Scaling EEG Foundation Models Improves Zero-Shot Transfer Across OmniEEG-Bench Task Families

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the scaling of EEG foundation model size impact zero-shot transfer accuracy across the six task families in OmniEEG-Bench compared to smaller domain-specific models. 16 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: OmniEEG-Bench: A Standardized Evaluation Benchmark for EEG Foundation Models. Research question: How does the scaling of EEG foundation model size impact zero-shot transfer accuracy across the six task families in OmniEEG-Bench compared to smaller domain-specific models?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.2/10.

3 Results

10 papers retrieved. 16 claims extracted; 1 independently verified. Quality review score: 5.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
OmniEEG-Bench includes 58 datasets for testing EEG foundation models.	✓	0.16
EEG foundation models pretrained on a greater number of datasets tend to achieve lower average ranks (better performance	×	0.09
EEG foundation models with a larger number of parameters tend to achieve lower average ranks (better performance) in lin	×	0.10
The per-dataset median Spearman correlation (ρ) between the number of pretrained datasets and model average rank is -0.2	×	0.03
The Wilcoxon signed-rank test p-value for the correlation between pretrained dataset count and model rank is 1.1e-07.	×	0.03
The per-dataset median Spearman correlation (ρ) between model parameter count and model average rank is -0.21.	×	0.05
The Wilcoxon signed-rank test p-value for the correlation between model parameter count and model rank is 7.0e-04.	×	0.04
The study implements four evaluation paradigms: cross-subject transfer, multi-subject adaptation, few-shot adaptation, a	×	0.06
54 EEG datasets were preprocessed using a standardized pipeline including downsampling, band-pass and notch filtering, c	×	0.05
Up to 40 samples per subject per class were randomly selected for linear probing based on variance stabilization analysi	×	0.03
Linear probing is used as the primary evaluation method, freezing the pretrained backbone and training only the classifi	×	0.05
Ten representative EEG foundation models were benchmarked: BENDR, BIOT, LaBraM, CBraMod, BrainOmni, FEMBA, Neuro-GPT, Ne	×	0.12
Models pretrained on multiple datasets outperform those pretrained on a single dataset overall in cross-subject linear-p	×	0.06
The Mann–Whitney U test p-value comparing single-dataset versus multi-dataset pretraining performance is 1.7e-04.	×	0.02
Transformer-family models show a per-dataset p-value of 4.8e-08 when compared to other model groups in cross-subject lin	×	0.04
Masked Reconstruction pretraining paradigms show a per-dataset p-value of 2.4e-02 when compared to other paradigms.	×	0.02

References

- http://arxiv.org/abs/2606.00815v1
- http://arxiv.org/abs/2506.13817v1
- http://arxiv.org/abs/2403.09832v1