Statistical Benchmark Report

Generated: 2025-11-17 03:29

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	0.5196
faith ci95 lo	0.4854
faith ci95 hi	0.5531
ndcg@k median	1.0000
faith median	0.5062
ndcg@k std	0.0026
faith std	0.1749
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

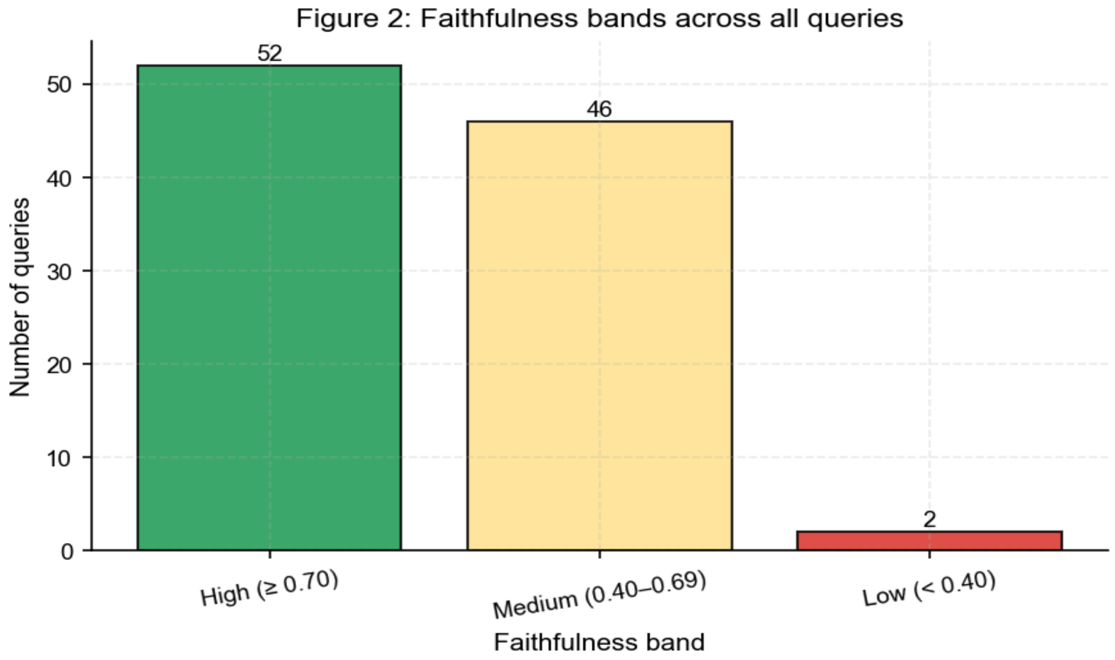
Metric Pair	Correlation	p-value
Pearson r	-0.085	4.023e-01
Spearman ρ	-0.023	8.179e-01

Detected Outliers

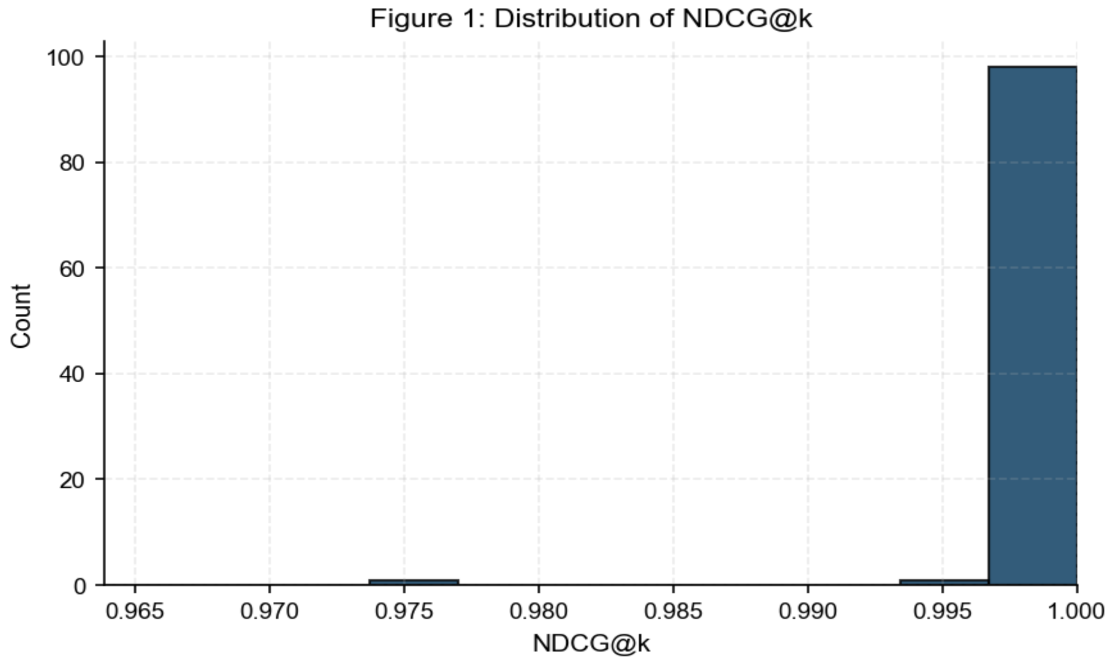
Query	NDCG@k	Faithfulness	z_ndcg	z_faith
Define_describe_the_notion_of_transparency_in_ai_decision_processes___describe_i	0.974	0.687	-9.82	0.95

