Benchmark Evaluation Report

Generated automatically on 2025-11-07 21:27

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	30.0000
ndcg@k mean	0.9905
ndcg@k ci95 lo	0.9814
ndcg@k ci95 hi	0.9976
faith mean	0.5985
faith ci95 lo	0.5757
faith ci95 hi	0.6209
ndcg@k median	1.0000
faith median	0.6195
ndcg@k std	0.0225
faith std	0.0647
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)	0.575	8.775e-04
Spearman (rank)	0.629	1.955e-04

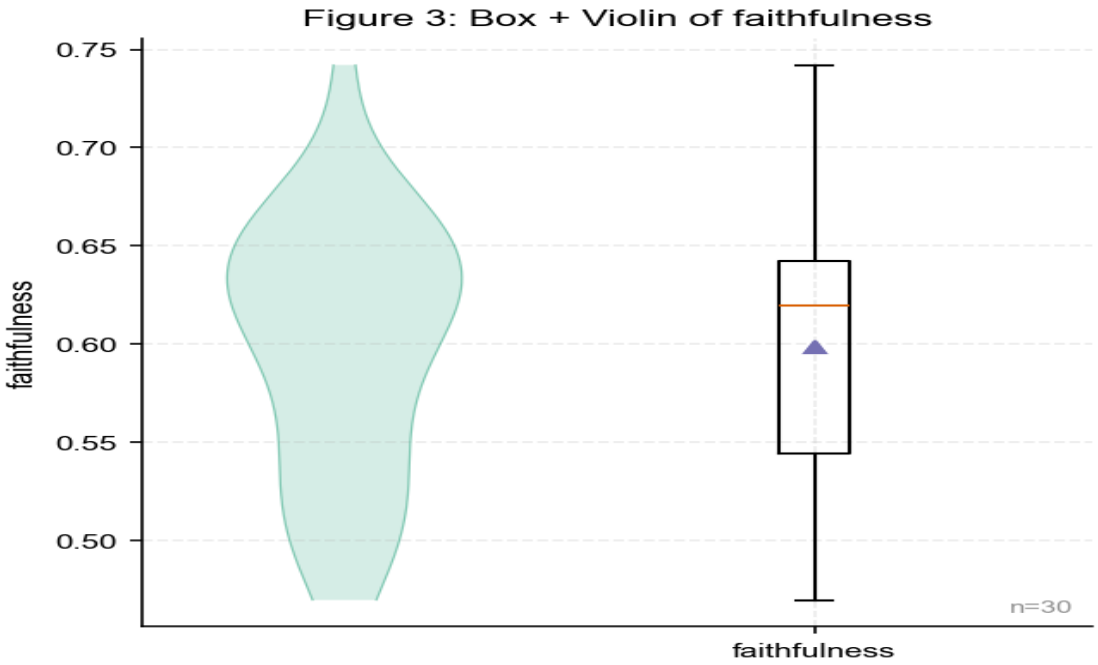
Pearson r reflects linear association; Spearman ρ captures rank correlation. Typical weak-to-moderate positive dependency indicates retrieval homogeneity, suggesting that Faithfulness could be further contrasted with semantic coherence metrics (e.g., BERTScore or FactScore).

Detected Outliers

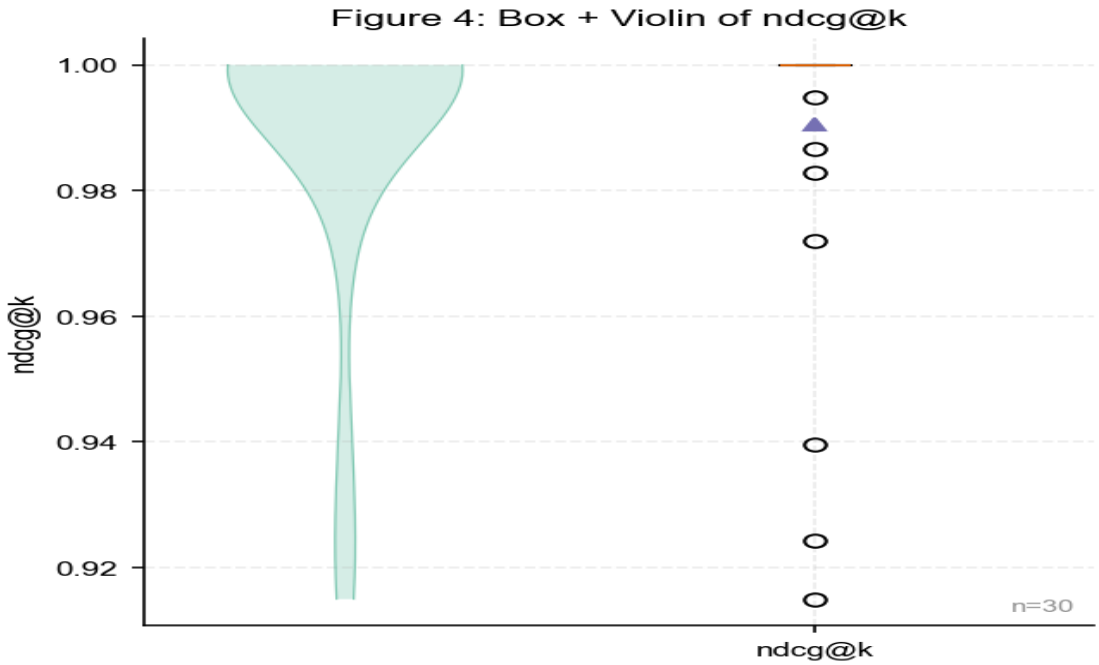
Query ID	NDCG@k	Faithfulness	z_ndcg	z_faith
Contrast_capability_framing_versus_ethic	0.915	0.469	-3.30	-1.96
Explain_knowledge_representation_in_expe	0.940	0.532	-2.23	-1.01
Explain_the_notion_of_machine-generated_	1.000	0.742	0.41	2.18
Summarize_how_context_completeness_affec	0.924	0.506	-2.89	-1.41

