

VQ-bench

A **Composable** Vector Quantization Framework

Ashwin Padaki

UPenn + Pinecone

Amir Ingber

Pinecone

Edo Liberty

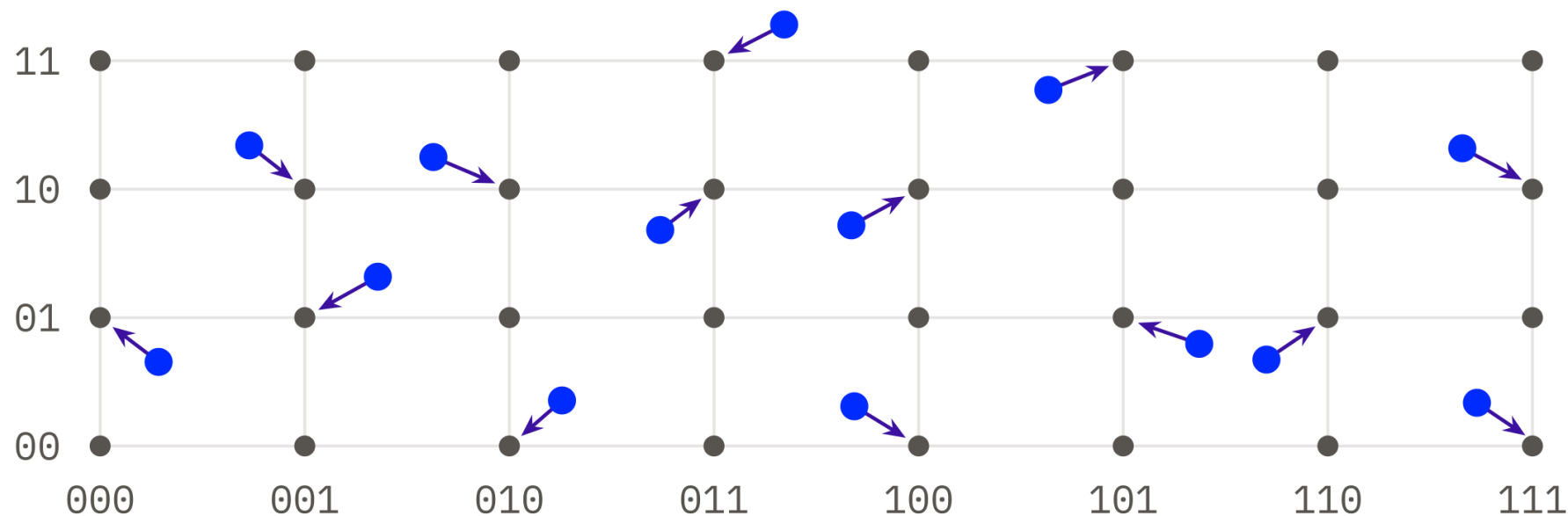
Pinecone

Quantization is crucial for vector databases

Storing vectors is expensive



Quantizers compress vectors into short codes



SimHash

PQ

ITQ

EDEN

QUIP

QUIP#

OPQ

QuaRot

TurboQuant

What is the **best** vector quantizer?