3. Visual Analytics

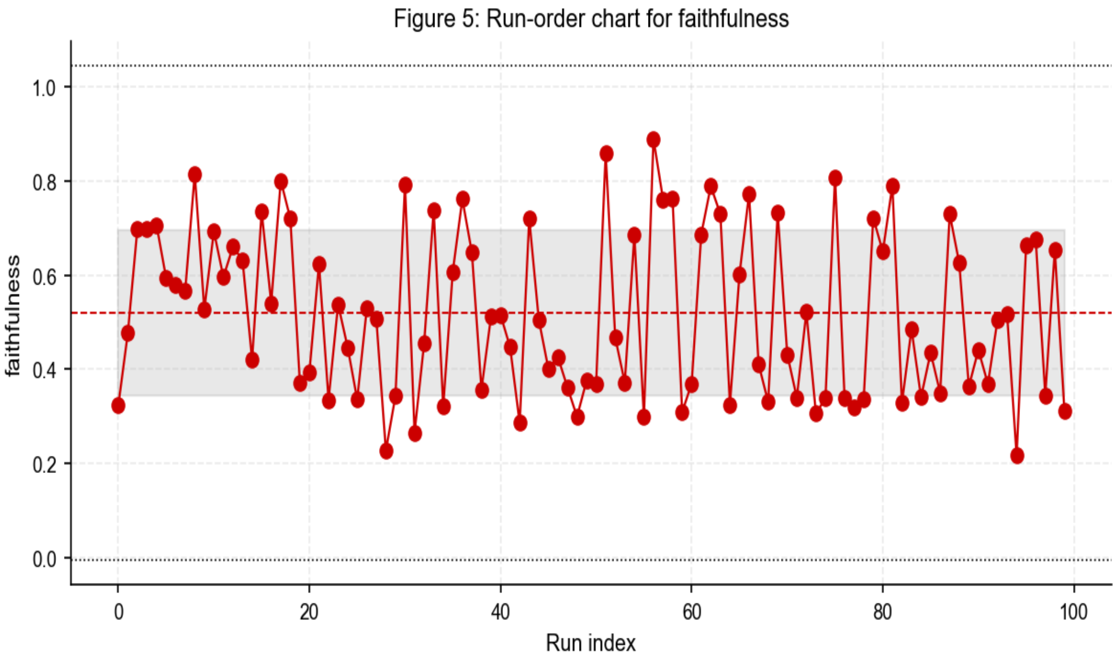
faithfulness bands global.png



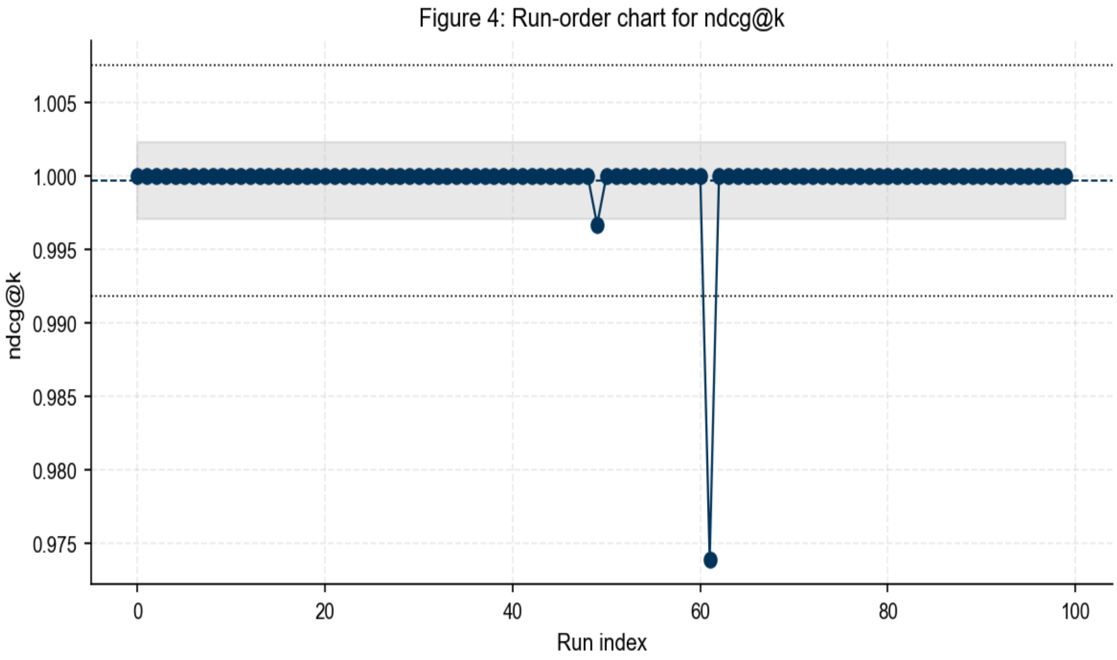
hist ndcg.png



run order faithfulness.png

