SOVEREIGN: How does the relative Mahalanobis distance (RMD) method compare to ODIN and other post-hoc OOD detectors on LL

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Out-of-distribution (OOD) detection is a rapidly growing field due to new robustness and security requirements driven by an increased number of AI-based systems. Existing OOD textual detectors often rely on an anomaly score (e.g., Mahalanobis distance) computed on the embedding output of the last layer of the encoder. In this work, we observe that OOD detection performance varies greatly depending on the task and layer output. More importantly, we show that the usual choice (the last layer) is rarely the best one for OOD detection and that far better results could be achieved if the best layer

1 Introduction

Analysis of: Unsupervised Layer-wise Score Aggregation for Textual OOD Detection. Research goal: How does the relative Mahalanobis distance (RMD) method compare to ODIN and other post-hoc OOD detectors on LLM embedding spaces for near-distribution outlier detection in large-scale text classification benchmarks like GLUE or SuperGLUE?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

7 papers retrieved. 8 claims extracted, 8 verified. Tribunal: 7.2/10 → APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
Existing OOD textual detectors often rely on an anomaly score computed on the embedding output of the last layer of the	✓	0.33
OOD detection performance varies greatly depending on the task and layer output.	✓	0.32
The usual choice (the last layer) is rarely the best one for OOD detection.	✓	0.29
Far better results could be achieved if the best layer were picked.	✓	0.26
The proposed method is a data-driven, unsupervised method to combine layer-wise anomaly scores.	✓	0.30
The proposed post-aggregation methods achieve robust and consistent results while removing manual feature selection alto	✓	0.32
Their performance achieves near oracle's best layer performance.	✓	0.27
The paper extends classical textual OOD benchmarks by including classification tasks with a greater number of classes (u	✓	0.28

References

- https://doi.org/10.48550/arxiv.2308.01222
- https://doi.org/10.48550/arxiv.2302.09852
- https://doi.org/10.48550/arxiv.2310.19852