SpinQuant

AdaRound

GPTQ

RaBitQ

KIVI

HQQ

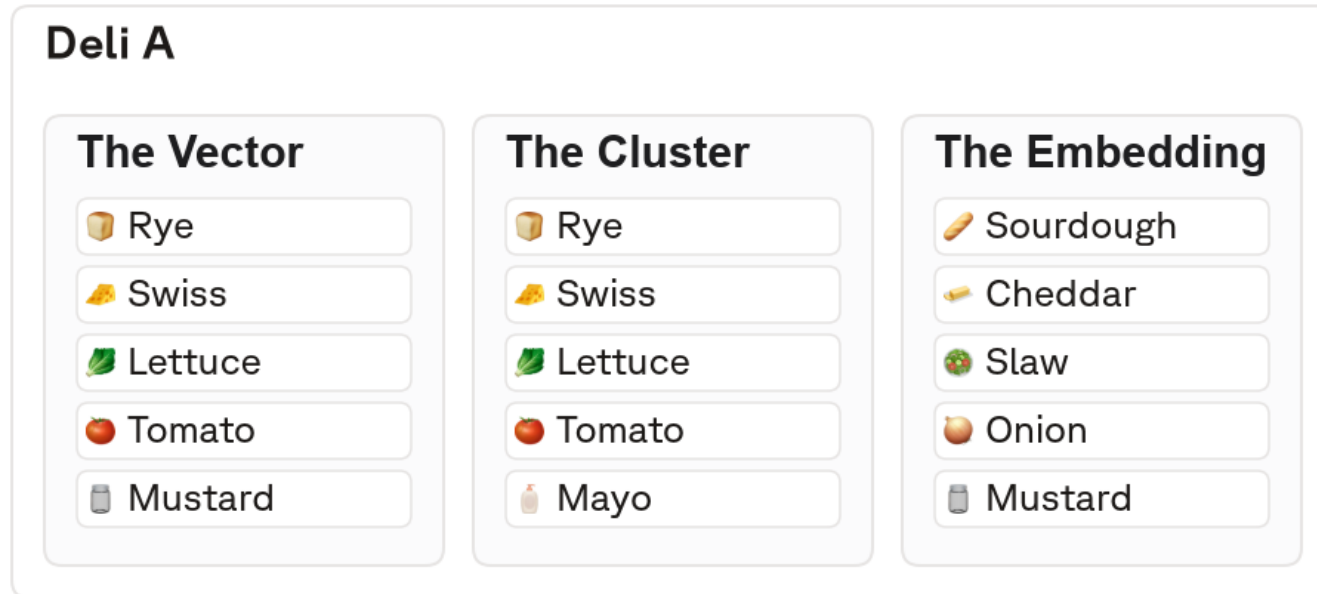
QJL

RVQ

QLoRA

GEAR

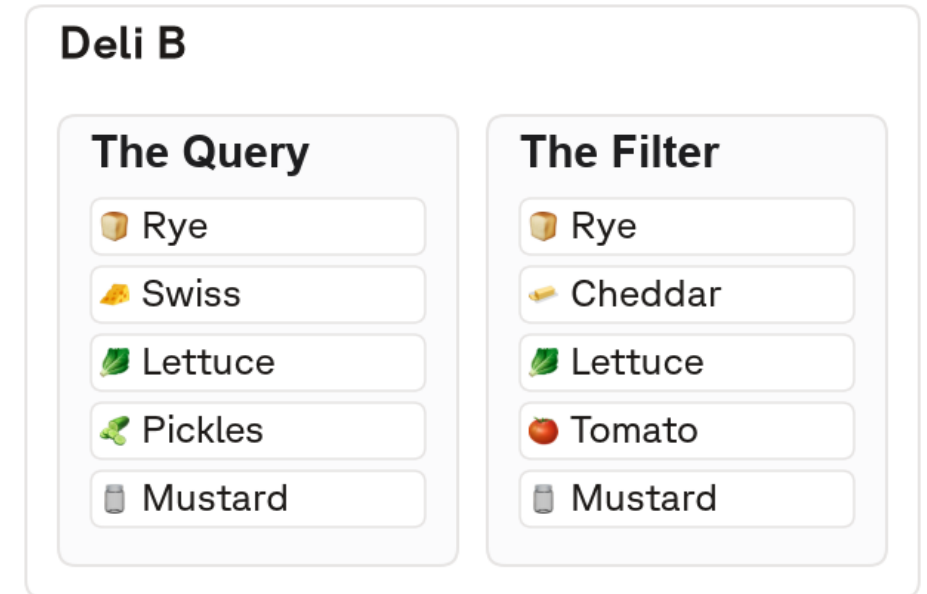
Best type of **sandwich** in NYC?



