

Statistical Benchmark Report

Generated: 2025-11-16 20:34

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	1.0000
ndcg@k ci95 lo	1.0000
ndcg@k ci95 hi	1.0000
faith mean	0.5600
faith ci95 lo	0.5246
faith ci95 hi	0.5965
ndcg@k median	1.0000
faith median	0.6073
ndcg@k std	0.0002
faith std	0.1817
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	0.090	3.732e-01
Spearman p	0.075	4.592e-01

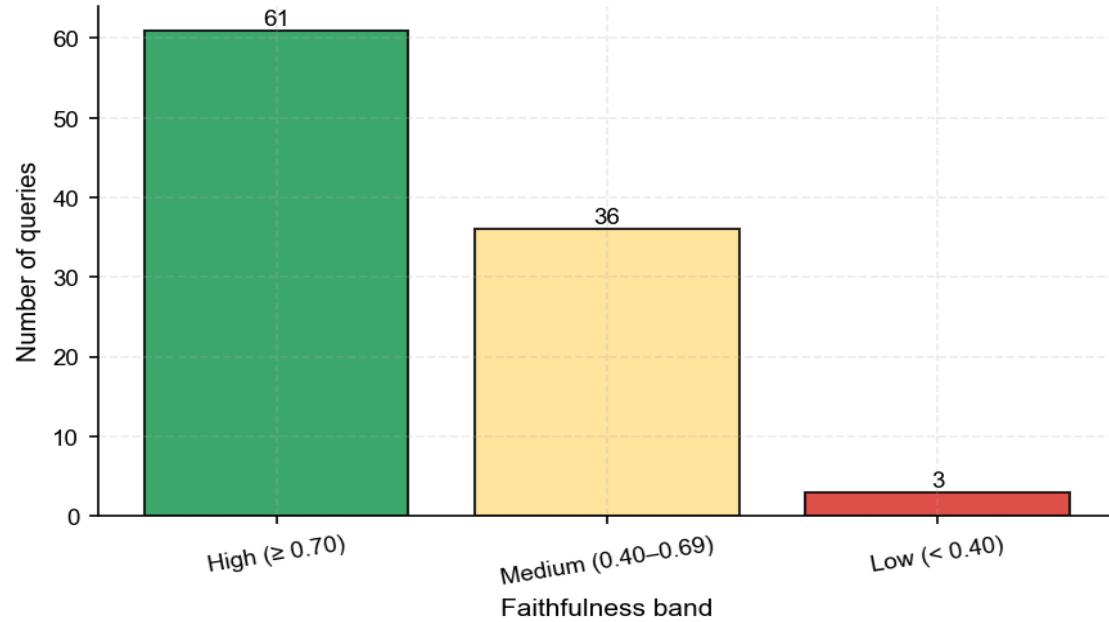
Detected Outliers

Query	NDCG@k	Faithfulness	z_ndcg	z_faith
Compare_and_contrast_the_main_theoretical_perspectives_on_summarize_the_emergenc	0.998	0.398	-9.84	-0.89

