Qwen3 and Qwen2-1.5B Robustness to Adversarial Docstring Perturbations in HumanEval-X

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How do Qwen3 and Qwen2-1.5B differ in robustness against adversarial docstring perturbations across diverse programming languages in the HumanEval-X dataset. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Deep Dive into Adversarial Robustness in Zero-Shot Learning. Research question: How do Qwen3 and Qwen2-1.5B differ in robustness against adversarial docstring perturbations across diverse programming languages in the HumanEval-X dataset?.

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 3.5/10.

3 Results

12 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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 CUB dataset has 312 attributes, 200 classes, and 11788 images.	×	0.03
The SUN dataset has 102 attributes, 717 classes, and 14340 images.	×	0.03
The AWA2 dataset has 85 attributes, 50 classes, and 37322 images.	×	0.03
The standard per-class top-1 accuracy is used for ZSL evaluation.	×	0.05
For GZSL, per-class top-1 accuracy values for seen and unseen classes are used to compute harmonic-scores.	×	0.04
The reproduced values of ALE are denoted as original, although there are slight variations compared to the original resu	×	0.02
The model used is the label-embedding model [1], specifically the Attribute-label embedding (ALE) model.	×	0.09
The compatibility function F() is parametrized by learnable weights W.	×	0.00
ALE is one of the earlier studies that showed direct mapping by exploiting data and auxiliary information is more effect	×	0.04
The results of ALE will be representative of the adversarial robustness of this family of ZSL approaches.	×	0.11

References

- http://arxiv.org/abs/2208.00428v1
- http://arxiv.org/abs/2008.07651v1
- http://arxiv.org/abs/1009.0305v1