Sequential Fine-Tuning Order Effects on CodeT5 Performance in MBPP Benchmark Tasks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the order of sequential fine-tuning (e.g., high-resource $\rightarrow$ low-resource vs. low-resource $\rightarrow$ high-resource) affect the downstream performance of CodeT5 on the MBPP benchmark across language. 18 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Transfer Learning Robustness in Multi-Class Categorization by Fine-Tuning Pre-Trained Contextualized Language Models. Research question: How does the order of sequential fine-tuning (e.g., high-resource $\rightarrow$ low-resource vs. low-resource $\rightarrow$ high-resource) affect the downstream performance of CodeT5 on the MBPP benchmark across language pairs?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

14 papers retrieved. 18 claims extracted; 2 independently verified. Quality review score: 4.5/10.

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
BERT consistently outperforms XLNet using identical hyperparameters on the entire range of class label quantities for ca	✓	0.39
BERT is more affordable than XLNet in terms of the computational cost (i.e., time and memory) required for training.	✓	0.29
All performance metrics on the test set decrease approximately linearly as the number of classes increases.	×	0.09
The linearly decreasing trend suggests that the probability of incorrectly categorizing an item is proportional to the n	×	0.04
There is a performance degradation rate of approximately 1% per additional class for all the models studied.	×	0.14
The macro-averaged recall should equal accuracy due to the use of balanced data for each class.	×	0.05
The precisions and F1 scores are similar to the accuracies.	×	0.03
Fitted lines result in coefficients of determination (R2) close to 1, indicating good fits in the range of interest.	×	0.01
The dispersion of the data points appears to increase with the number of classes.	×	0.04
Extrapolating performance would be less accurate with a higher number of classes.	×	0.05
Deviations from the fitted line could be due to a variety of factors, including underfitting, overfitting, violating the	×	0.04
Multi-label tendencies refer to an item reasonably having more than one label.	×	0.04
Some classes perform more poorly than others, dragging down the macro-averaged metrics below the fitted line, most evide	×	0.02
Binary classification involves two possible outcomes, whereas multinomial (i.e., multi-class) classification involves th	×	0.08
Such models are based on the assumption of independence of irrelevant alternatives (McFadden, Tye, and Train 1977), impl	×	0.02
In practice, this assumption can be violated when a newly introduced class label correlates with any of the existing lab	×	0.07
Many interesting classification problems can involve a plethora of class labels for categorization.	×	0.09
One would expect a classification model's performance to decrease as the number of classes increases.	×	0.08

References

- http://arxiv.org/abs/1909.03564v2
- http://arxiv.org/abs/2508.11281v3
- http://arxiv.org/abs/2402.04177v3