

Benchmark Evaluation Report

Generated automatically on 2025-11-07 22:11

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	100.0000
ndcg@k mean	0.9951
ndcg@k ci95 lo	0.9921
ndcg@k ci95 hi	0.9980
faith mean	0.5698
faith ci95 lo	0.5571
faith ci95 hi	0.5833
ndcg@k median	1.0000
faith median	0.5691
ndcg@k std	0.0156
faith std	0.0681
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)	0.186	6.344e-02
Spearman (rank)	0.238	1.699e-02

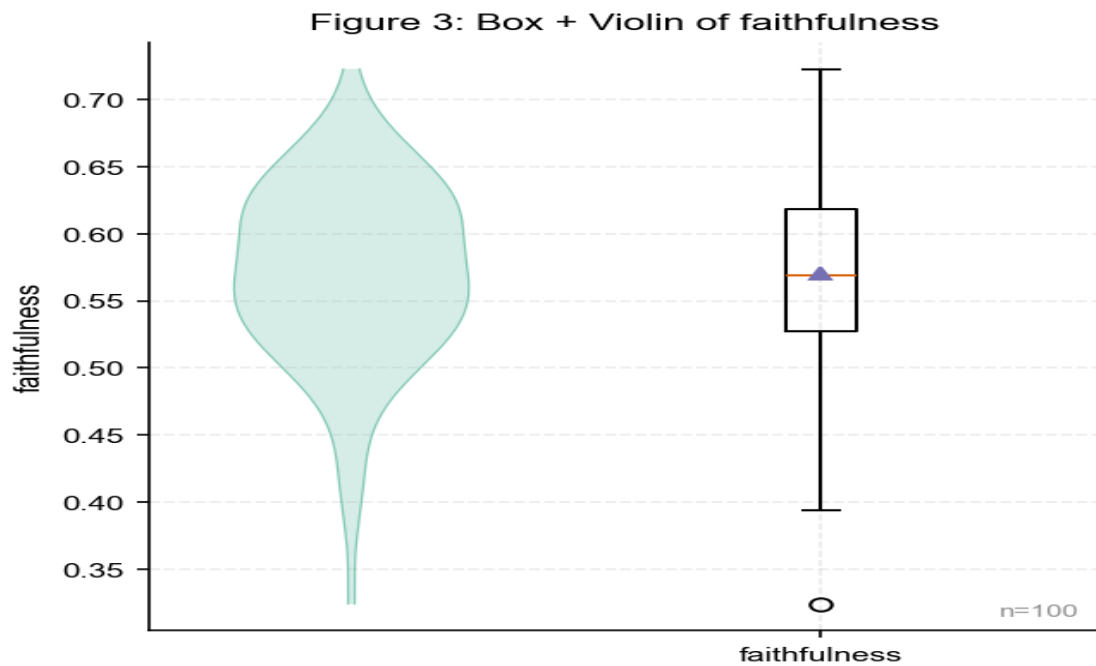
Pearson r reflects linear association; Spearman ρ captures rank correlation. Typical weak-to-moderate positive dependency indicates retrieval homogeneity, suggesting that Faithfulness could be further contrasted with semantic coherence metrics (e.g., BERTScore or FactScore).

