CodeT5 Scaling and Robustness to Identifier Renaming in MBPP Benchmarks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does scaling the model size of CodeT5 affect its robustness to identifier renaming perturbations on MBPP, measured by exact match scores across different model variants. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Scaling Law with Learning Rate Annealing. Research question: How does scaling the model size of CodeT5 affect its robustness to identifier renaming perturbations on MBPP, measured by exact match scores across different model variants?.

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 2.3/10.

3 Results

14 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 2.3/10.

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
Different warmup steps can result in different loss curves in training from scratch.	$\times$	0.07
High gradient norms are usually observed during the LR warmup stage, especially in the initial steps of training.	$\times$	0.06
We use standard experimental setups for LLM pre-training.	$\times$	0.03
The training dataset is Fineweb (Penedo et al., 2024) and the validation dataset is RedPajama-CC (Computer, 2023).	$\times$	0.04
We train a 594M non-embedding parameters LLAMA architecture-like model (Touvron et al., 2023) from scratch.	$\times$	0.06
We use AdamW optimizer (Loshchilov & Hutter, 2017) with β1 = 0.9 and β2 = 0.95.	$\times$	0.01
The weight decay is set as 0.1 and gradient clip is set as 1.0.	$\times$	0.02
We set maximal learning rate as 2 $\times$ 10-4 and batch size as 4M tokens.	$\times$	0.07
We adopt the decay factor or learning rate annealing λ = 0.999 in our all experiments.	$\times$	0.13
We minimize the Huber loss (Huber, 1964) between the predicted and the observed log loss values using the L-BFGS algorit	$\times$	0.03
We use the implementation of the minimize method provided by the scipy library.	$\times$	0.01
We set the δ value of Huber loss to 1.0$\times$10-3.	$\times$	0.03
We fit Eq. 1 on the loss curves under constant and cosine LRS with 20K total steps.	$\times$	0.07
We predict the full loss curves under several unseen LRS with 60K total steps.	$\times$	0.10
The results show an almost perfect fit, achieving a coefficient of determination (R2) greater than 0.999.	$\times$	0.04
Mean Prediction Error = 0.146%.	$\times$	0.02

References

- http://arxiv.org/abs/2403.09832v1
- http://arxiv.org/abs/2408.11029v2
- http://arxiv.org/abs/2212.02035v1