

Statistical Benchmark Report

Generated: 2025-11-18 14:39

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	0.4608
faith ci95 lo	0.4348
faith ci95 hi	0.4872
ndcg@k median	1.0000
faith median	0.4518
ndcg@k std	0.0110
faith std	0.1329
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	-0.018	8.596e-01
Spearman p	0.025	8.058e-01

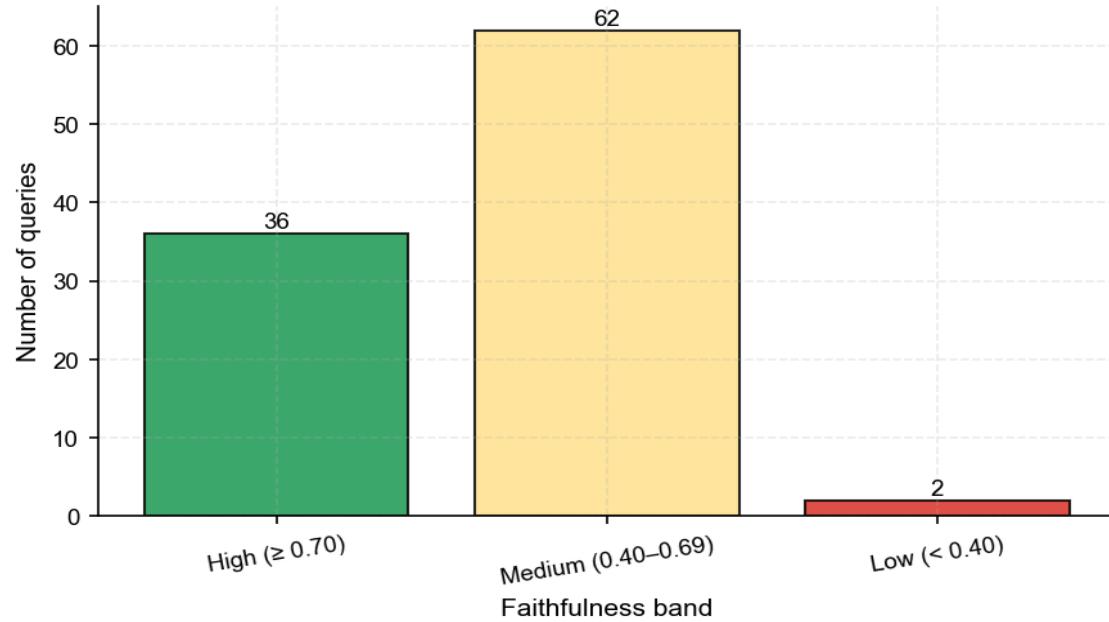
Detected Outliers

Query	NDCG@k	Faithfulness	z_ndcg	z_faith
Define_describe_the_notion_of_transparency_in_ai_decision_processes___describe_i	0.910	0.586	-7.96	0.94
Define_explain_the_concept_of_heuristic_search_as_described_in_the_corpus___desc	0.953	0.308	-4.00	-1.15