Detected Outliers

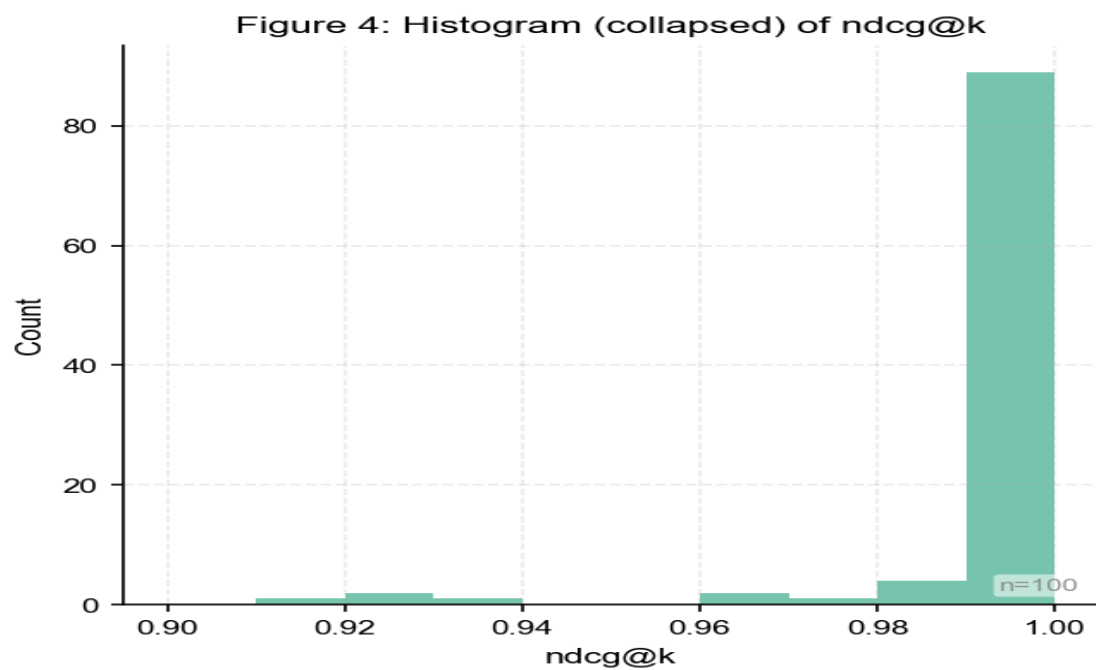
Query ID	NDCG@k	Faithfulness	z_ndcg	z_faith
Compare_shallow_learning_and_deep_hierar	0.961	0.456	-2.18	-1.67
Contrast_capability_framing_versus_ethic	0.915	0.523	-5.11	-0.69
Describe_how_fairness_and_bias_are_quant	1.000	0.324	0.30	-3.60
Discuss_how_symbolic_AI_and_connectionis	1.000	0.394	0.30	-2.58
Explain_knowledge_representation_in_expe	0.940	0.566	-3.55	-0.06
Explain_the_concept_of_heuristic_search_	1.000	0.424	0.30	-2.14
Explain_the_notion_of_machine-generated_	1.000	0.723	0.30	2.23
Summarize_how_bias_is_introduced_through	1.000	0.403	0.30	-2.44
Summarize_how_context_completeness_affec	0.924	0.562	-4.52	-0.12
Summarize_how_reinforcement_learning_ext	0.961	0.508	-2.18	-0.90
Summarize_how_search_algorithms_contribu	0.928	0.472	-4.32	-1.43

