Scaling of Bug Localization Accuracy in Contrastively Pretrained CodeT5 versus MLM on BigBench Debugging

Assignee Research

June 12, 2026

Abstract

Specialized foundation models are beginning to emerge in various medical subdomains, but pretraining methodologies and parametric scaling with the size of the pretraining dataset are rarely assessed systematically and in a like-for-like manner. This work focuses on foundation models for electrocardiography (ECG) data, one of the most widely captured physiological time series world-wide. We present a comprehensive assessment of pretraining methodologies, covering five different contrastive and non-contrastive self-supervised learning objectives for ECG foundation models, and investigate their s

1 Introduction

This paper examines: Pretraining Strategies and Scaling for ECG Foundation Models: A Systematic Study. Research question: How does the bug localization accuracy of contrastively pretrained CodeT5 scale with model size on the BigBench debugging subset compared to MLM-based pretraining?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.2/10.

3 Results

15 papers retrieved. 8 claims extracted; 6 independently verified. Quality review score: 7.2/10.

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
The study covers five different pretraining methodologies trained on over 11M samples.	×	0.12
State space models are confirmed as the superior architecture choice across all pretraining paradigms.	✓	0.18
CPC shows the strongest and most transferable representations across diverse clinical tasks.	×	0.14
Data2vec consistently lags behind across all evaluation modes and scaling regimes.	✓	0.19
Lower pretraining loss correlates with small residual errors in downstream tasks.	✓	0.20
The S4 backbone with model dimension 512 consistently outperforms larger and alternative configurations.	✓	0.26
The study investigates five self-supervised pretraining objectives spanning contrastive, predictive, and clustering-base	✓	0.19
The default backbone adopted is the 4-layer S4 with dimension 512.	✓	0.17

References

- http://arxiv.org/abs/2108.07435v2
- http://arxiv.org/abs/2605.12241v1
- http://arxiv.org/abs/2211.14875v3