The Integration Of Domain-Specific Terminology During Pretraining Performance On The Zero-Shot Performance Of Codet5 On

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the integration of domain-specific terminology during pretraining affect the zero-shot performance of CodeT5 on low-resource domain languages, as measured by BLEU score comparisons between. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. 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: How does the integration of domain-specific terminology during pretraining affect the zero-shot performance of CodeT5 on low-resource domain languages, as measured by BLEU score comparisons between models with and without domain-augmented pretraining?.

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 3.2/10.

3 Results

14 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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
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	×	0.05
Standalone LLMs perform worse than human annotators in data annotation tasks.	×	0.04
In the implementation described, queries were generated using GPT-4o from randomly selected documents containing at leas	×	0.02
The 'Combined' method achieved a Krippendorff's alpha of 67.03 on Dataset A, compared to 50.30 for the Ensemble method.	×	0.04
The 'Combined' method achieved an F1-score of 53.42 on Dataset A, which is higher than the Ensemble method's score of 33	×	0.04
On Dataset B, the 'Combined' method achieved a Krippendorff's alpha of 68.69.	×	0.02
The average relevance score for the 'Comb.' (Combined) method is 50.2, which is higher than the 'Ens.' (Ensemble) averag	×	0.04
The 'azure-text-embedding-3-large' model achieved an nDCG@10 score of 69.	×	0.03
The 'sentence-transformers/multi-qa-mpnet-base-cos-v1' model achieved an average score of 35.29 across evaluated metrics	×	0.02
The dataset used for evaluation contains 79.6K documents, 20 queries, and 406 relevant documents.	×	0.03

References

- http://arxiv.org/abs/2402.12317v2
- http://arxiv.org/abs/2412.10008v1
- http://arxiv.org/abs/2602.17425v1