Synthetic Claude 2 Data Training Enhances Small Language Model Robustness in Adversarial NLI

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: To what extent does training on synthetic data from Claude 2 improve the cross-domain robustness of small language models on adversarial NLI datasets compared to ChatGPT-3.5-Turbo. Natural Language Inference (NLI) remains an important benchmark task for LLMs. NLI datasets are a springboard for transfer learning to other semantic tasks, and NLI models are standard tools for identifying the faithfulness of model-generated text. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A synthetic data approach for domain generalization of NLI models. Research question: To what extent does training on synthetic data from Claude 2 improve the cross-domain robustness of small language models on adversarial NLI datasets compared to ChatGPT-3.5-Turbo?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

16 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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 MNLI dataset contains 392,000 training examples.	×	0.03
The ANLI dataset contains 162,000 training examples.	×	0.03
The WANLI dataset contains 102,000 training examples.	×	0.03
WANLI and GNLI utilized MNLI exemplars for data generation via few-shot or supervised learning.	×	0.03
The study trained T5 models in three sizes: small (60M parameters), large (770M parameters), and XXL (11B parameters).	×	0.06
For the WANLI dataset, which lacks a validation split, the MNLI validation data was used for hyper-parameter tuning.	×	0.04
Binary classifiers were created by converting neutral and contradiction labels into a single non-entailment label.	×	0.02
The TRUE benchmark consists of 11 evaluation datasets containing human annotations for factual consistency.	×	0.03
The TRUE benchmark includes datasets for abstractive summarization: FRANK, SummEval, MNBM, Wang et al. (2020), QAGS-CNND	×	0.06
The TRUE benchmark includes datasets for dialogue generation: BEGIN, Q2, and DialFact.	×	0.04
The TRUE benchmark includes datasets for fact verification: FEVER and VitaminC.	×	0.04
The TRUE benchmark includes the PAWS dataset for paraphrase detection.	×	0.03
The TRUE benchmark standardizes annotations to binary labels indicating whether text is factually consistent with ground	×	0.05
The total size of the GNLI dataset split described in the table is 684,929 examples.	×	0.03
The GNLI training split contains 670,739 examples with 237,325 entailment, 208,676 contradiction, and 224,738 neutral la	×	0.02
The GNLI dataset includes a human-annotated test set of 490 examples.	×	0.03

References

- http://arxiv.org/abs/2402.12368v2
- http://arxiv.org/abs/2410.02152v1
- http://arxiv.org/abs/2306.17555v2