PQWebANN Benchmark Results

Generated: 2025-12-06T08:13:53.995Z

Tests: 15/15 passed

Config: searchK=20, rerankK=10, efSearch=64, iterations=100

Latency Summary (Average)

Index	Total	WASM	IO	Rerank	P50	P95	P99	Recall
arxiv_100k_f32	1.85 ± 1.78 ms	1.37 ± 1.52 ms	0.45 ± 0.44 ms	0.02 ± 0.05 ms	1.40 ms	3.20 ms	9.50 ms	98.9 ± 4.9%
arxiv_100k_int4	1.90 ± 1.33 ms	1.34 ± 1.16 ms	0.52 ± 0.29 ms	0.03 ± 0.05 ms	1.70 ms	3.00 ms	7.60 ms	97.4 ± 11.8%
arxiv_100k_int8	1.82 ± 1.34 ms	1.37 ± 1.22 ms	0.42 ± 0.19 ms	0.03 ± 0.05 ms	1.50 ms	2.90 ms	7.50 ms	98.1 ± 6.5%
arxiv_1k_f32	1.07 ± 0.52 ms	0.71 ± 0.38 ms	0.32 ± 0.24 ms	0.02 ± 0.04 ms	0.90 ms	2.00 ms	2.90 ms	99.8 ± 1.4%
arxiv_1k_int4	1.05 ± 0.75 ms	0.72 ± 0.55 ms	0.30 ± 0.30 ms	0.03 ± 0.05 ms	0.70 ms	2.30 ms	4.40 ms	99.9 ± 1.0%

Index	Total	WASM	IO	Rerank	P50	P95	P99	Recall
arxiv_1k_int8	1.56 ± 1.63 ms	0.75 ± 0.50 ms	0.78 ± 1.35 ms	0.02 ± 0.04 ms	1.00 ms	5.00 ms	6.20 ms	99.4 ± 2.4%
finance_13k_f32	1.67 ± 0.99 ms	1.25 ± 0.73 ms	0.38 ± 0.38 ms	0.04 ± 0.05 ms	1.40 ms	3.60 ms	5.20 ms	98.7 ± 3.9%
finance_13k_int4	1.35 ± 0.79 ms	0.90 ± 0.50 ms	0.40 ± 0.44 ms	0.04 ± 0.05 ms	1.10 ms	2.90 ms	3.90 ms	97.2 ± 11.2%
finance_13k_int8	1.70 ± 0.93 ms	1.19 ± 0.72 ms	0.45 ± 0.28 ms	0.05 ± 0.06 ms	1.50 ms	3.00 ms	6.00 ms	97.8 ± 7.2%
wiki_480k_f32	5.20 ± 3.98 ms	4.31 ± 3.75 ms	0.84 ± 0.43 ms	0.04 ± 0.05 ms	4.00 ms	10.80 ms	19.70 ms	83.0 ± 26.0%
wiki_480k_int4	3.38 ± 2.84 ms	2.66 ± 2.69 ms	0.66 ± 0.22 ms	0.04 ± 0.05 ms	2.50 ms	9.90 ms	15.50 ms	82.3 ± 27.0%
wiki_480k_int8	4.14 ± 3.58 ms	3.37 ± 3.29 ms	0.72 ± 0.34 ms	0.04 ± 0.05 ms	3.10 ms	11.40 ms	20.30 ms	83.2 ± 25.8%
wiki_60k_f32	2.40 ± 1.93 ms	1.94 ± 1.71 ms	0.41 ± 0.28 ms	0.04 ± 0.05 ms	2.00 ms	4.10 ms	8.50 ms	93.3 ± 15.1%
wiki_60k_int4	1.79 ±	1.36 ± 0.99	0.38 ±	0.04 ± 0.05	1.60 ms	3.00 ms	6.20 ms	92.7 ± 19.1%

Index	Total	WASM	IO	Rerank	P50	P95	P99	Recall
	1.14 ms	ms	0.19 ms	ms				
wiki_60k_int8	2.25 ± 1.55 ms	1.70 ± 0.82 ms	0.50 ± 0.95 ms	0.04 ± 0.05 ms	2.00 ms	3.60 ms	7.20 ms	93.6 ± 17.3%

IO Latency Variance (IndexedDB)