3. Visual Analytics

box violin faithfulness.png

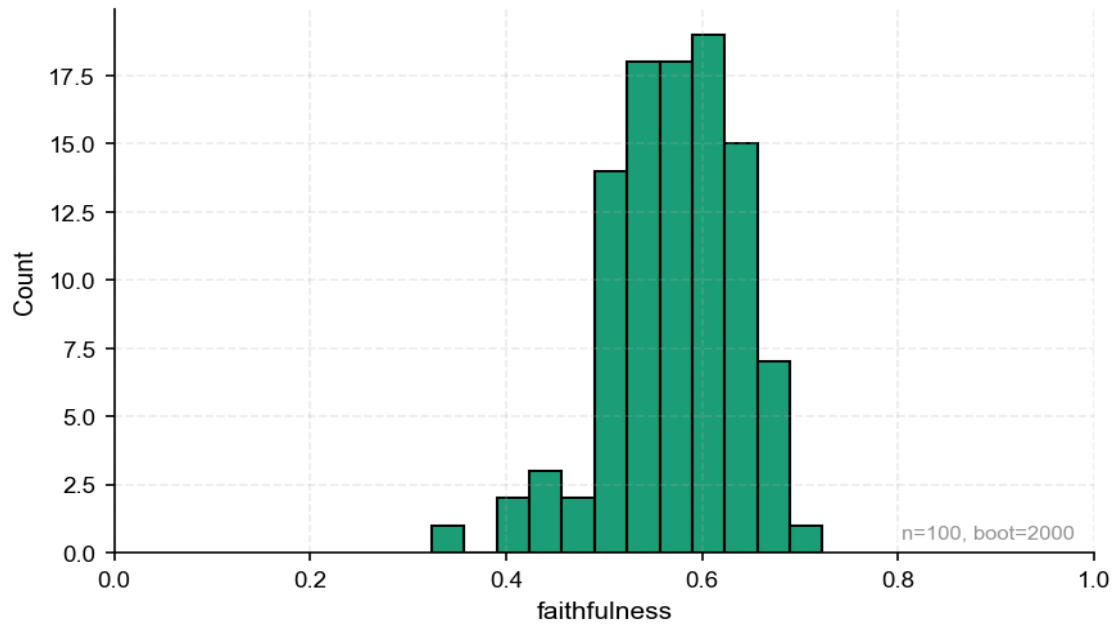


box violin ndcg@k.png



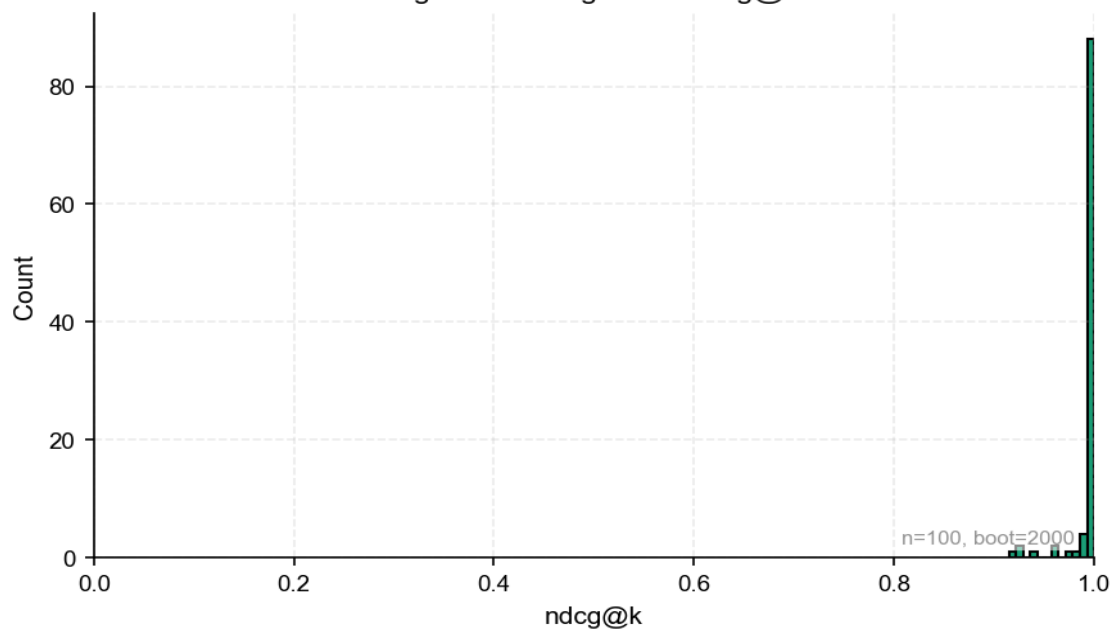
hist faithfulness.png

Figure 2: Histogram of faithfulness



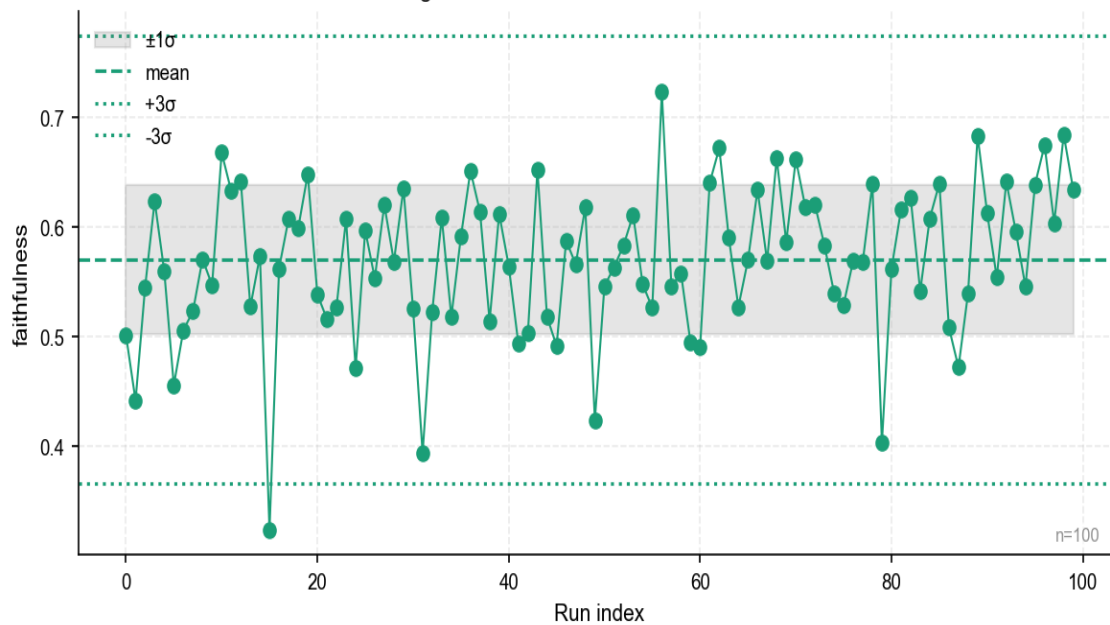
hist ndcg@k.png

Figure 1: Histogram of ndcg@k



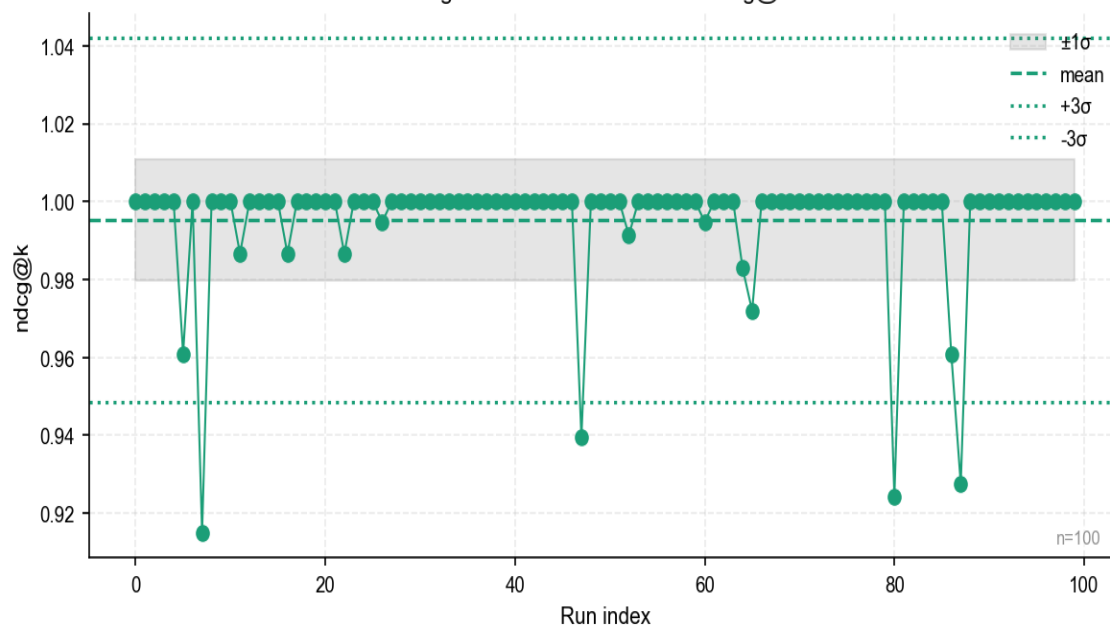