3. Visual Analytics

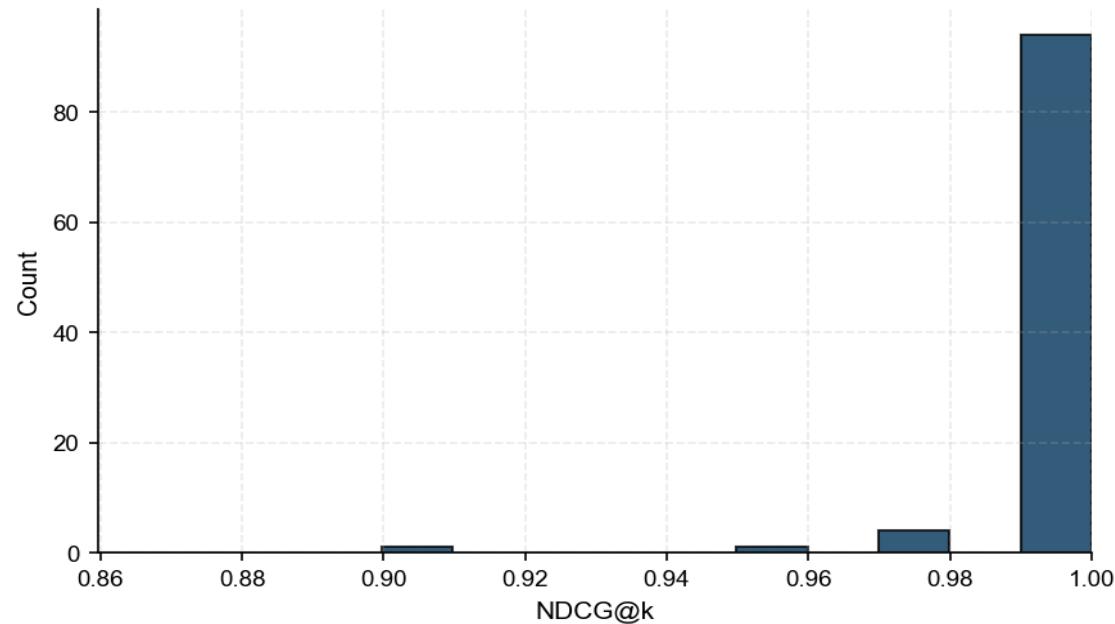
faithfulness bands global.png

Figure 2: Faithfulness bands across all queries



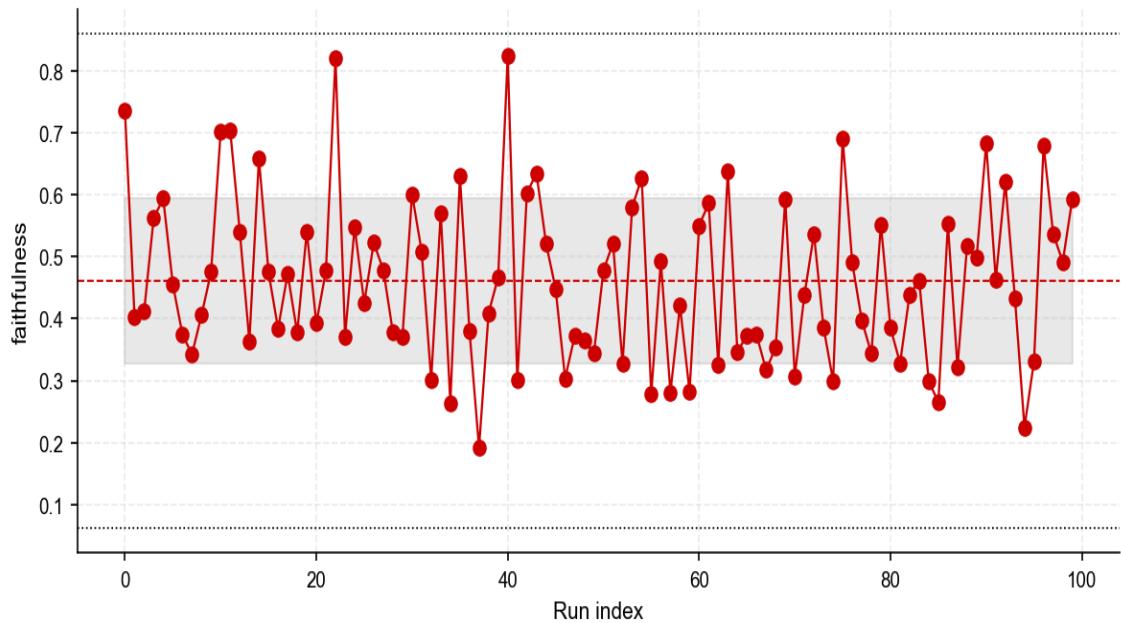
hist ndcg.png

Figure 1: Distribution of NDCG@k



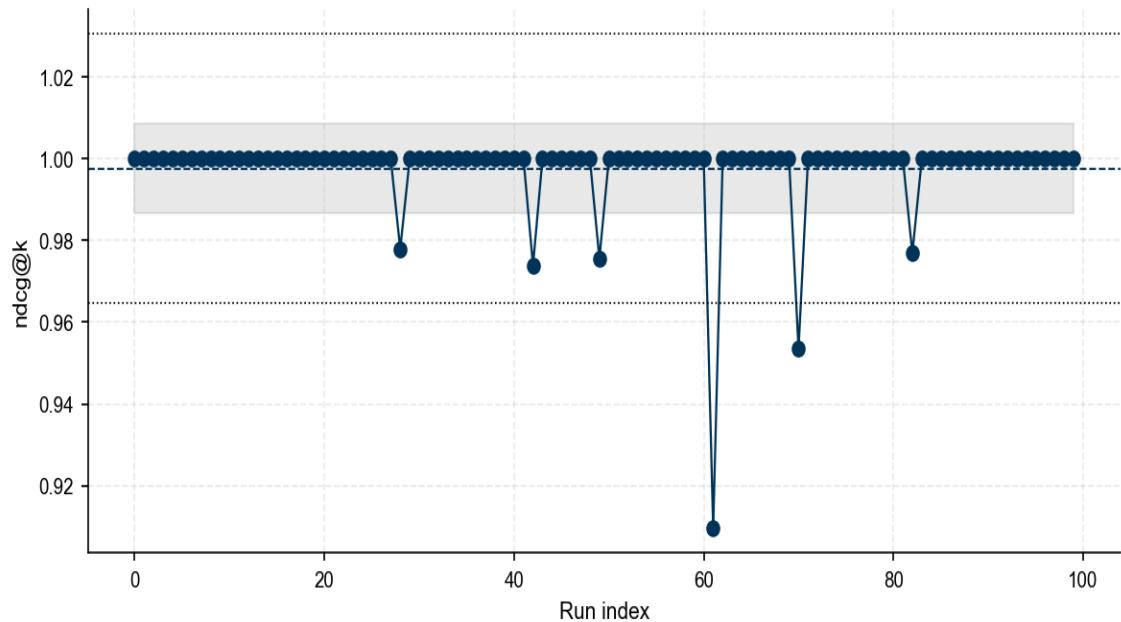
run order faithfulness.png

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 ($r=-0.018$, $p=0.025$)