3. Visual Analytics

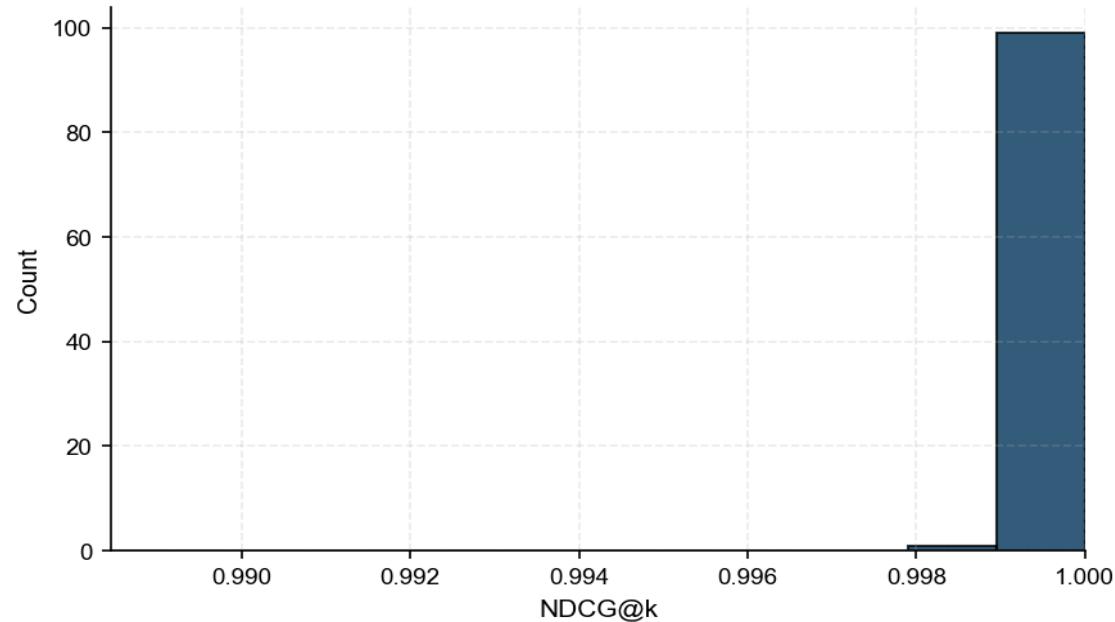
faithfulness bands global.png

Figure 2: Faithfulness bands across all queries



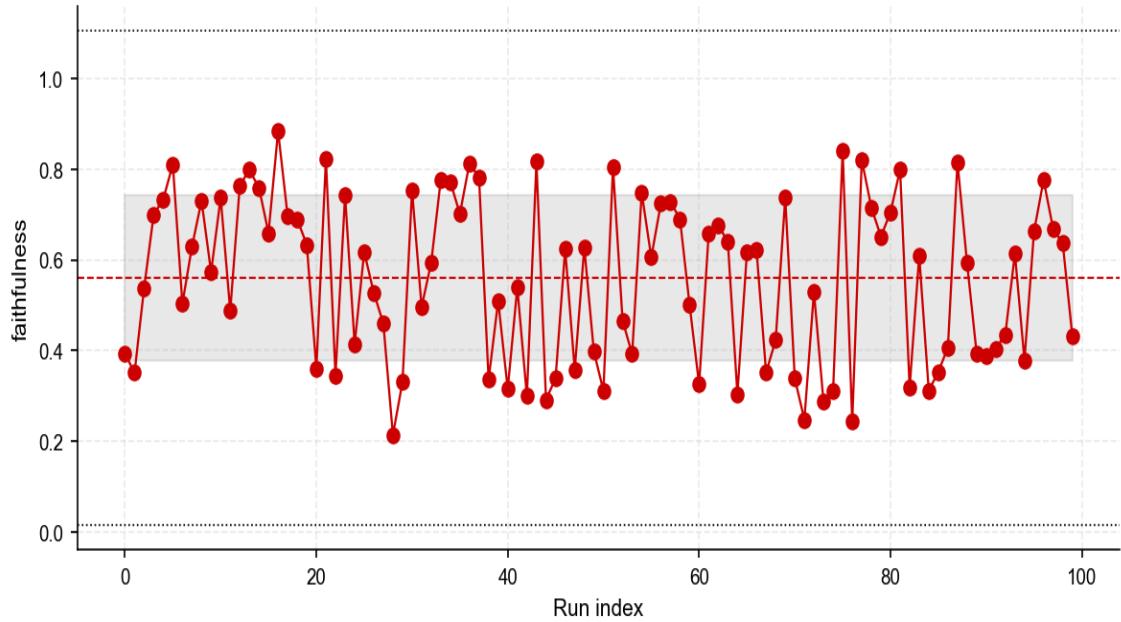
hist ndcg.png

Figure 1: Distribution of NDCG@k



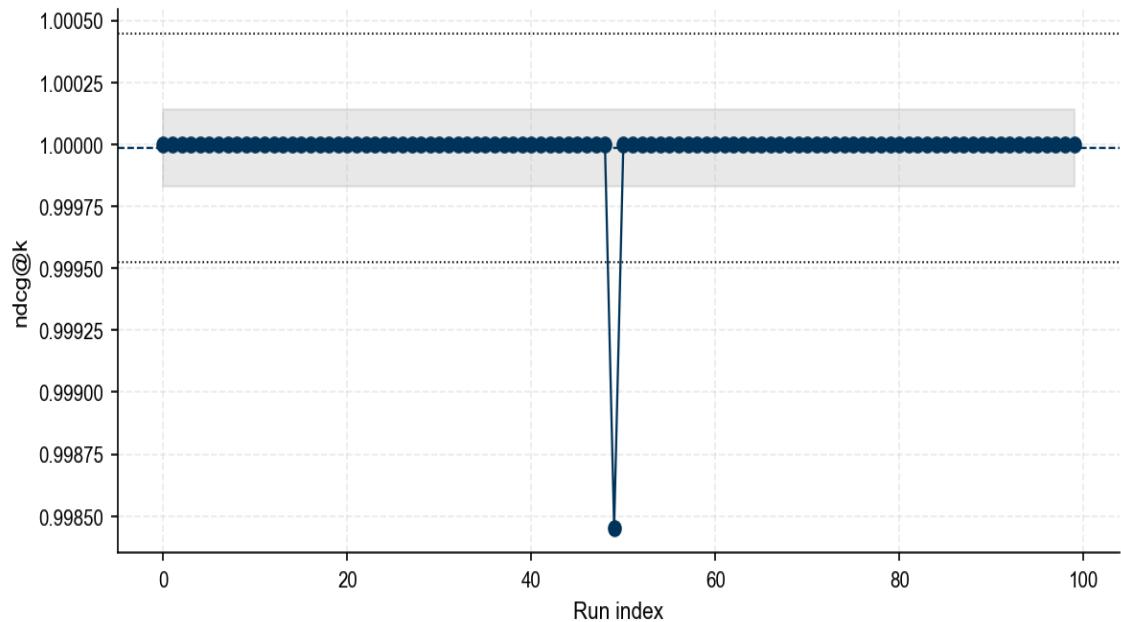
run order faithfulness.png

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 ($r=0.090$, $p=0.075$)