Index	IO Avg	IO P50	IO P90	IO P95	IO P99
arxiv_100k_f32	0.45 ± 0.44 ms	0.40 ms	0.70 ms	0.80 ms	1.40 ms
arxiv_100k_int4	0.52 ± 0.29 ms	0.40 ms	0.90 ms	1.10 ms	1.50 ms
arxiv_100k_int8	0.42 ± 0.19 ms	0.40 ms	0.60 ms	0.90 ms	1.20 ms
arxiv_1k_f32	0.32 ± 0.24 ms	0.20 ms	0.70 ms	0.90 ms	1.10 ms
arxiv_1k_int4	0.30 ± 0.30 ms	0.20 ms	0.50 ms	0.90 ms	1.60 ms
arxiv_1k_int8	0.78 ± 1.35 ms	0.30 ms	1.80 ms	3.80 ms	4.40 ms
finance_13k_f32	0.38 ± 0.38 ms	0.30 ms	0.60 ms	0.80 ms	1.40 ms
finance_13k_int4	0.40 ± 0.44 ms	0.30 ms	0.70 ms	1.20 ms	2.10 ms
finance_13k_int8	0.45 ± 0.28 ms	0.40 ms	0.80 ms	1.00 ms	1.50 ms
wiki_480k_f32	0.84 ± 0.43 ms	0.70 ms	1.40 ms	1.60 ms	2.60 ms
wiki_480k_int4	0.66 ± 0.22 ms	0.60 ms	0.80 ms	1.00 ms	1.50 ms
wiki_480k_int8	0.72 ± 0.34 ms	0.70 ms	0.90 ms	1.10 ms	1.80 ms
wiki_60k_f32	0.41 ± 0.28 ms	0.30 ms	0.70 ms	0.80 ms	1.60 ms
wiki_60k_int4	0.38 ± 0.19 ms	0.30 ms	0.60 ms	0.70 ms	1.00 ms
wiki_60k_int8	0.50 ± 0.95 ms	0.40 ms	0.70 ms	0.80 ms	1.20 ms

Setup Time

Index	Total Setup	Index Build	IndexedDB Store
arxiv_100k_f32	11400 ms	143 ms	6321 ms
arxiv_100k_int4	8979 ms	119 ms	7104 ms
arxiv_100k_int8	8789 ms	120 ms	6801 ms
arxiv_1k_f32	191 ms	6 ms	76 ms
arxiv_1k_int4	140 ms	6 ms	69 ms
arxiv_1k_int8	164 ms	8 ms	86 ms
finance_13k_f32	2826 ms	28 ms	1501 ms
finance_13k_int4	2299 ms	19 ms	1921 ms
finance_13k_int8	1958 ms	19 ms	1538 ms
wiki_480k_f32	118332 ms	1555 ms	69911 ms
wiki_480k_int4	82802 ms	597 ms	70175 ms
wiki_480k_int8	82710 ms	595 ms	69532 ms
wiki_60k_f32	13065 ms	94 ms	7129 ms
wiki_60k_int4	8581 ms	70 ms	7028 ms
wiki_60k_int8	9368 ms	76 ms	7688 ms

Recall Analysis

Index	Recall Avg	Min	Max	Std
arxiv_100k_f32	98.9 ± 4.9%	70.0%	100.0%	σ = 4.9%
arxiv_100k_int4	97.4 ± 11.8%	20.0%	100.0%	σ = 11.8%
arxiv_100k_int8	98.1 ± 6.5%	70.0%	100.0%	σ = 6.5%
arxiv_1k_f32	99.8 ± 1.4%	90.0%	100.0%	σ = 1.4%
arxiv_1k_int4	99.9 ± 1.0%	90.0%	100.0%	σ = 1.0%

Index	Recall Avg	Min	Max	Std
arxiv_1k_int8	99.4 ± 2.4%	90.0%	100.0%	σ = 2.4%
finance_13k_f32	98.7 ± 3.9%	80.0%	100.0%	σ = 3.9%
finance_13k_int4	97.2 ± 11.2%	0.0%	100.0%	σ = 11.2%
finance_13k_int8	97.8 ± 7.2%	50.0%	100.0%	σ = 7.2%
wiki_480k_f32	83.0 ± 26.0%	0.0%	100.0%	σ = 26.0%
wiki_480k_int4	82.3 ± 27.0%	0.0%	100.0%	σ = 27.0%
wiki_480k_int8	83.2 ± 25.8%	0.0%	100.0%	σ = 25.8%
wiki_60k_f32	93.3 ± 15.1%	0.0%	100.0%	σ = 15.1%
wiki_60k_int4	92.7 ± 19.1%	10.0%	100.0%	σ = 19.1%
wiki_60k_int8	93.6 ± 17.3%	0.0%	100.0%	σ = 17.3%

See `benchmark_results.json` for full raw data.