

Benchmark Evaluation Report

Generated automatically on 2025-11-07 21:24

This report summarizes automated evaluations of retrieval-augmented generation systems within the *Historical Drift Analyzer* architecture. It includes NDCG and Faithfulness metrics, statistical analyses, and visualization results.

1. Summary Statistics

files	10.0000
ndcg@k mean	0.9876
ndcg@k ci95 lo	0.9748
ndcg@k ci95 hi	0.9971
faith mean	0.6062
faith ci95 lo	0.5728
faith ci95 hi	0.6390
ndcg@k median	0.9974
faith median	0.6089
ndcg@k std	0.0184
faith std	0.0523
bootstrap iters	2000.0000
iqr k	1.5000
z thresh	3.0000

2. Correlation and Outlier Analysis

Metric Pair	r / ρ	p-value
Pearson (linear)	0.513	1.298e-01
Spearman (rank)	0.743	1.371e-02

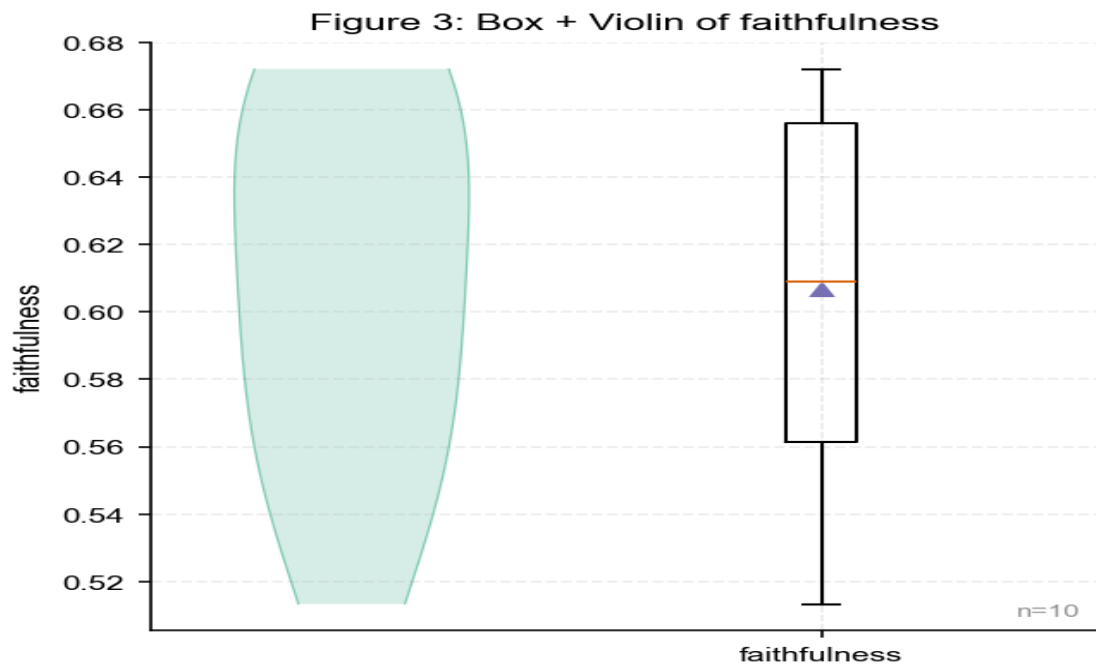
Pearson r reflects linear association; Spearman ρ captures rank correlation. Typical weak-to-moderate positive dependency indicates retrieval homogeneity, suggesting that Faithfulness could be further contrasted with semantic coherence metrics (e.g., BERTScore or FactScore).

