A COMPARATIVE ANALYSIS OF PERFORMANCE AND SEMANTIC ACCURACY IN TEXT EMBEDDING ARCHITECTURES VIA C# AND ONNX

Pavel Kyurkchiev, Anton Iliev, Asen Rahnev, Viktor Matanski

Abstract. *The rapid adoption of Retrieval-Augmented Generation (RAG) has made text embeddings central to modern software. However, most AI benchmarks focus exclusively on Python, leaving a critical gap for enterprise .NET developers who require local, offline inference for data privacy and cost-efficiency. This paper presents a comprehensive comparative analysis of nine open-source text embedding architectures executed directly in C# using ONNX Runtime. We evaluate models across computational performance (CPU latency) and semantic accuracy. To rigorously test accuracy, we introduce a custom dataset of 25 "lexical traps"–queries and targets that exhibit high keyword overlap but possess opposing meanings. Experimental results demonstrate that INT8 quantization (e.g., all-MiniLM-L6-v2) reduces latency by 36% (5.24 ms to 3.36 ms) with negligible accuracy degradation. For systems requiring deep semantic comprehension, the prompt-instructed* `e5-small-v2` *emerges as optimal, successfully avoiding 48% of traps and maintaining a positive semantic margin at a 10.72 ms latency. Conversely, while heavy architectures like* `MPNet-Base` *matched this accuracy, they exhibited severe diminishing returns, increasing latency by nearly 200% (31.24 ms). These findings highlight the superiority of asymmetric prompting over raw parameter count for robust offline NLP in .NET applications.*

Key words: Text Embeddings, ONNX Runtime, C#, .NET, Semantic Search, Quantization, RAG

Acknowledgments

This study is financed by the project No FP25-FMI-010 "Innovative Interdisciplinary Research in Informatics, Mathematics, and Pedagogy of Education" of the Scientific Fund of the Paisii Hilendarski University of Plovdiv, Bulgaria.

Pavel Kyurkchiev[1,*], Anton Iliev[1], Asen Rahnev[1], Viktor Matanski[1]
[1] Paisii Hilendarski University of Plovdiv,
Faculty of Mathematics and Informatics,
236 Bulgaria Blvd., 4027 Plovdiv, Bulgaria
Corresponding author: `pkyurkchiev@uni-plovdiv.bg`