

# Statistical Benchmark Report

Generated: 2025-11-17 10:07

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	0.9997
ndcg@k ci95 lo	0.9991
ndcg@k ci95 hi	1.0000
faith mean	0.5212
faith ci95 lo	0.4854
faith ci95 hi	0.5587
ndcg@k median	1.0000
faith median	0.4947
ndcg@k std	0.0026
faith std	0.1845
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	-0.029	7.743e-01
Spearman p	0.012	9.090e-01

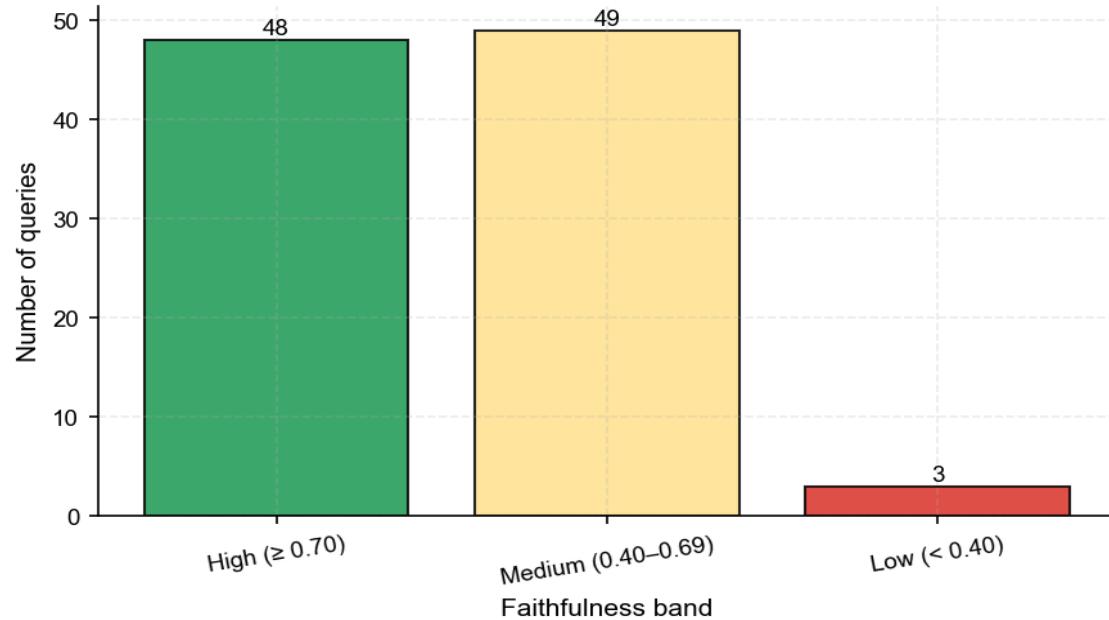
### Detected Outliers

Query	NDCG@k	Faithfulness	z_ndcg	z_faith
Define_describe_the_notion_of_transparency_in_ai_decision_processes___describe_i	0.974	0.593	-9.82	0.38

