

VQ-bench A Composable Vector Quantization Framework

Ashwin Padaki^{1,2} • Amir Ingber¹ • Edo Liberty¹

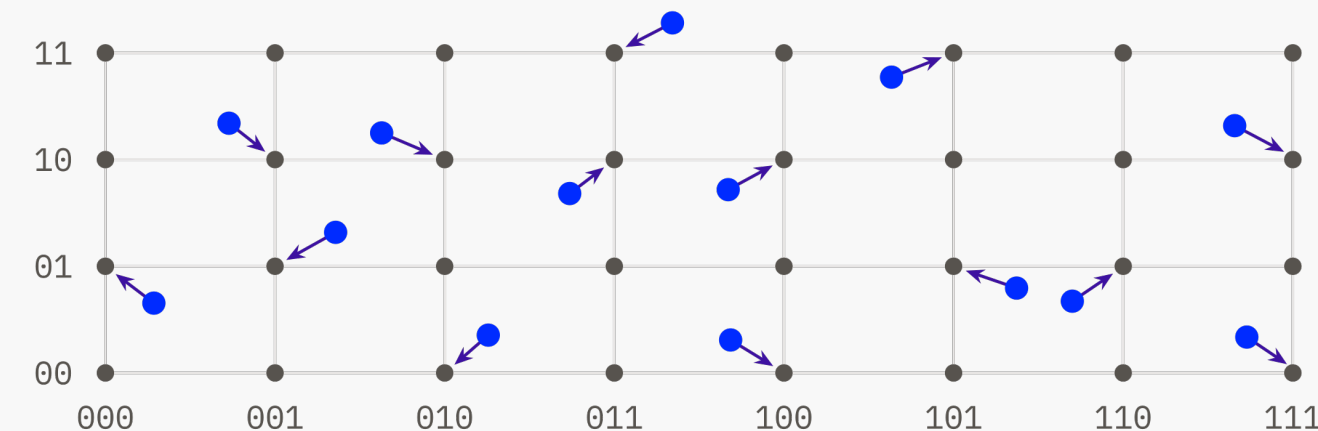
¹Pinecone • ²University of Pennsylvania



VecDB @ VLDB 2026

Quantization is crucial for VecDBs / LLMs

- Quantizers compress vectors into short codes



(a simple grid quantizer)

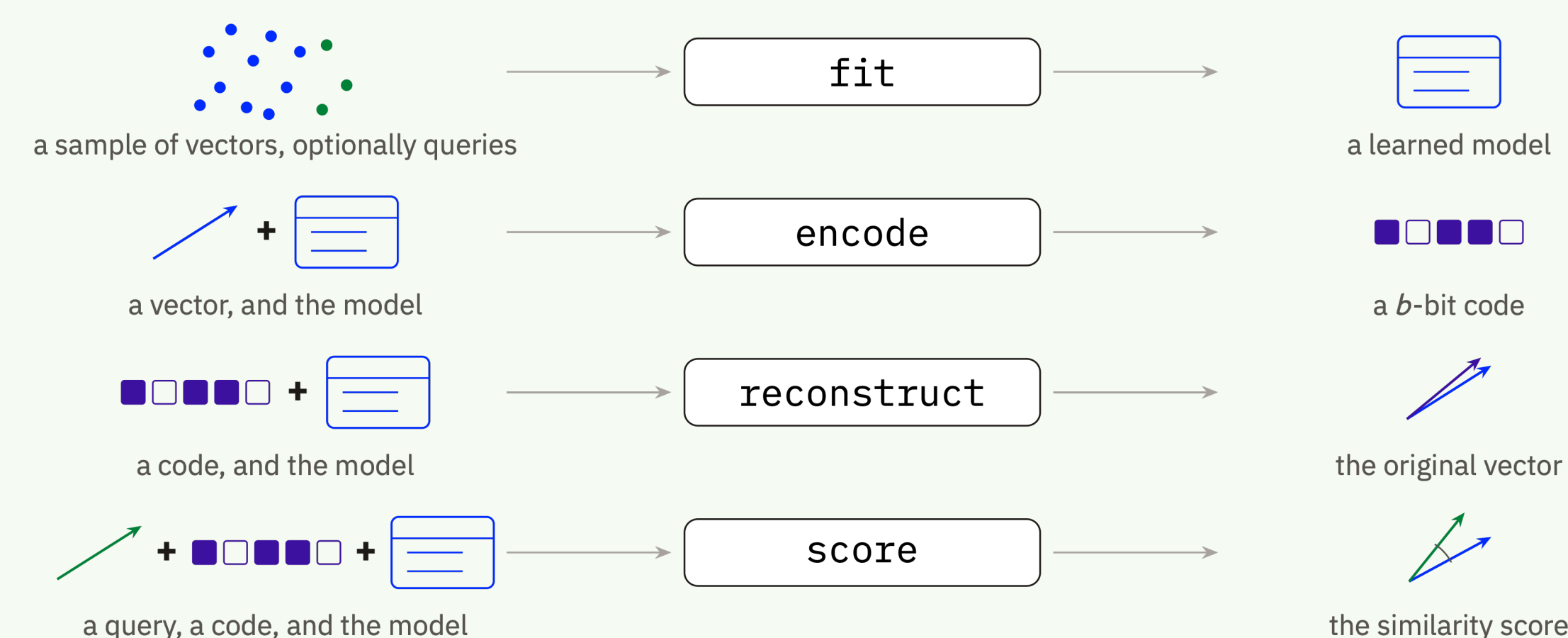
There are many good quantizers, but no good way to compare them