Detected Outliers

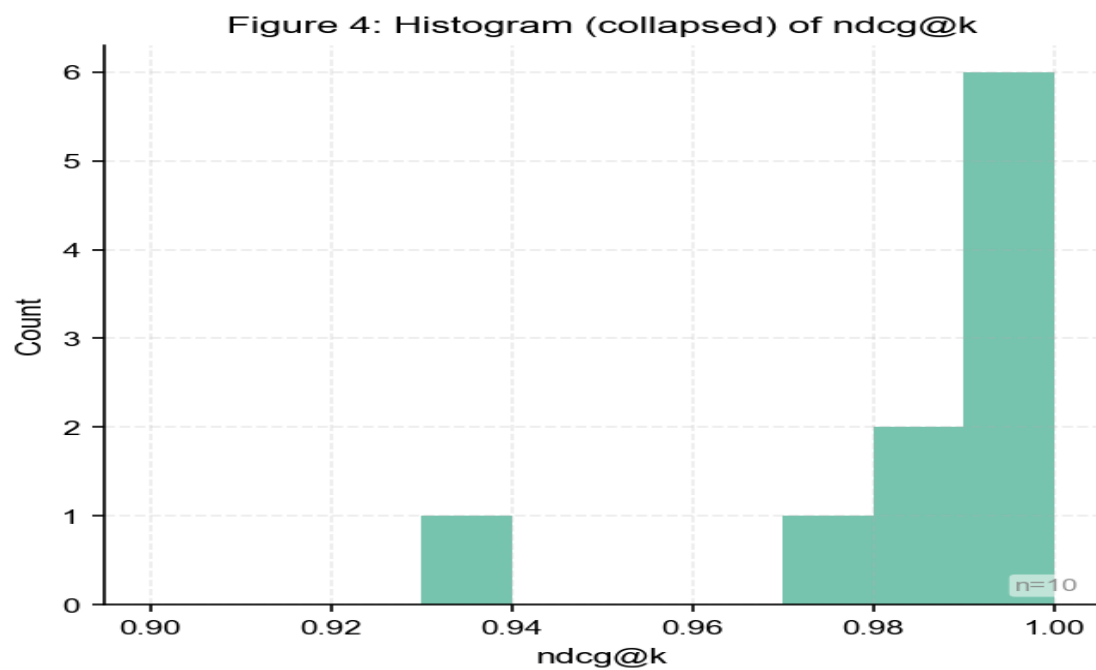
Query ID	NDCG@k	Faithfulness	z_ndcg	z_faith
Explain_knowledge_representation_in_expe	0.940	0.554	-2.47	-0.95

