PQWebANN Benchmark Results

Generated: 2025-12-06T08:02:28.596Z

Tests: 15/15 passed

Config: searchK=200, rerankK=100, efSearch=64, iterations=100

Latency Summary (Average)

Index	Total	WASM	IO	Rerank	P50	P95	P99	Recall
arxiv_100k_f32	7.67 ± 2.64 ms	4.98 ± 2.38 ms	2.43 ± 0.42 ms	0.24 ± 0.13 ms	7.00 ms	11.10 ms	17.30 ms	97.4 ± 3.9%
arxiv_100k_int4	6.41 ± 2.50 ms	3.58 ± 2.04 ms	2.55 ± 0.91 ms	0.25 ± 0.11 ms	5.80 ms	8.80 ms	16.90 ms	97.8 ± 3.5%
arxiv_100k_int8	6.81 ± 2.31 ms	3.96 ± 1.75 ms	2.57 ± 1.11 ms	0.25 ± 0.12 ms	6.20 ms	12.30 ms	15.30 ms	96.8 ± 5.4%
arxiv_1k_f32	3.54 ± 1.53 ms	1.92 ± 1.21 ms	1.35 ± 0.43 ms	0.24 ± 0.12 ms	3.00 ms	6.90 ms	10.60 ms	100.0 ± 0.2%
arxiv_1k_int4	3.35 ± 1.42 ms	1.67 ± 1.00 ms	1.39 ± 0.55 ms	0.25 ± 0.12 ms	2.90 ms	6.70 ms	9.30 ms	99.9 ± 0.3%
arxiv_1k_int8	3.36 ± 2.04 ms	1.71 ± 1.59 ms	1.37 ± 0.55 ms	0.25 ± 0.12 ms	2.90 ms	4.70 ms	10.90 ms	99.9 ± 0.3%

Index	Total	WASM	IO	Rerank	P50	P95	P99	Recall
finance_13k_f32	5.71 ± 1.47 ms	3.36 ± 1.29 ms	1.83 ± 0.44 ms	0.48 ± 0.22 ms	5.40 ms	8.70 ms	10.40 ms	98.8 ± 2.3%
finance_13k_int4	5.01 ± 1.71 ms	2.65 ± 1.38 ms	1.85 ± 0.49 ms	0.47 ± 0.22 ms	4.50 ms	8.30 ms	10.80 ms	98.8 ± 2.5%
finance_13k_int8	5.31 ± 2.22 ms	2.93 ± 1.68 ms	1.89 ± 0.73 ms	0.46 ± 0.22 ms	4.80 ms	7.70 ms	13.90 ms	98.9 ± 1.6%
wiki_480k_f32	19.07 ± 11.20 ms	10.63 ± 6.45 ms	7.93 ± 8.60 ms	0.47 ± 0.22 ms	16.30 ms	28.70 ms	64.60 ms	86.4 ± 20.5%
wiki_480k_int4	13.64 ± 3.33 ms	7.80 ± 2.81 ms	5.33 ± 0.86 ms	0.47 ± 0.22 ms	13.20 ms	18.70 ms	22.80 ms	87.9 ± 11.2%
wiki_480k_int8	13.21 ± 3.33 ms	7.63 ± 2.92 ms	5.07 ± 0.74 ms	0.48 ± 0.22 ms	13.10 ms	18.10 ms	26.80 ms	88.0 ± 13.1%
wiki_60k_f32	8.39 ± 2.03 ms	5.48 ± 1.78 ms	2.43 ± 0.37 ms	0.45 ± 0.22 ms	8.30 ms	9.90 ms	15.70 ms	95.7 ± 4.6%
wiki_60k_int4	7.61 ± 2.03 ms	4.44 ± 1.70 ms	2.67 ± 0.56 ms	0.48 ± 0.22 ms	7.30 ms	9.80 ms	14.40 ms	95.5 ± 5.0%

Index	Total	WASM	IO	Rerank	P50	P95	P99	Recall
wiki_60k_int8	9.74 ± 2.19 ms	6.08 ± 1.82 ms	3.13 ± 0.94 ms	0.49 ± 0.23 ms	9.60 ms	13.50 ms	16.90 ms	95.8 ± 5.1%

IO Latency Variance (IndexedDB)

