Quantization Beyond Uniform Bit Allocation

K. S. Sreeramji[1,*,†], Sabyasachi Basu[2], Ravishankar Krishnaswamy[2], Kirankumar Shiragur[2], Yujia Wang[3]

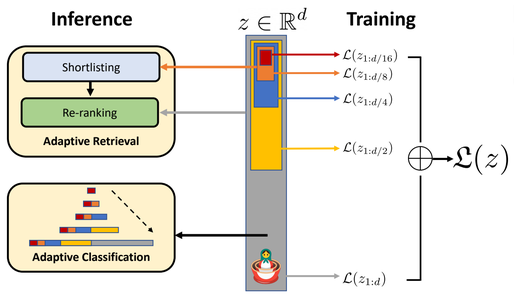
[1]Indian Institute of Science [2]Microsoft Research [3]Microsoft STCA

Motivation

- **Memory Bottleneck:** Modern embedding models produce thousands of dimensions ($D = 3072$). Storing billions of vectors in fast memory is prohibitive ($\approx 12\,\text{KB}$ / vector).
- **Capitalizing on Embedding Structure:** Modern models exhibit rich geometric structure (Matryoshka representation), yet conventional quantization treats all dimensions equally with uniform bit budgets.

"Can a variable bit allocation scheme allow for better quantization of vectors at the same compression rate?"

MRL and Variance Decay



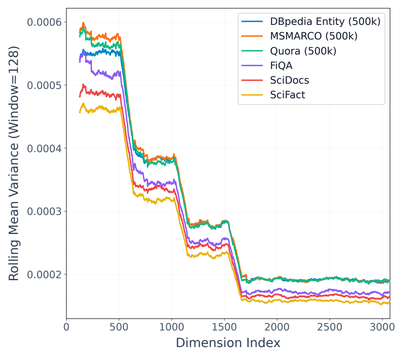
[Fig. credit: Kusupati et al. [1]]

Matryoshka Representation Learning (MRL) [1]:

- Nested loss functions enforce high semantic density in prefix dimensions.
- Enables multi-granular retrieval and adaptive truncation.

Variance Plateaus:

- Coordinate variance decays in sharp step-like tiers across 3072 dimensions.
- Empirical proxy for dimensional signal importance.



OpenAI text-embedding-3-large, $D = 3072$

Core Insight: Uniform quantization starves high-variance leading coordinates while wasting bits on near-constant trailing ones.

Variable Product Quantization

- **Bucket Decomposition:** Partition $D = 3072$ dimensions into $K = 8$ contiguous buckets $B_1, \ldots, B_8$.
- **Variable Subvector Density:** Allocate b_k subvectors to each bucket B_k (1 byte / subvector), concentrating density in high-variance buckets.
- **Zero Query Overhead:** Apply a single offline allocation vector $\vec{b} = (b_1, \ldots, b_K)$ globally, maintaining standard PQ query efficiency.

Bit Allocation & Empirical Recall Wins

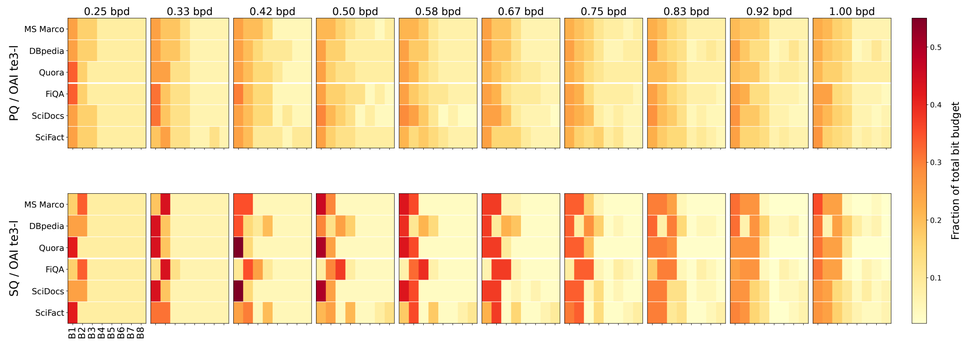


Figure 1: Discovered bit allocations across 8 buckets for OpenAI `text-embedding-3-large` (0.25 – 1.00 bpd) on 6 BEIR benchmarks. Bits heavily concentrate in prefix buckets $B_1 - B_3$.

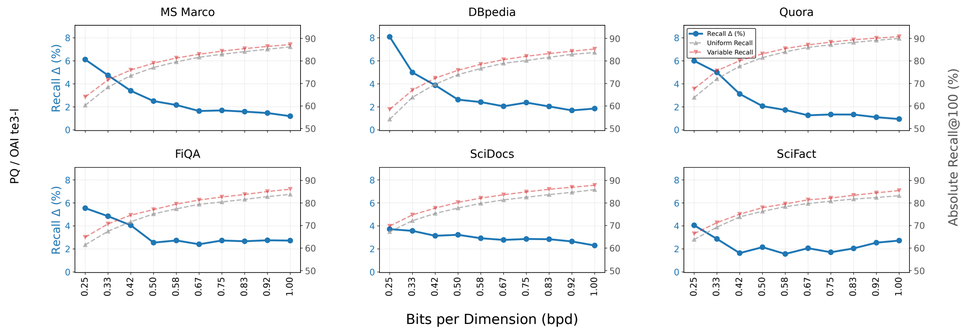


Figure 2: Recall improvements (Δ) and absolute Recall@100 of Variable PQ over Uniform PQ across 6 BEIR benchmarks (OpenAI `text-embedding-3-large`).

- **Strict Dominance:** Variable allocation strictly outperforms uniform PQ at identical memory footprints.
- **Low-Bitrate Advantage:** Relative recall gains are largest in highly constrained compression regimes (≤ 1.00 bpd).

Greedy Bit Allocation Search

1. Start Uniform
Initialize baseline vector $\vec{b}^{(0)} = (b_{\text{init}}, \ldots, b_{\text{init}})$.

2. Probe Candidates
Test adding δ bytes to each bucket $k \in \{1, \ldots, K\}$ to compute recall gain.

3. Greedily Assign
Update $b_{k^*} \leftarrow b_{k^*} + \delta$ for the bucket k^* yielding maximum recall improvement.

Repeat until memory budget is met

Greedy search serves purely as a proof of concept to demonstrate that non-uniform bit allocation strictly dominates uniform baselines.

Open Directions

- **Closed-Form Allocation:** Derive analytical formulas for optimal allocation directly from embedding properties.
- **Structure-Aware Quantizers:** Design algorithms where non-uniform allocation emerges natively, bypassing expensive heuristics.
- **Systems Challenges:** Address memory-access and SIMD alignment bottlenecks caused by non-uniform data types.

Key Takeaways:

- **Free Recall Gains:** Variable allocation improves recall at zero extra memory cost.
- **MRL Alignment:** Matryoshka models naturally benefit from variable quantization.

References:

[1] A. Kusupati, G. Bhatt, A. Ananyeva, et al. *Matryoshka Representation Learning*. NeurIPS 2022.

*Work done during internship at Microsoft Research. †Supported by the Walmart Center for Tech Excellence at IISc.

sreeramjiks@iisc.ac.in