Statistical Benchmark Report

Generated: 2025-11-18 12:14

This report provides quantitative evaluation results for NDCG@k and faithfulness, correlation diagnostics, outlier detection, and statistical visualizations.

1. Summary Statistics

files	100.0000
ndcg@k mean	0.9977
ndcg@k ci95 lo	0.9952
ndcg@k ci95 hi	0.9995
faith mean	0.4682
faith ci95 lo	0.4419
faith ci95 hi	0.4951
ndcg@k median	1.0000
faith median	0.4598
ndcg@k std	0.0110
faith std	0.1385
bootstrap iters	2000.0000
iqr k	1.5000
z thresh	3.0000
faith high thr	0.5000
faith mid thr	0.2500

2. Correlation and Outlier Analysis

Metric Pair	Correlation	p-value
Pearson r	0.012	9.023e-01
Spearman ρ	0.110	2.765e-01
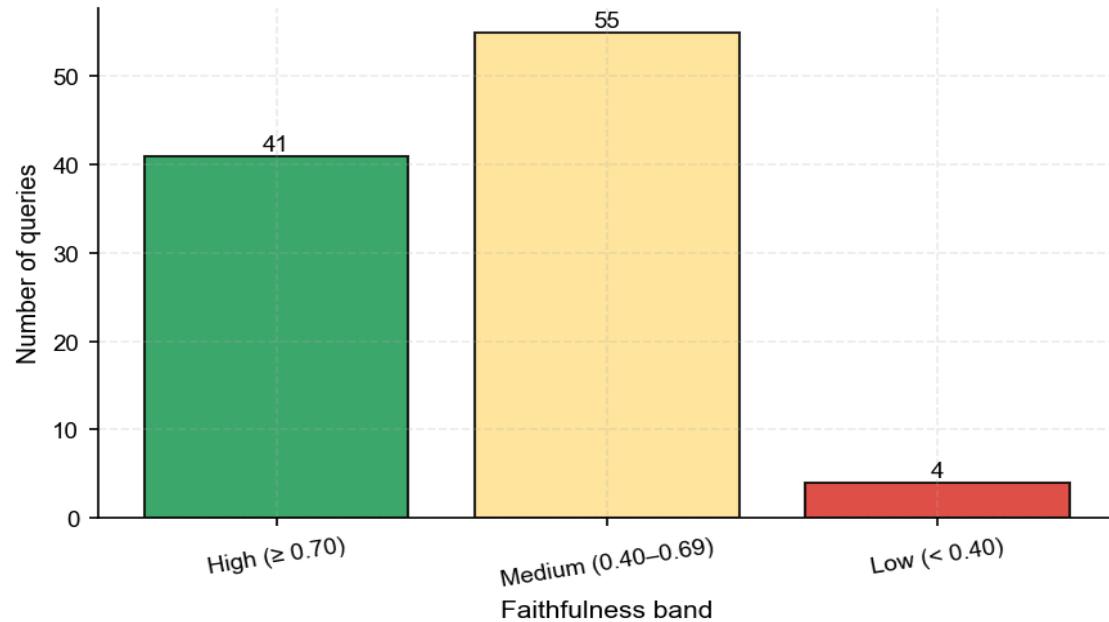
Detected Outliers

Query	NDCG@k	Faithfulness	z_ndcg	z_faith
Define_describe_the_notion_of_transparency_in_ai_decision_processes___describe_i	0.910	0.604	-7.96	0.97
Define_explain_the_concept_of_heuristic_search_as_described_in_the_corpus___desc	0.953	0.335	-4.00	-0.96