Index	IO Avg	IO P50	IO P90	IO P95	IO P99
arxiv_100k_f32	2.43 ± 0.42 ms	2.40 ms	2.80 ms	2.90 ms	3.70 ms
arxiv_100k_int4	2.55 ± 0.91 ms	2.40 ms	2.70 ms	2.90 ms	6.60 ms
arxiv_100k_int8	2.57 ± 1.11 ms	2.40 ms	2.70 ms	3.00 ms	7.10 ms
arxiv_1k_f32	1.35 ± 0.43 ms	1.20 ms	1.70 ms	2.30 ms	3.10 ms
arxiv_1k_int4	1.39 ± 0.55 ms	1.20 ms	1.90 ms	2.10 ms	3.80 ms
arxiv_1k_int8	1.37 ± 0.55 ms	1.20 ms	1.70 ms	2.10 ms	3.20 ms
finance_13k_f32	1.83 ± 0.44 ms	1.70 ms	2.20 ms	2.50 ms	3.40 ms
finance_13k_int4	1.85 ± 0.49 ms	1.70 ms	2.20 ms	2.40 ms	4.10 ms
finance_13k_int8	1.89 ± 0.73 ms	1.70 ms	2.10 ms	2.90 ms	4.60 ms
wiki_480k_f32	7.93 ± 8.60 ms	6.00 ms	9.50 ms	14.80 ms	39.60 ms
wiki_480k_int4	5.33 ± 0.86 ms	5.30 ms	6.20 ms	7.00 ms	7.60 ms
wiki_480k_int8	5.07 ± 0.74 ms	5.20 ms	5.90 ms	6.00 ms	6.20 ms
wiki_60k_f32	2.43 ± 0.37 ms	2.40 ms	2.70 ms	2.90 ms	3.40 ms
wiki_60k_int4	2.67 ± 0.56 ms	2.60 ms	2.90 ms	3.00 ms	5.30 ms
wiki_60k_int8	3.13 ± 0.94 ms	2.90 ms	4.10 ms	4.80 ms	7.30 ms

Setup Time

Index	Total Setup	Index Build	IndexedDB Store
arxiv_100k_f32	11398 ms	133 ms	6275 ms

Index	Total Setup	Index Build	IndexedDB Store
arxiv_100k_int4	8213 ms	113 ms	6395 ms
arxiv_100k_int8	8430 ms	118 ms	6452 ms
arxiv_1k_f32	173 ms	5 ms	70 ms
arxiv_1k_int4	131 ms	5 ms	64 ms
arxiv_1k_int8	153 ms	5 ms	65 ms
finance_13k_f32	2724 ms	26 ms	1477 ms
finance_13k_int4	1770 ms	19 ms	1423 ms
finance_13k_int8	1786 ms	18 ms	1407 ms
wiki_480k_f32	118458 ms	1419 ms	69661 ms
wiki_480k_int4	78303 ms	594 ms	65572 ms
wiki_480k_int8	80909 ms	579 ms	68012 ms
wiki_60k_f32	13263 ms	97 ms	7303 ms
wiki_60k_int4	8519 ms	71 ms	6950 ms
wiki_60k_int8	8915 ms	73 ms	7282 ms

Recall Analysis

Index	Recall Avg	Min	Max	Std
arxiv_100k_f32	97.4 ± 3.9%	74.0%	100.0%	σ = 3.9%
arxiv_100k_int4	97.8 ± 3.5%	85.0%	100.0%	σ = 3.5%
arxiv_100k_int8	96.8 ± 5.4%	68.0%	100.0%	σ = 5.4%
arxiv_1k_f32	100.0 ± 0.2%	99.0%	100.0%	σ = 0.2%
arxiv_1k_int4	99.9 ± 0.3%	98.0%	100.0%	σ = 0.3%
arxiv_1k_int8	99.9 ± 0.3%	99.0%	100.0%	σ = 0.3%
finance_13k_f32	98.8 ± 2.3%	88.0%	100.0%	σ = 2.3%
finance_13k_int4	98.8 ± 2.5%	84.0%	100.0%	σ = 2.5%

Index	Recall Avg	Min	Max	Std
finance_13k_int8	98.9 ± 1.6%	90.0%	100.0%	σ = 1.6%
wiki_480k_f32	86.4 ± 20.5%	0.0%	100.0%	σ = 20.5%
wiki_480k_int4	87.9 ± 11.2%	43.0%	100.0%	σ = 11.2%
wiki_480k_int8	88.0 ± 13.1%	23.0%	100.0%	σ = 13.1%
wiki_60k_f32	95.7 ± 4.6%	81.0%	100.0%	σ = 4.6%
wiki_60k_int4	95.5 ± 5.0%	68.0%	100.0%	σ = 5.0%
wiki_60k_int8	95.8 ± 5.1%	78.0%	100.0%	σ = 5.1%

See `benchmark_results.json` for full raw data.