Q-Tab: Quantized Tabular Data Generator

Julian Wustl[1], Philipp Haid[1], Yarema Okhrin[1], Claudius Schnörr[2]

[1]*Department of Statistics and Data Science, University of Augsburg, Germany*
[2]*Department of Computer Science and Mathematics, Munich University of Applied Sciences, Germany*

Abstract

Codebook-based generators built on masked language model (MLM) transformers have become highly effective in text and vision, yet remain underused for tabular data. This is because codebooks typically act as information bottlenecks, whereas synthetic tabular generation requires a code space larger than the training sample, with additional codes trained to support new tabular rows. We address this gap with Q-Tab, a codebook-based tabular generator that uses lookup-free quantization (LFQ) with residual corruption to jointly tokenize numerical variables, categorical variables are directly one-hot tokenized. A BERT-style MLM captures dependencies in the token space and can then be sampled from. Corruption propagates reconstruction supervision across the numerical code space, but under joint encoder–decoder training induces a moving-target regression problem whose difficulty depends on the corruption structure. This motivates residual LFQ as the quantization mechanism, balancing broader supervision with locality. Q-Tab achieves state-of-the-art predictive utility and label prediction, while matching the distributional fidelity of diffusion-based generators.

Keywords tabular data, generation, lookup-free quantization