### 3. Visual Analytics

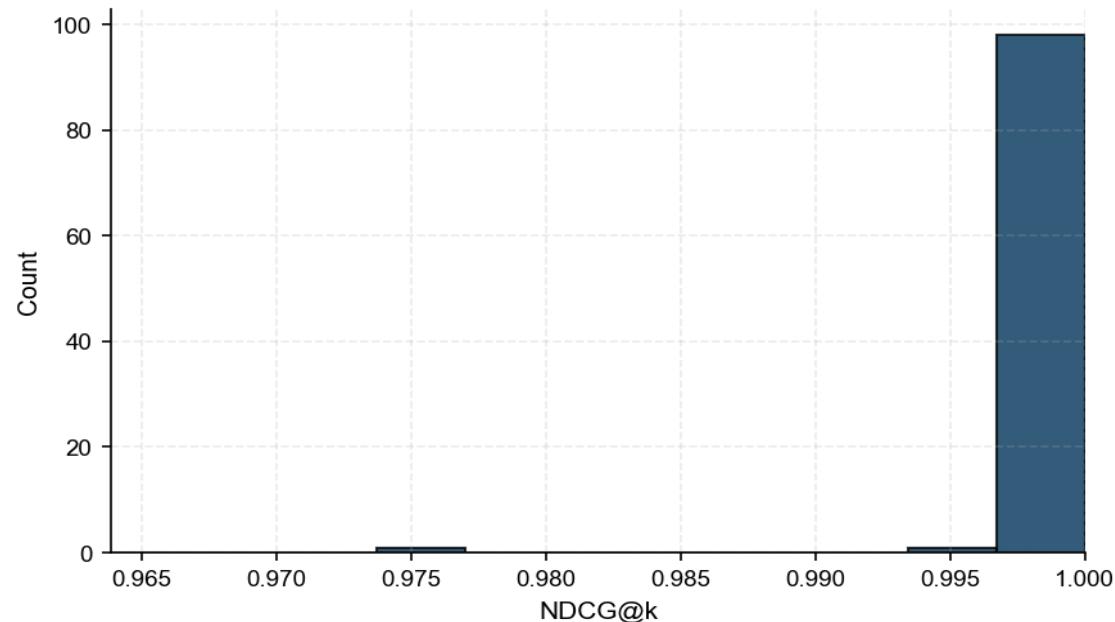
**faithfulness bands global.png**

Figure 2: Faithfulness bands across all queries



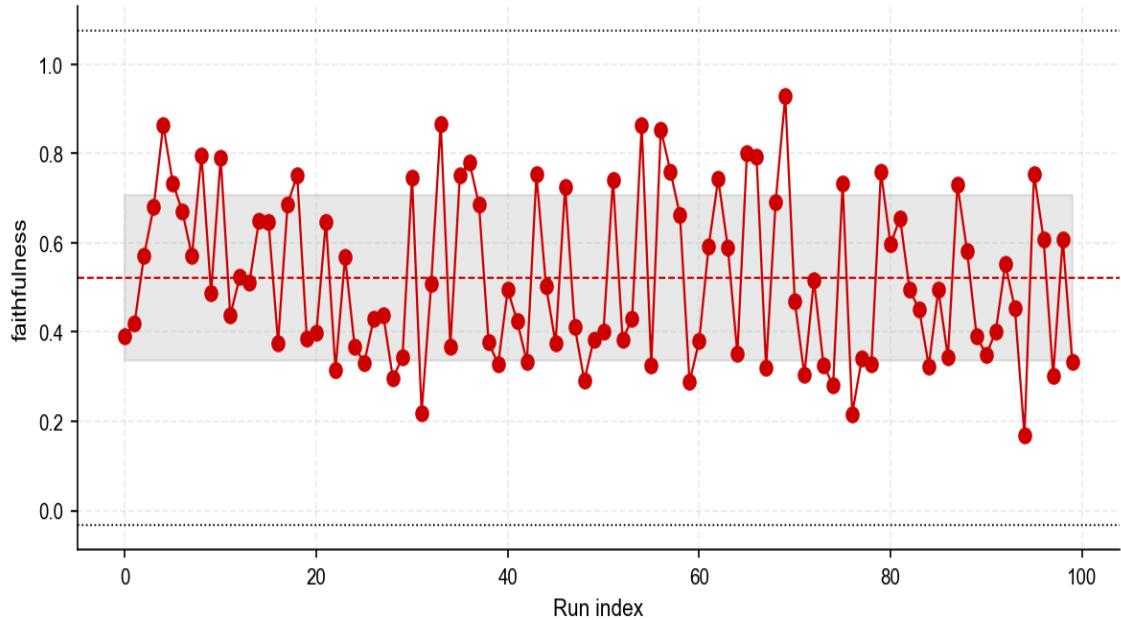
**hist ndcg.png**

Figure 1: Distribution of NDCG@k



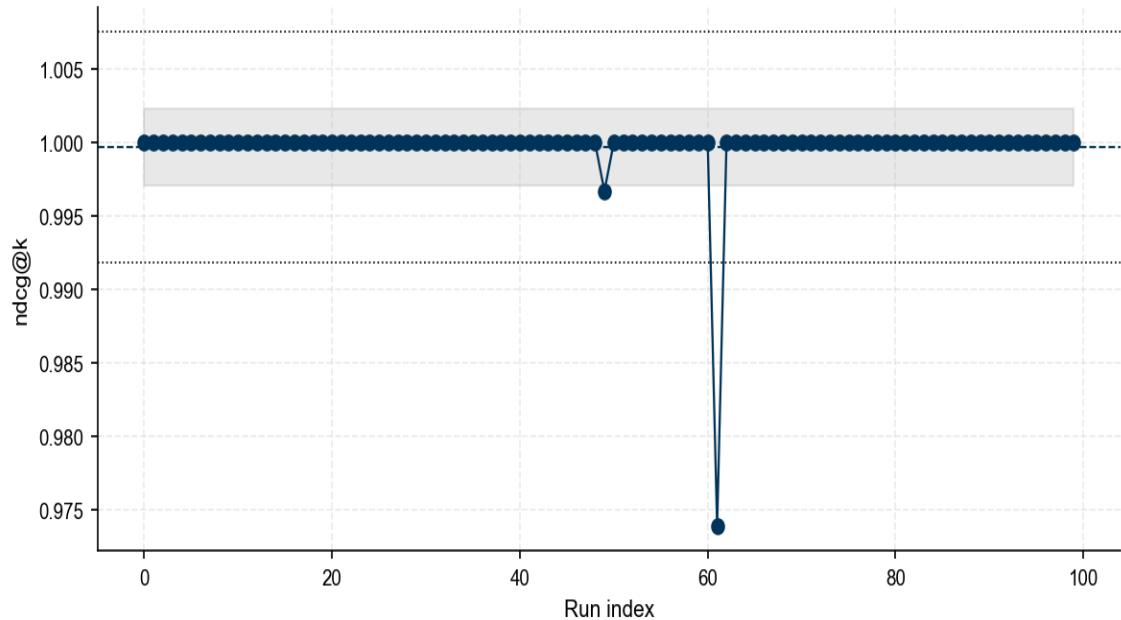
**run order faithfulness.png**

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 ( $r=-0.029$ ,  $p=0.012$ )

