically classified as small, intermediate, and large. Some marginal improvements can be observed for certain tasks and models, although these variants were developed with a fixed amount of pretraining data and were evaluated within study-specific experimental and methodological contexts. Notably, LaBraM investigated the combined effects of pretraining data and model scale on downstream TUAB/TUEV classifications. Overall, it is unclear whether a clear and strong trend exists with model scaling, especially within the current EEG-FM parameter regime ranging from 3.3M (BIOT) to 1.7B (NeuroLM, largest variant).

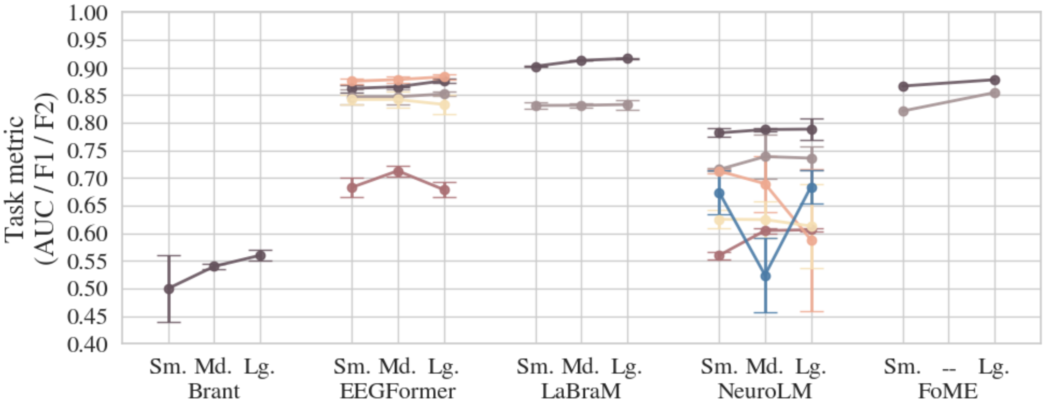


Figure 5: Model scaling and performance gains. Model sizes are specific to each study. *‘Sm.’* - smallest variant, *‘Md.’* - intermediate variant, *‘Lg.’* - largest variant.