Fine-Tuning CodeT5 with Automated Domain-Specific Pairs in Low-Resource Languages

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: Does fine-tuning CodeT5 on automatically generated domain-specific pairs yield higher evaluation metric scores than manual annotation-based fine-tuning for low-resource language tasks. 17 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Automated Collection of Evaluation Dataset for Semantic Search in Low-Resource Domain Language. Research question: Does fine-tuning CodeT5 on automatically generated domain-specific pairs yield higher evaluation metric scores than manual annotation-based fine-tuning for low-resource language tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

11 papers retrieved. 17 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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
Combining multiple encoders with a generative LLM (GPT-4o) to reassess relevance scores increases inter-coder agreement	×	0.11
The approach of combining multiple encoders with GPT-4o improves the F1-score by 1.5 times.	×	0.04
Ensemble learning improves machine learning performance by combining predictions from multiple models, enhancing accurac	×	0.06
Standalone LLMs perform worse than human annotators in data annotation tasks.	×	0.05
In the implementation described, queries were generated using GPT-4o from randomly selected documents containing at leas	×	0.03
Dataset A contains 17,053 documents, 30 queries, and 2,739 verified retrieved candidates.	×	0.01
In Dataset A, the Combined method achieved a Krippendorff's alpha of 67.03.	×	0.01
In Dataset A, the Combined method achieved an F1-score of 53.42.	×	0.02
In Dataset A, the Ensemble method achieved a Krippendorff's alpha of 50.30.	×	0.05
Dataset B contains 14,065 documents, 30 queries, and 2,022 verified retrieved candidates.	×	0.01
In Dataset B, the Combined method achieved a Krippendorff's alpha of 68.69.	×	0.01
In Dataset B, the Combined method achieved an F1-score of 49.96.	×	0.02
Dataset C contains 129,345 documents, 30 queries, and 2,166 verified retrieved candidates.	×	0.01
The average relevance score for the Combined method across ratings 0-3 is 50.2.	×	0.04
The azure-text-embedding-3-large model achieved an nDCG@10 score of 69.	×	0.02
The sentence-transformers/multi-qa-mpnet-base-cos-v1 model achieved an average score of 35.29 across evaluated metrics.	×	0.02
The evaluation dataset used for model comparison consisted of 79.6K documents, 20 queries, and 406 relevant documents.	×	0.03

References

- http://arxiv.org/abs/2411.04573v1
- http://arxiv.org/abs/2412.10008v1
- http://arxiv.org/abs/2411.00727v2