- Dozens of quantizers are published each year
- Better algorithm or better engineering?
- Papers vary in datasets, metrics, hardware

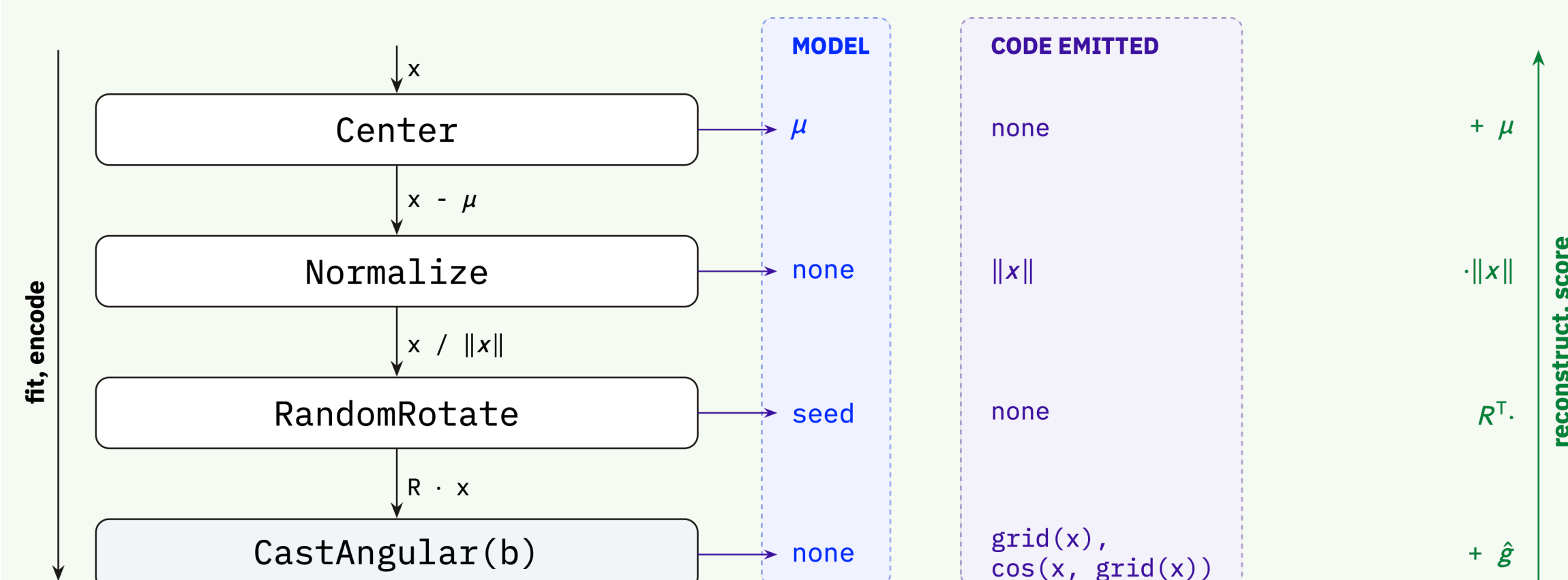
VQ-bench offers a solution

- A compositional framework for quantization
- An open-source library for quantizers and their underlying primitives
- A public benchmark at vq-bench.com

Quantizers encode vectors and recover info



Most quantizers decompose into primitive stages



We implement several primitives (in Rust)

Conditioners transform the data and pass it forward

```
center normalize minmax minmax_dim absmax
scale resize random_rotate random_hadamard
pca_rotate balance_parts optimize_signs optimize_pq
```

Rounders cast to a codebook and pass the residual

```
cast_uint cast_normal cast_angular
cast_sign cast_hamming kmeans
```

