

# Quantized RWKV Layers: Embedding Quality and Inference Speed Trade-offs Versus Full-Precision Models

Assignee Research

June 12, 2026

## Abstract

This paper investigates the efficacy of RWKV, a novel language model architecture known for its linear attention mechanism, for generating sentence embeddings in a zero-shot setting. I conduct a layer-wise analysis to evaluate the semantic similarity captured by embeddings from different hidden layers of a pre-trained RWKV model. The performance is assessed on the Microsoft Research Paraphrase Corpus (MRPC) dataset using Spearman correlation and compared against a GloVe-based baseline. My results indicate that while RWKV embeddings capture some semantic relatedness, they underperform compared

## 1 Introduction

This paper examines: Exploring RWKV for Sentence Embeddings: Layer-wise Analysis and Baseline Comparison for Semantic Similarity. Research question: What is the impact of quantizing RWKV layers on the trade-off between embedding quality (measured by Spearman correlation on MRPC) and inference speed (measured in tokens/second) compared to full-precision models?.

## 2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.2/10.

## 3 Results

10 papers retrieved. 13 claims extracted; 11 independently verified. Quality review score: 7.2/10.

## 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
Spearman correlation was used as the primary evaluation metric to quantify the monotonic relationship between cosine sim	✓	0.33
Inference time was measured as the average time taken to process a sentence pair for both RWKV and the GloVe baseline.	✓	0.25
Peak GPU memory usage was recorded during embedding generation to assess resource consumption.	✓	0.22
Experiments were conducted on a Google Colab environment equipped with a Tesla T4 GPU.	✓	0.18
The study utilized the pre-trained RWKV-v6-Finch-1B6-HF model loaded via Hugging Face Transformers.	✓	0.24
The study utilized GloVe 6B 50d embeddings as a baseline.	×	0.11
Sentence embeddings were generated from RWKV hidden layers 1, 3, 5, 7, 9, and 11.	×	0.14
Embeddings were generated for a subset of 1000 samples from the MRPC training set and the full 408 samples from the MRPC	✓	0.18
Sentence embeddings were computed by averaging the hidden states across all tokens in the sentence (average pooling).	✓	0.20
Cosine similarity was calculated for each sentence pair’s embeddings.	✓	0.18
Spearman correlation was computed between similarity scores and MRPC labels using the SciPy library.	✓	0.25
Inference time and GPU memory usage were recorded using PyTorch utilities.	✓	0.22
Table 1 presents the Spearman correlation coefficients for RWKV layers and the GloVe baseline on the MRPC dataset.	✓	0.26

## References

- <http://arxiv.org/abs/2411.10958v7>
- <http://arxiv.org/abs/2301.12017v2>

- <http://arxiv.org/abs/2502.14620v1>