SOVEREIGN: How does the layer-wise score aggregation method generalize across different domains when evaluated on out-of-

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, Steffen Eger. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.

1 Introduction

Analysis of: MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. Research goal: How does the layer-wise score aggregation method generalize across different domains when evaluated on out-of-distribution samples from MNLI and QQP tasks?.

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

11 papers retrieved. 2 claims extracted, 2 verified. Tribunal: 8.2/10 $\rightarrow$ 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
MoverScore uses contextualized embeddings and Earth Mover Distance for text generation evaluation.	✓	0.30
MoverScore was presented at the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th Internat	✓	0.52

References

- https://doi.org/10.1016/j.inffus.2023.101805
- https://doi.org/10.18653/v1/d19-1053
- https://doi.org/10.1145/3639372