varying ingredient quality



a common pantry



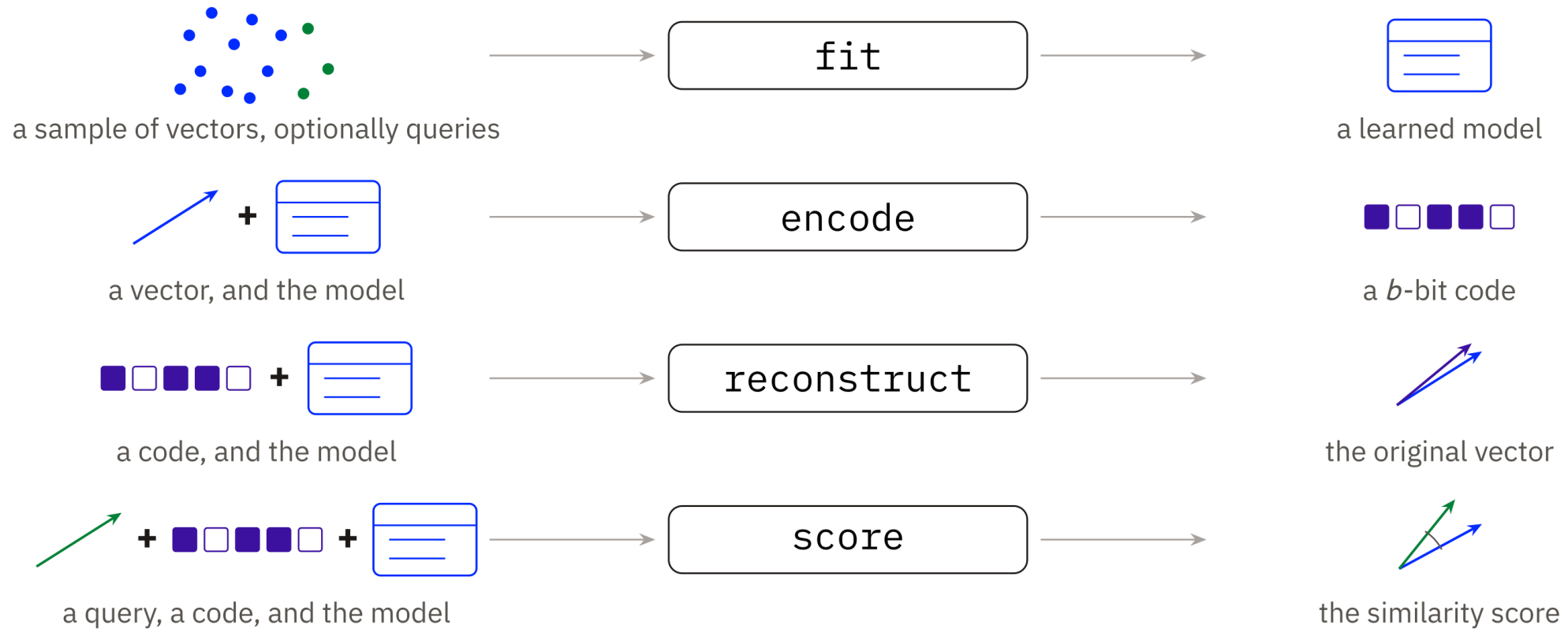
varying execution



a common cookbook

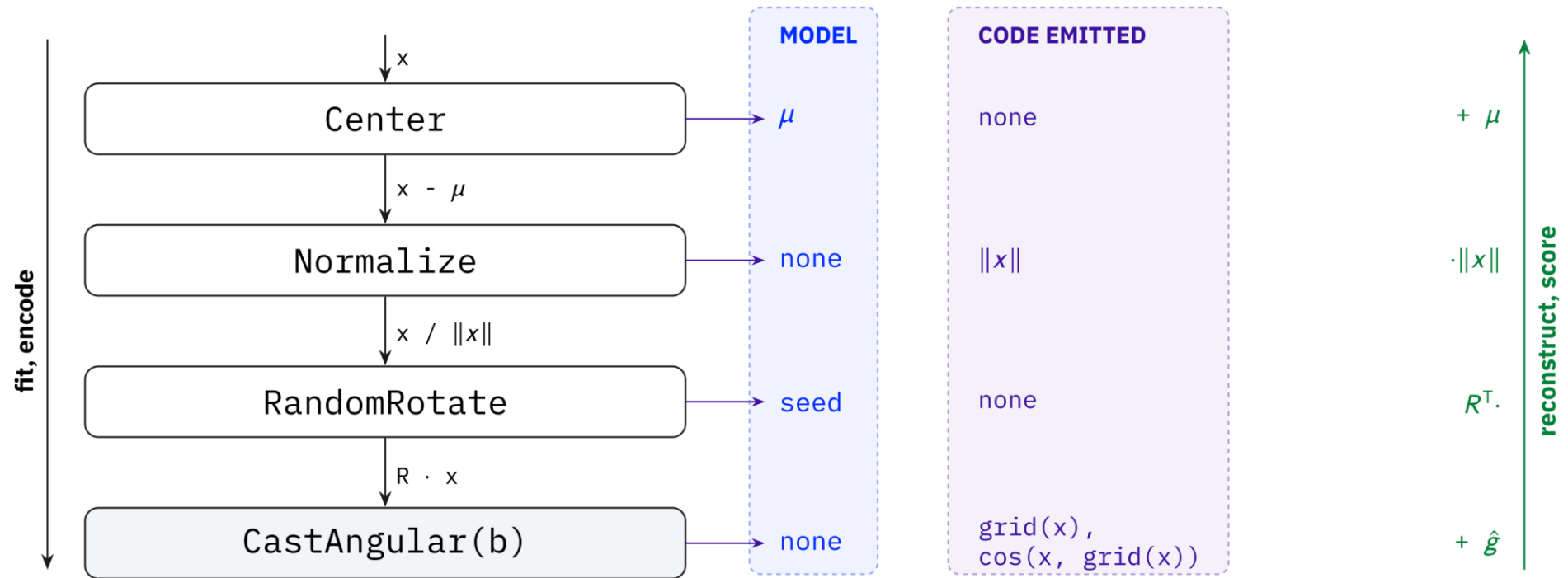
We built a **pantry** and a **cookbook**
for vector quantization

A quantizer supports **four functions**



Most quantizers are chains of **primitives**

E-RaBitQ



Primitives are special types of quantizers that compose with each other

The primitive **pantry**

Conditioners transform the data and pass it downstream.

