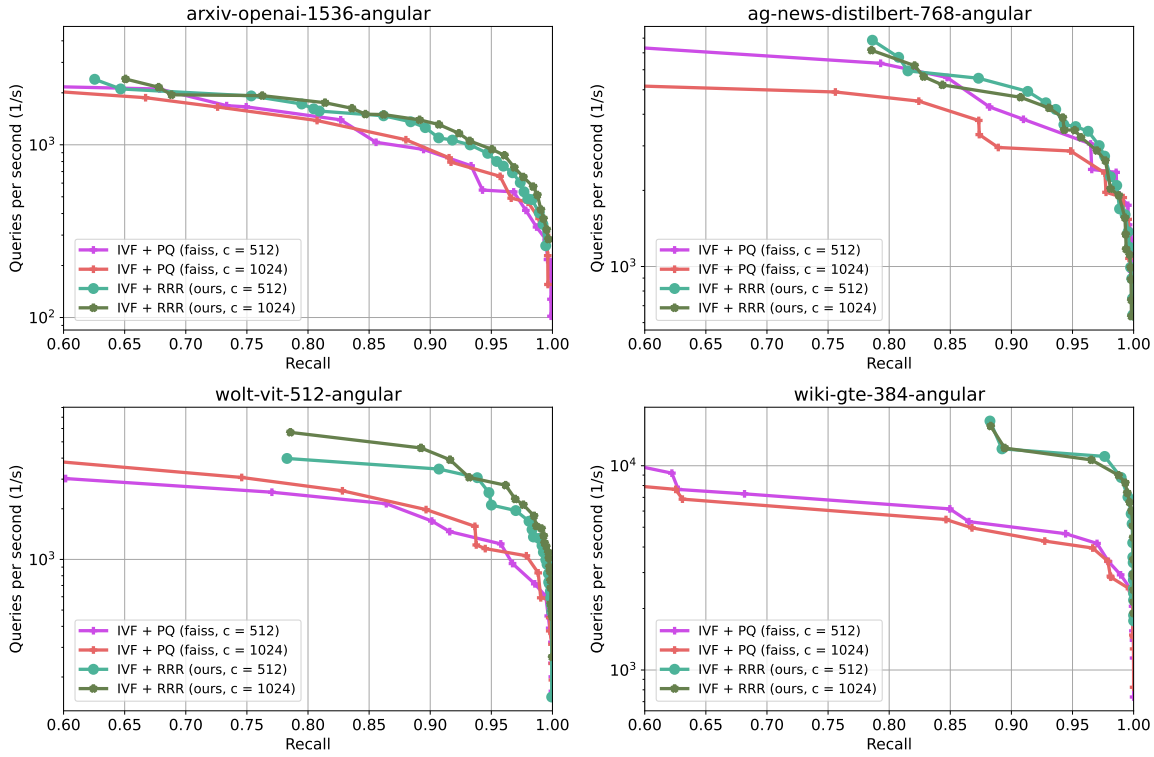

Figure 1: Performance comparison of IVF+reduced-rank regression (our supervised score computation method) and IVF+product quantization (Faiss implementation of IVF-PQ). The k-means clustering is kept constant to directly compare the effect of the score computation method (here c denotes the number of clusters). To directly compare the score computation methods, we do not use global dimensionality reduction. However, we use 8-bit integer quantization to implement reduced-rank regression, since Faiss also uses 4-bit integer quantization for its PQ implementation. The proposed score computation method outperforms the baseline method (PQ) on all of the data sets.

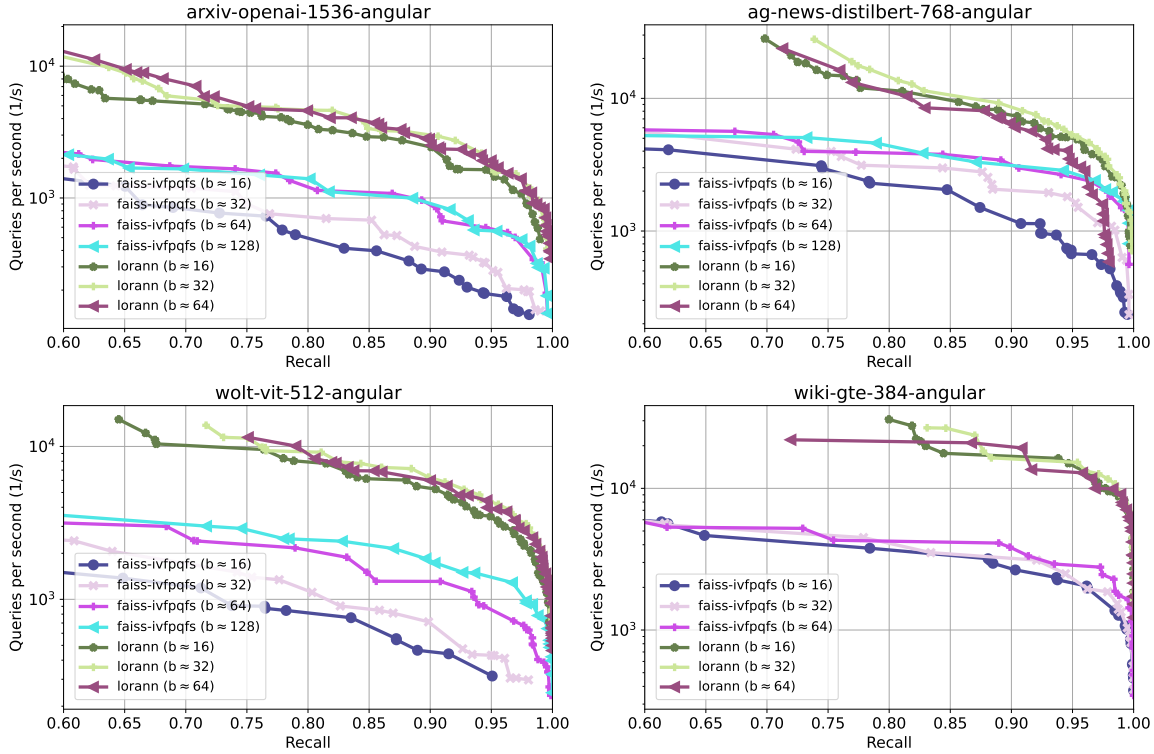

Figure 2: Performance comparison of LoRANN and IVF-PQ (Faiss) for different memory consumptions. We vary the rank parameter r for LoRANN and the code size for PQ such that b, bytes per vector, is similar for both. For fixed memory consumption, LoRANN has better performance compared to PQ that is a typical choice in memory-limited use cases.