# Vendi-RAG Diversity-Weight Parameter Effects on FLAN-T5-xl Robustness in Adversarial QA

Assignee Research

May 30, 2026

**Abstract**

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does the diversity-weight parameter in Vendi-RAG influence the robustness of FLAN-T5-xl against adversarial attacks (e.g., ANLI) in knowledge-intensive QA, and what is the correlation between. Machine learning (ML) systems have introduced significant advances in various fields, due to the introduction of highly complex models. Despite their success, it has been shown multiple times that machine learning models are prone to imperceptible perturbations that can severely. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. 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 does the diversity-weight parameter in Vendi-RAG influence the robustness of FLAN-T5-xl against adversarial attacks (e.g., ANLI) in knowledge-intensive QA, and what is the correlation between retrieval diversity and robustness metrics?.

## 2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.0/10.

## 3 Results

11 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.0/10.

## 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. | × | 0.02 |
| The SUN dataset has 102 attributes, 717 classes, and 14340 images. | × | 0.02 |
| The AWA2 dataset has 85 attributes, 50 classes, and 37322 images. | × | 0.02 |
| The standard per-class top-1 accuracy is used for ZSL evaluation. | × | 0.04 |
| 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 | × | 0.05 |
| The ALE model is formulated as F(x, y; W) = $\theta$(x)W T $\varphi$(y), where $\theta$(x) is the visual and $\varphi$(y) is the class embeddings. | × | 0.02 |
| The ALE model is one of the earlier studies that showed direct mapping by exploiting data and auxiliary information is m | × | 0.02 |
| The ALE model is selected for its stability and competitiveness in modern benchmarks. | × | 0.03 |
| The ALE model is representative of the adversarial robustness of the family of ZSL approaches. | × | 0.10 |

## References


- http://arxiv.org/abs/2008.07651v1
- http://arxiv.org/abs/2502.11228v2
- http://arxiv.org/abs/2103.15670v3