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

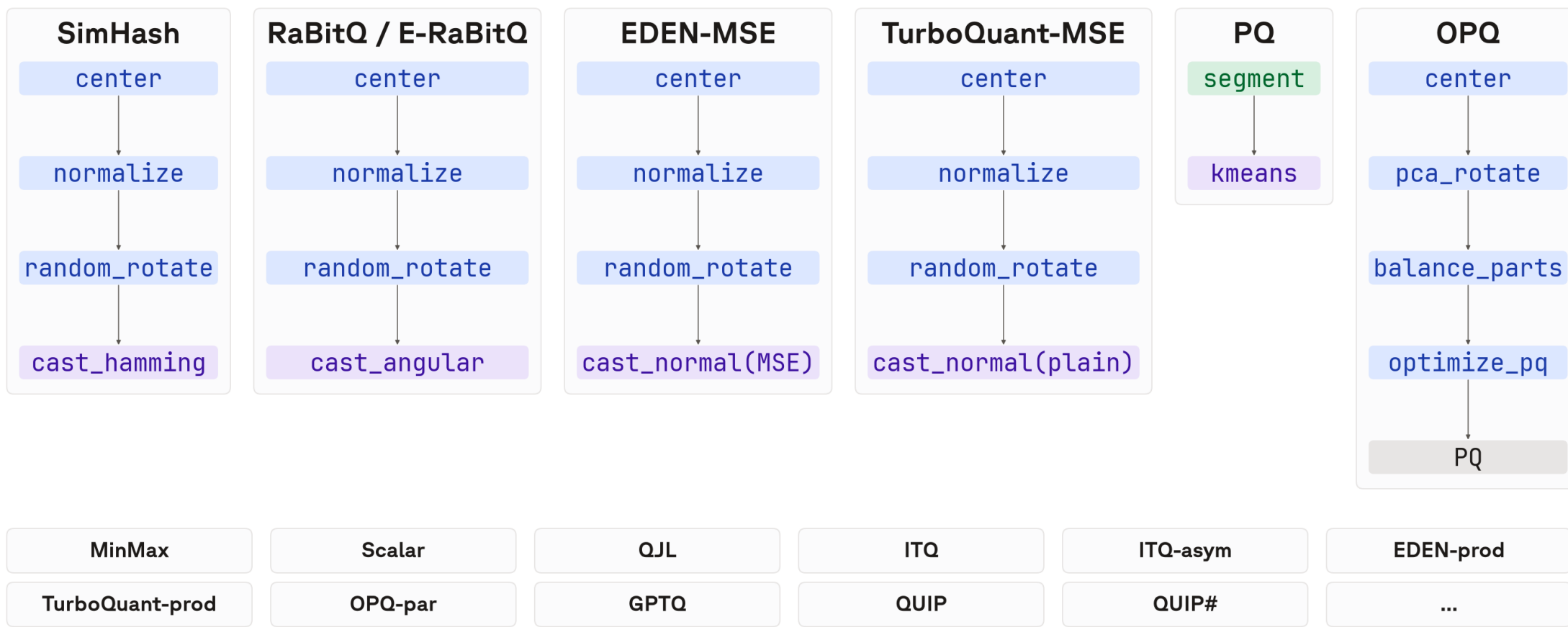
Rounders cast to a finite codebook, and pass down the residual.

cast_uint cast_normal cast_angular cast_sign cast_hamming kmeans

Splitters split the data and open a sub-pipeline for each part.

segment

The quantizer **cookbook**



The **chef**: VQ-bench harness

Build each quantizer

SimHash RaBitQ E-RaBitQ

EDEN-MSE EDEN-prod

TurboQuant-MSE

TurboQuant-prod PQ OPQ

OPQ-par MinMax Scalar

QJL ITQ ITQ-asym GPTQ

QUIP QUIP#

Evaluate on datasets (VIBE, Jääsaari et al., 2025)

ArXiv CCNews COCO

HotpotQA ImageNet

LAION Landmark

Llama MS MARCO

Yahoo

Report metrics

recall@k SOS@k

exp-SOS@k

reconstruction MSE

score MSE

KL divergence

total variation

bits/dim encode time

score time peak heap

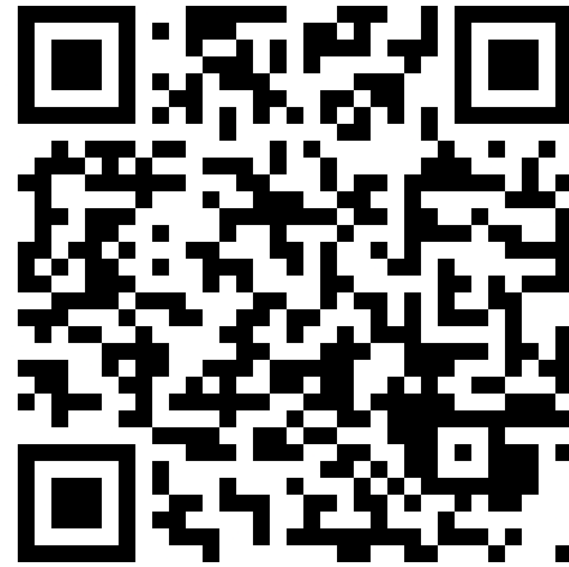
It's all **open source** – please contribute!

view the public benchmark



vq-bench.com

add a new quantizer/primitive



github.com/pinecone-io/vq-bench