Causal Data Augmentation Enhances Calibration Under Distribution Shift in Large Language Models

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: Does causal data augmentation improve the calibration metrics of large language models under distribution shift more effectively than non-causal synthetic data generation. 17 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Quaternion Generative Adversarial Networks. Research question: Does causal data augmentation improve the calibration metrics of large language models under distribution shift more effectively than non-causal synthetic data generation?.

2 Methodology

Systematic literature search across multiple databases yielded 12 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

12 papers retrieved. 17 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
The CelebA-HQ dataset contains 27,000 images for training and 3,000 images for testing.	×	0.03
The 102 Oxford Flowers dataset contains approximately 7,000 images for training and fewer than 1,000 images for testing.	×	0.03
Image samples from both datasets were reshaped to 128 $\times$ 128 resolution.	×	0.02
The Adam optimizer was used with a learning rate of 0.0002, beta1 equal to 0.0, and beta2 equal to 0.9.	×	0.01
Experiments were conducted with critic iterations set to 1 and 5.	×	0.02
The batch size was fixed to 64 for every experiment.	×	0.04
Training consisted of 100,000 iterations for CelebA-HQ and 50,000 iterations for 102 Oxford Flowers.	×	0.01
Adding a gradient penalty to SNGAN and QSNGAN did not result in any observed improvement.	×	0.03
The QSNGAN generator initial fully connected layer takes noise of size 128 as input and is composed of 4 $\times$ 4 $\times$ 1024 neur	×	0.02
The QSNGAN generator residual blocks stack 1024, 512, 256, 128, and 64 filters respectively.	×	0.02
The first residual block includes an upsampling module with a scale factor of 2.	×	0.02
The final output image is bounded to the interval [-1, 1] using a split Tanh function.	×	0.03
Quaternion convolutions in the generator use a kernel size of 3, stride of 1, and padding of 1.	×	0.02
The shortcut in the generator residual block uses a quaternion convolution with kernel size 1 and null padding.	×	0.02
The QSNGAN model has fewer than 10 million free parameters.	×	0.03
The real-valued SNGAN counterpart has 32 million parameters.	×	0.08
The QSNGAN checkpoint for inference saves more than 70% of disk memory compared to the real-valued counterpart.	×	0.05

References

- http://arxiv.org/abs/2411.15497v3
- http://arxiv.org/abs/2002.11318v5
- http://arxiv.org/abs/2104.09630v2