3. Visual Analytics

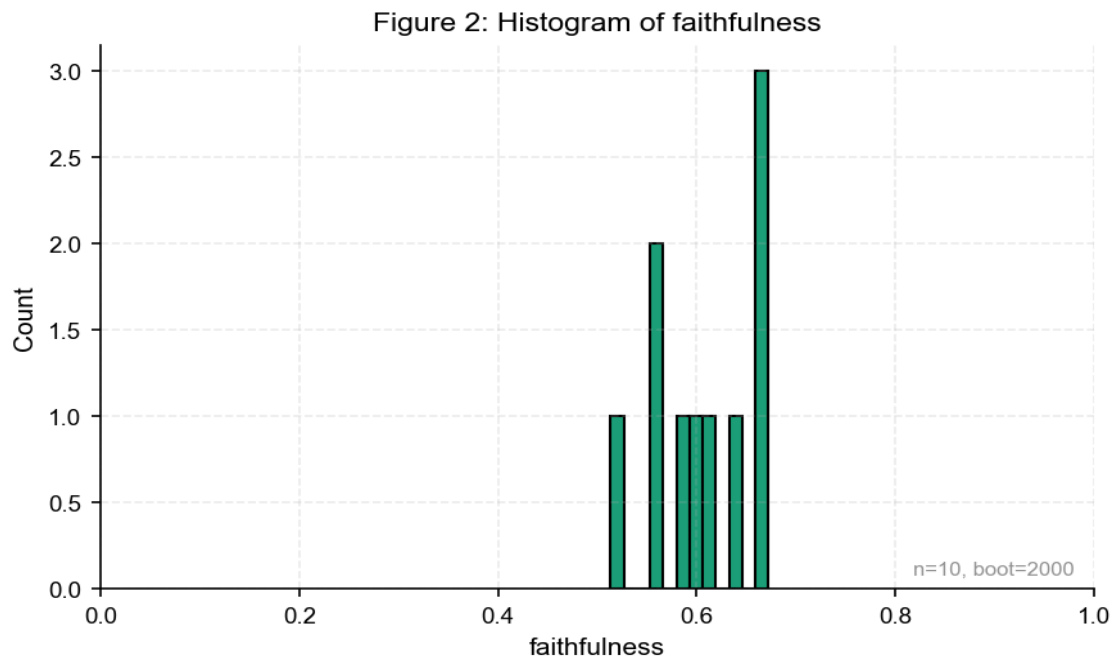
box violin faithfulness.png



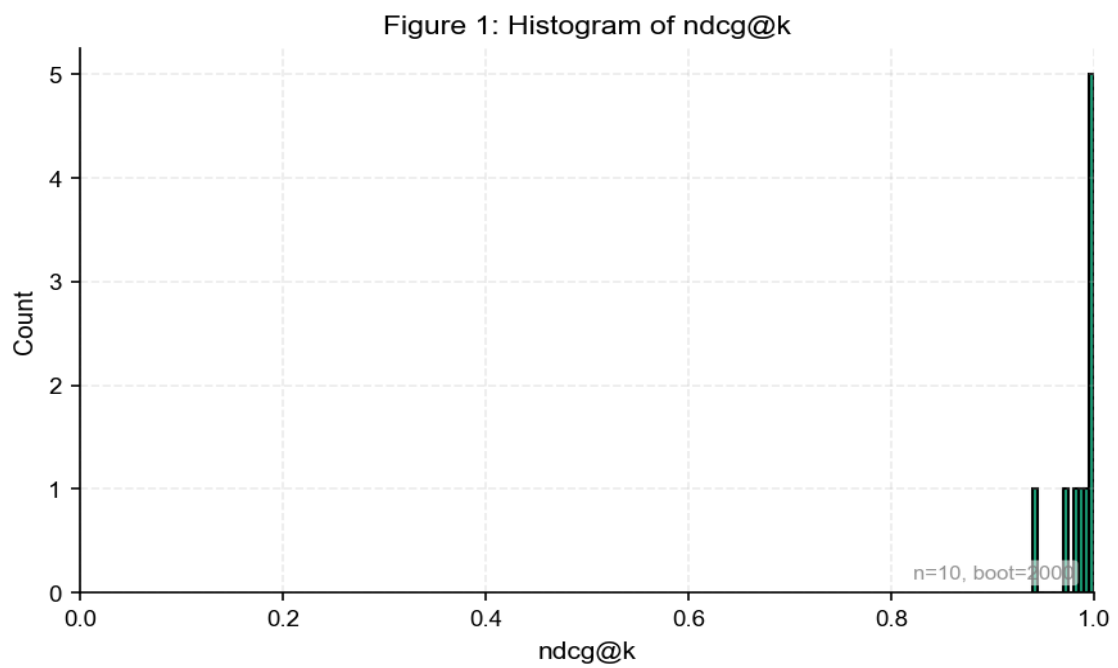
box violin ndcg@k.png



hist faithfulness.png

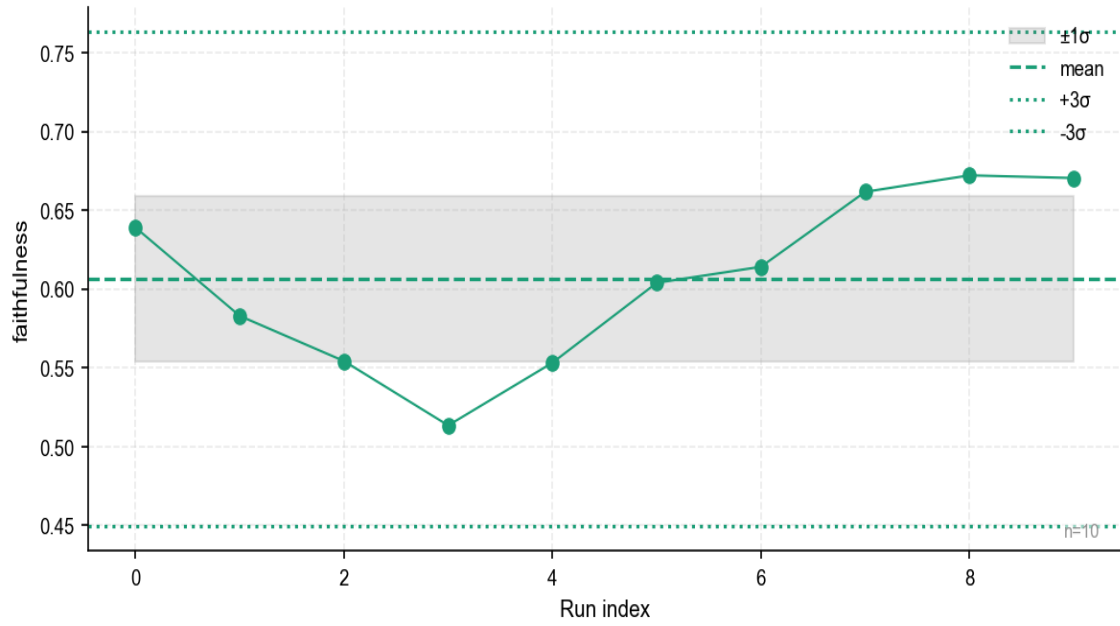


hist ndcg@k.png



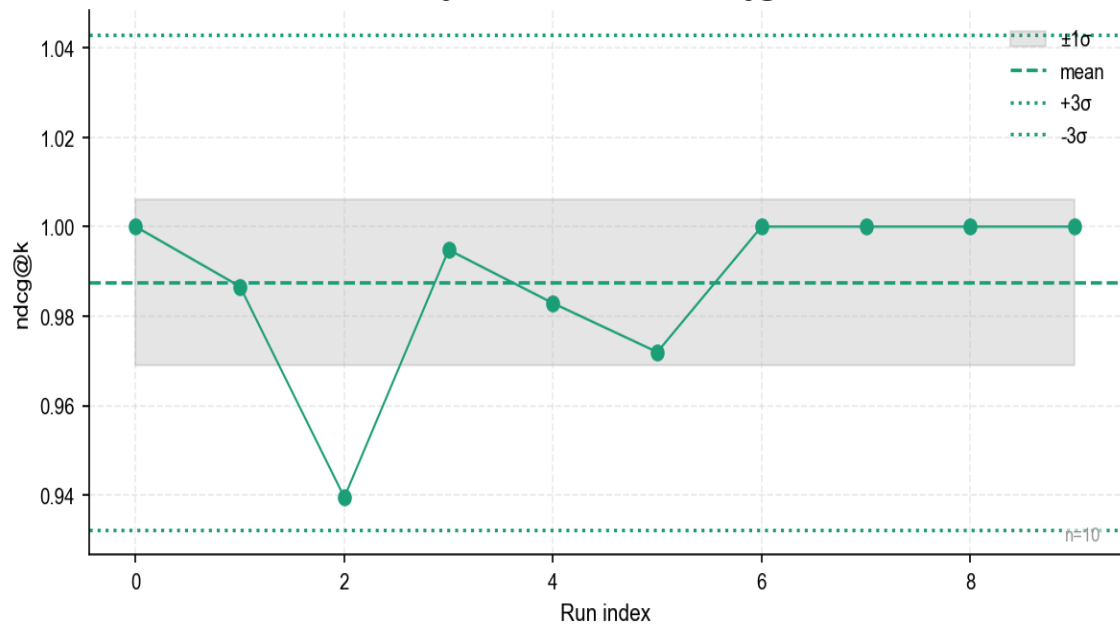
run order faithfulness.png

Figure 7: Run-order chart for faithfulness



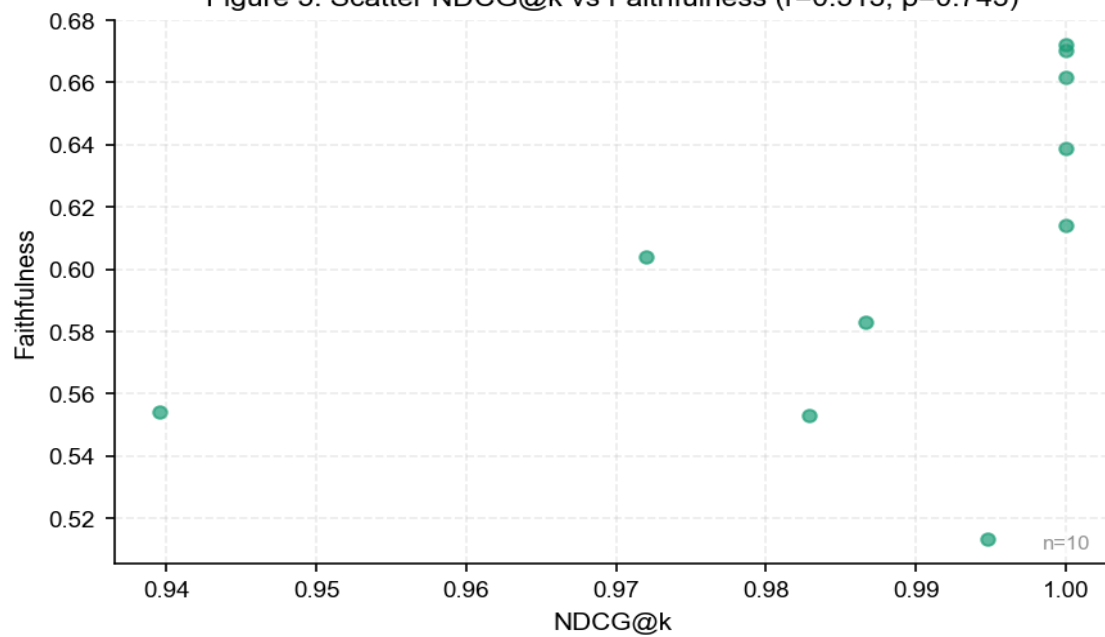
run order ndcg@k.png

Figure 6: Run-order chart for ndcg@k



scatter ndcg vs faithfulness.png

Figure 5: Scatter NDCG@k vs Faithfulness ($r=0.513$, $\rho=0.743$)



4. Interpretation and Next Steps

The retrieval architecture is in its target state. Remaining limitations are primarily in the Faithfulness level rather than in retrieval quality. Future experiments should contrast temporal weighting (`temporal_mode=True` vs. `False`) and perform decade-based query analyses after year-detection enhancement.

A convergence plot (mean $\pm 95\%$ CI vs. n) should be used to demonstrate statistical stabilization as sample size increases. All results are reproducible by rerunning the benchmark scripts with identical configurations.