SOVEREIGN: How does COCO-DR's continuous contrastive pretraining on target corpora affect zero-shot accuracy on BEIR benc

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

In recent years, foundation models have become very popular due to their exceptional performance, mainly in natural language (NLP) tasks where they were first introduced. These models usually consist of hundreds of millions, or even billions, of parameters, making them resource-intensive during training and in production systems, leading to increased costs. This paper focuses on the reduction of a foundation's model size when applied to music information retrieval (MIR) tasks. Our research combines the Branchformer architecture with SummaryMixing, which were first applied in speech recognition

1 Introduction

Analysis of: Linear Complexity Self-Supervised Learning for Music Understanding with Random Quantizer. Research goal: How does COCO-DR's continuous contrastive pretraining on target corpora affect zero-shot accuracy on BEIR benchmark datasets compared to standard dense retrieval models like Contriever and DPR?.

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

3 papers retrieved. 9 claims extracted, 0 verified. Tribunal: 3.2/10 → REJECT (revision_round=0). Policy: ESCALATE_TO_OWNER.

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
The Music4All dataset contains around 910 hours of audio collected from online platforms and sanitized for high quality.	×	0.02
The FMA dataset is a dump of the free and open library Free Music Archive and contains audio alongside low & high-level	×	0.03
The FMA-large subset contains around 900 hours of audio.	×	0.02
The proprietary dataset is around 200k hours in size.	×	0.06
The proprietary dataset ensures high quality and contains a significantly larger amount of music samples.	×	0.03
The model performance is evaluated on music tagging, key detection, genre classification, emotion regression, instrument	×	0.05
The model uses SummaryMixing instead of multi-head self-attention module in the experiments.	×	0.12
The proprietary dataset size is comparable to other private datasets mentioned in the literature, such as the one in MER	×	0.10
The private dataset used is larger than 160k hours.	×	0.02

References

- https://arxiv.org/abs/2601.09603

- https://www.semanticscholar.org/paper/4391bd955c6f31cccf2072c72eb0038487626943
- https://www.semanticscholar.org/paper/35a1306a965d07fdef271017b55627d021ccf3c9