Impact of Synthetic Data Generators on Alignment Robustness in Multimodal Models with Corrupted Tabular-Text Pairs

Assignee Research

June 11, 2026

Abstract

Large language models (LLMs) have been recently leveraged as training data generators for various natural language processing (NLP) tasks. While previous research has explored different approaches to training models using generated data, they generally rely on simple class-conditional prompts, which may limit the diversity of the generated data and inherit systematic biases of LLM. Thus, we investigate training data generation with diversely attributed prompts (e.g., specifying attributes like length and style), which have the potential to yield diverse and attributed generated data. Our inves

1 Introduction

This paper examines: Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias. Research question: What is the impact of synthetic data generators on the alignment robustness of multimodal models when processing corrupted tabular-text pairs?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.1/10.

3 Results

13 papers retrieved. 8 claims extracted; 8 independently verified. Quality review score: 8.1/10.

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
Previous research on training models using LLM-generated data generally relies on simple class-conditional prompts.	✓	0.27
Simple class-conditional prompts may limit the diversity of generated data.	✓	0.33
Simple class-conditional prompts may inherit systematic biases of Large Language Models.	✓	0.27
Attributed prompts outperform simple class-conditional prompts in terms of the resulting model's performance on datasets	✓	0.38
Synthetic datasets generated by simple prompts exhibit significant biases, such as regional bias.	✓	0.30
Attribute diversity plays a pivotal role in enhancing model performance.	✓	0.26
Attributed prompts achieve the performance of simple class-conditional prompts while utilizing only 5% of the querying c	✓	0.35
The data and code for this study are available at https://github.com/yueyu1030/AttrPrompt.	✓	0.18

References

- https://doi.org/10.48550/arxiv.2306.15895
- https://doi.org/10.48550/arxiv.2312.10997
- https://doi.org/10.1038/s41586-024-08328-6