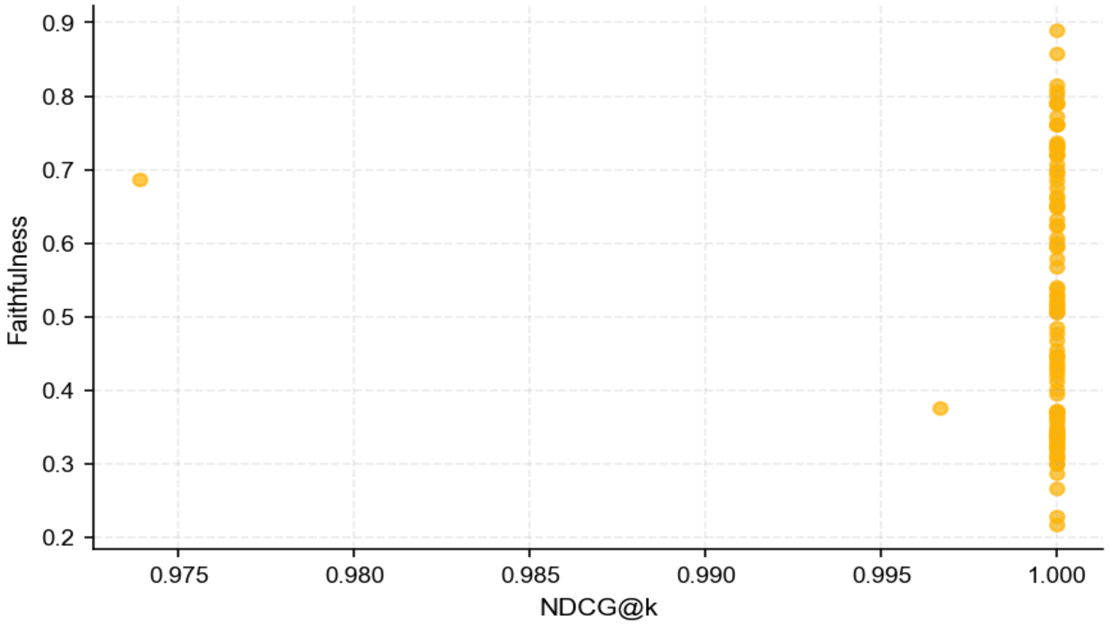


run order ndcg@k.png



scatter ndcg vs faithfulness.png

Figure 3: Scatter NDCG@k vs Faithfulness (r=-0.085, ρ=-0.023)


Faithfulness
NDCG@k