3. Visual Analytics
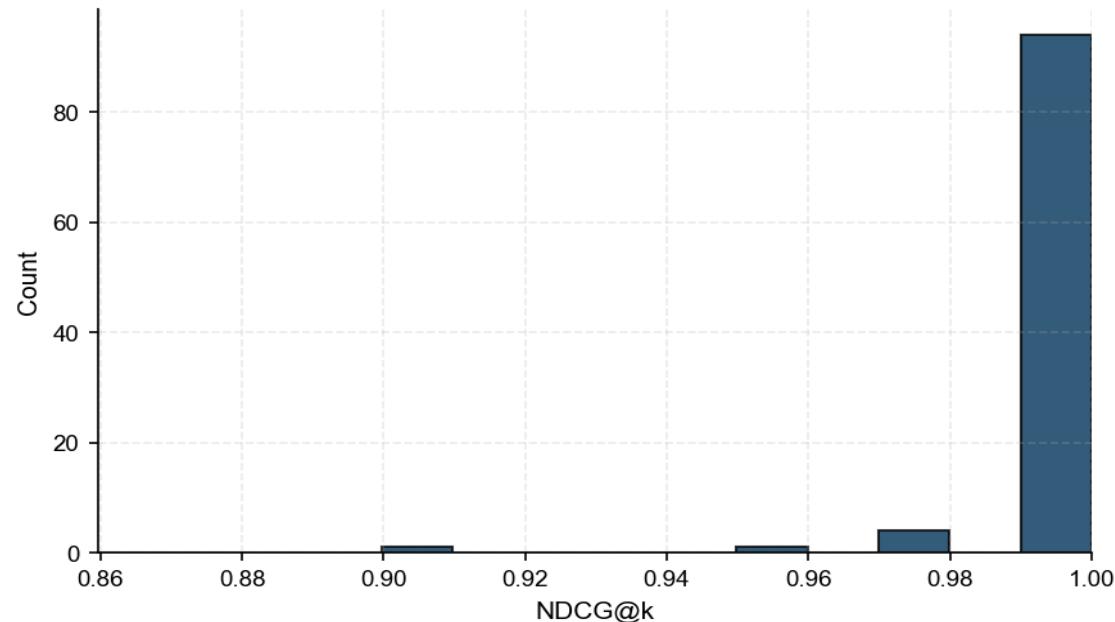
faithfulness bands global.png

Figure 2: Faithfulness bands across all queries
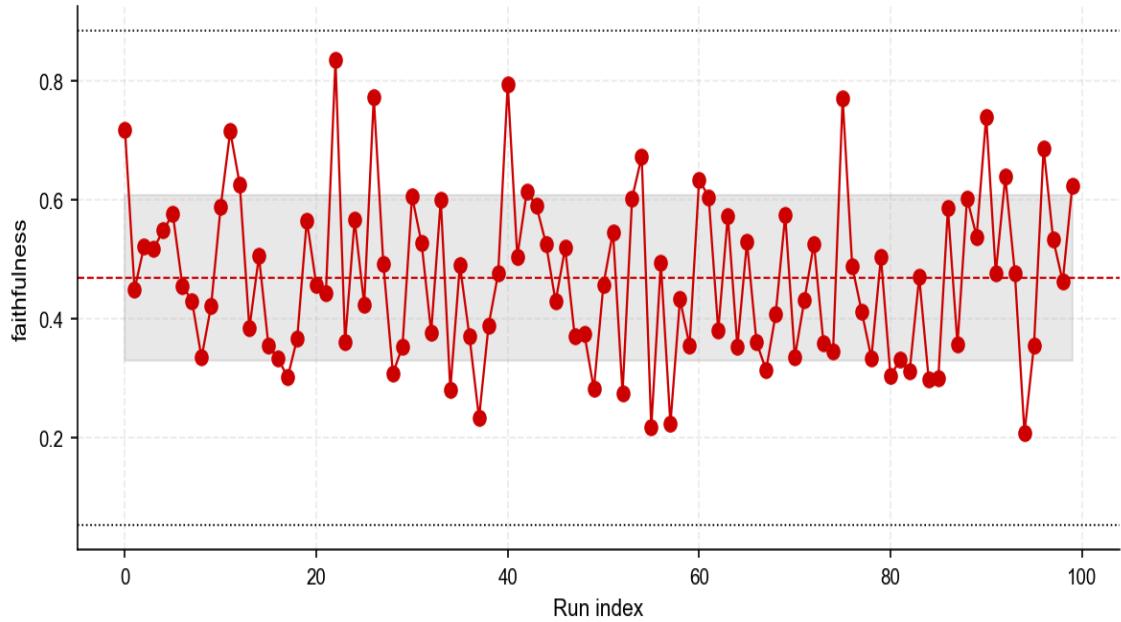
hist ndcg.png

Figure 1: Distribution of NDCG@k
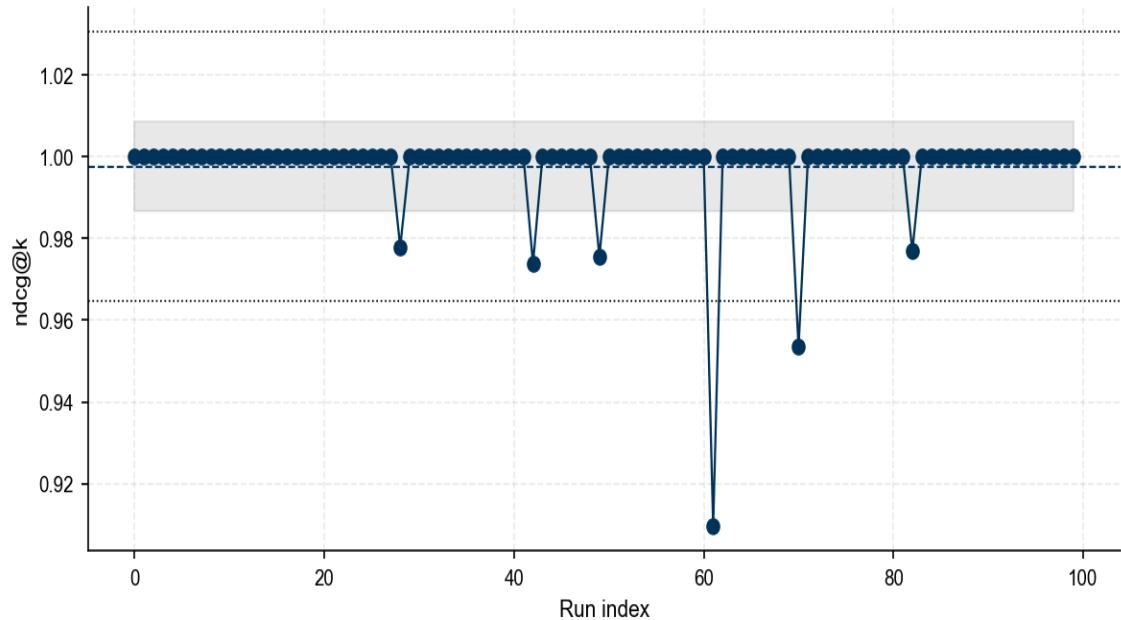
run order faithfulness.png

Figure 5: Run-order chart for faithfulness

run order ndcg@k.png

Figure 4: Run-order chart for ndcg@k

scatter ndcg vs faithfulness.png

Figure 3: Scatter NDCG@k vs Faithfulness (r=0.012, ρ=0.110)