Splitters open a pipeline for each part of the dataset segment

Now, coding a quantizer is easy

SimHash	<code>center.normalize.random_hadamard.cast_hamming</code>
E-RaBitQ	<code>center.normalize.random_hadamard.cast_angular</code>
TurboQuant	<code>center.normalize.random_hadamard.cast_normal</code>
EDEN	<code>center.normalize.random_hadamard.cast_normal(MSE)</code>
PQ	<code>segment.kmeans</code>
OPQ	<code>center.pca_rotate.balance_parts.optimize_pq.PQ</code>

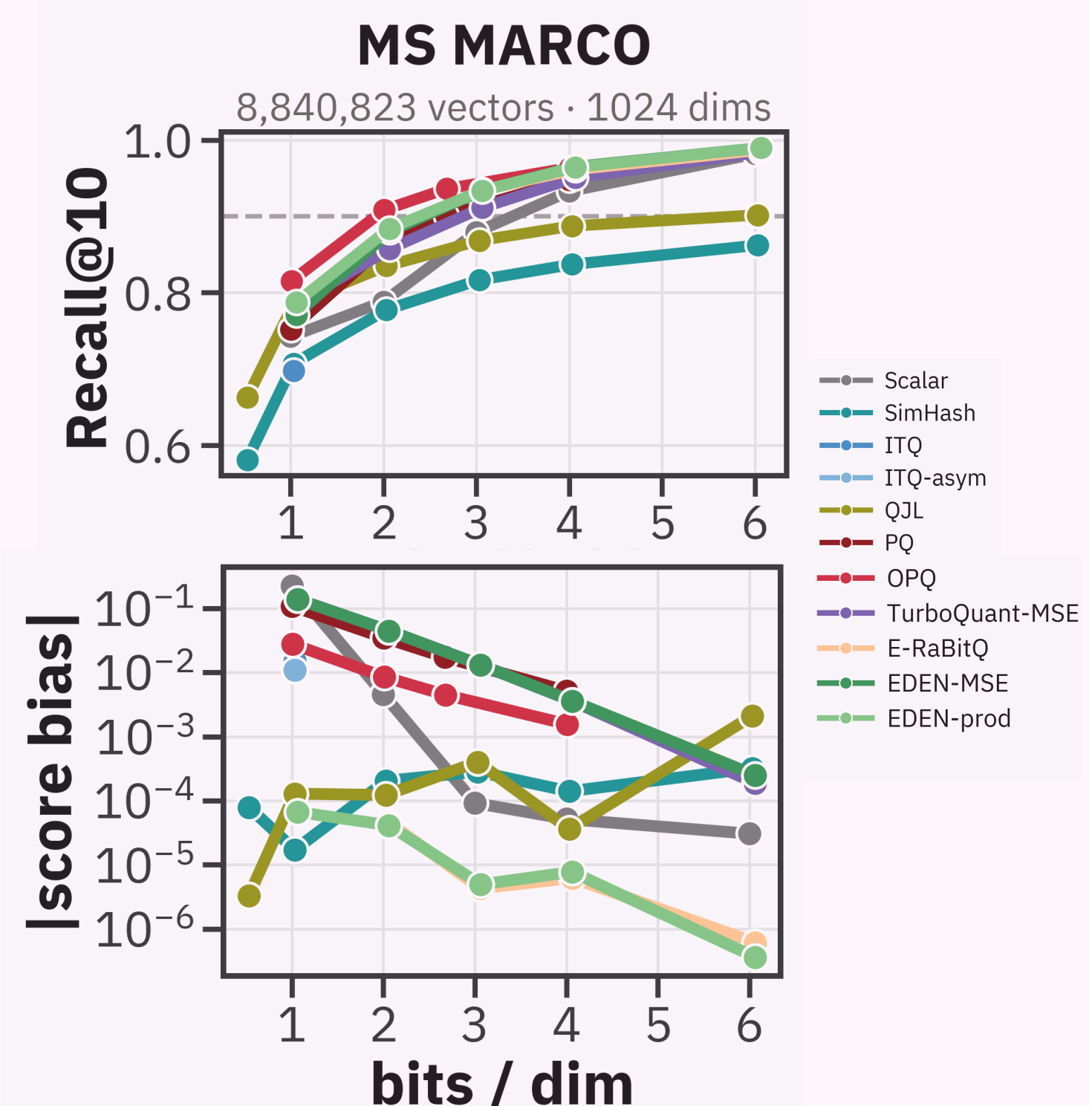
Also, TurboQuant-prod, EDEN-prod, ITQ, QJL, MinMax, Scalar etc.

It's all open source - help us add more quantizers!

We publish a benchmark

10+ datasets from VIBE, and we measure:

- Compression quality (recall, score bias, etc.)
- Efficiency (bits per dim, encode time, etc.)



Full benchmark and guidelines for contributing: vq-bench.com