run order faithfulness.png

Figure 7: Run-order chart for faithfulness



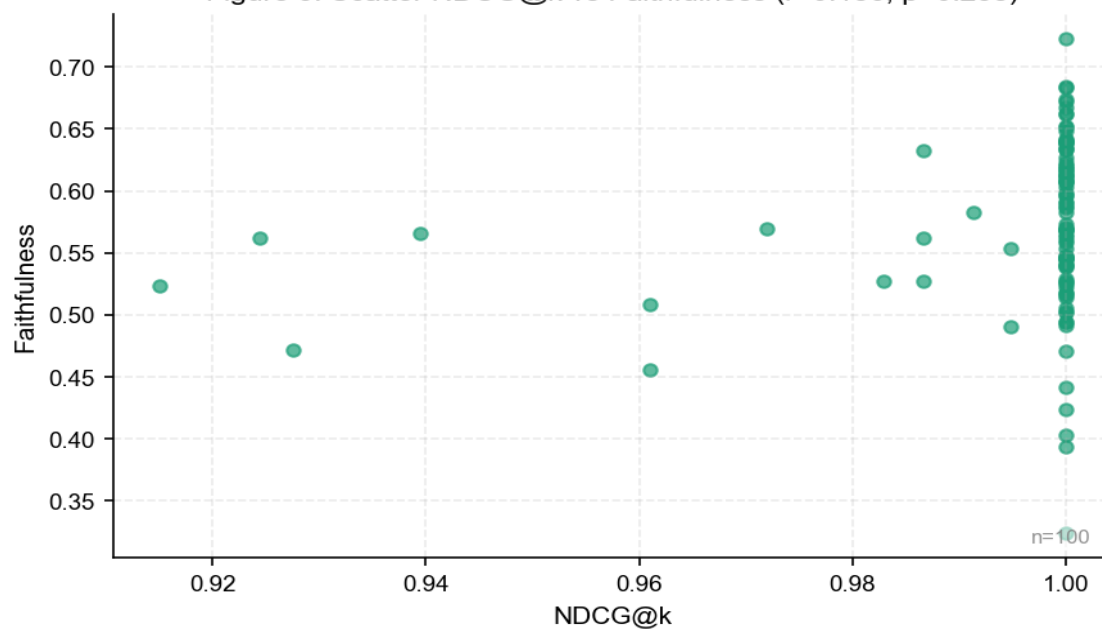
run order ndcg@k.png

Figure 6: Run-order chart for ndcg@k



scatter ndcg vs faithfulness.png

Figure 5: Scatter NDCG@k vs Faithfulness ($r=0.186$, $\rho=0.238$)



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.