3. Visual Analytics

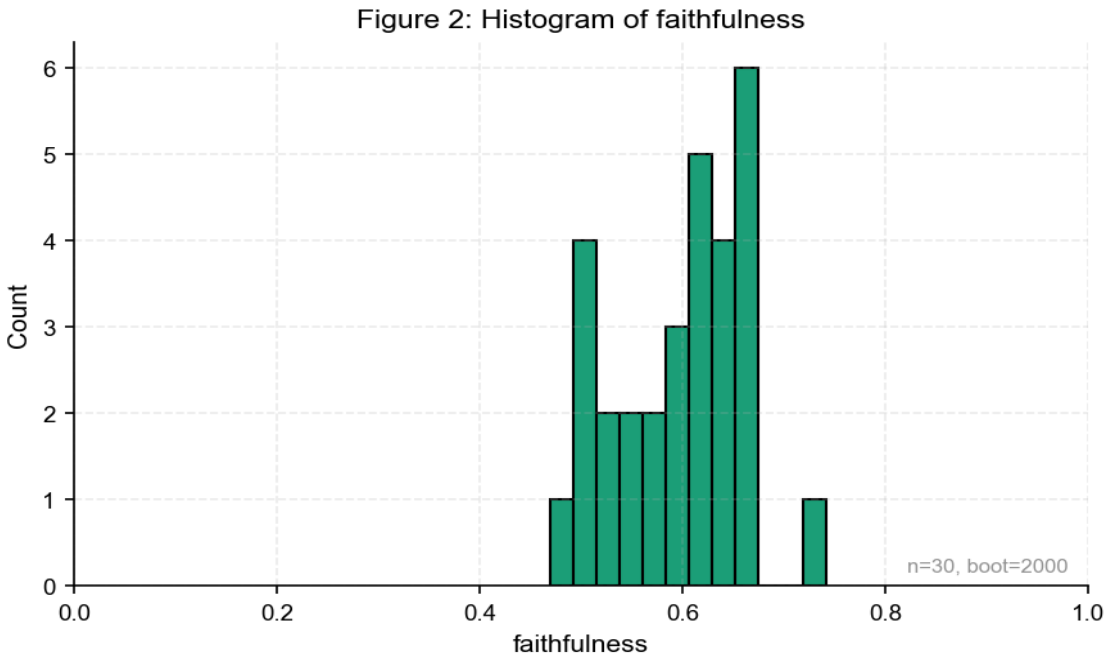
box violin faithfulness.png



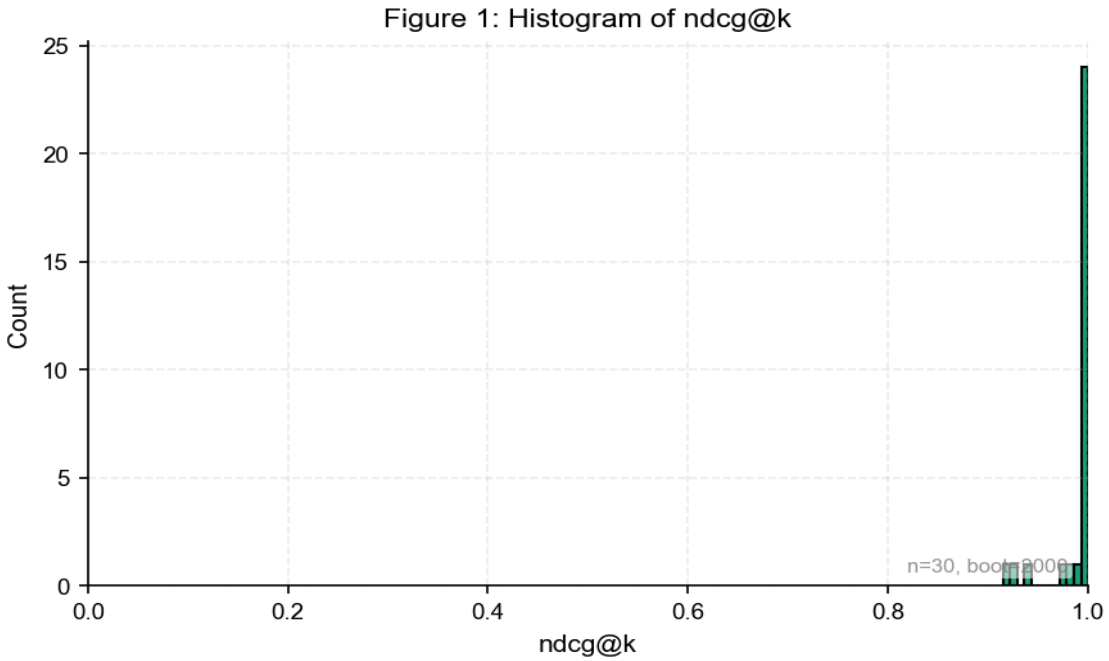
box violin ndcg@k.png



hist faithfulness.png

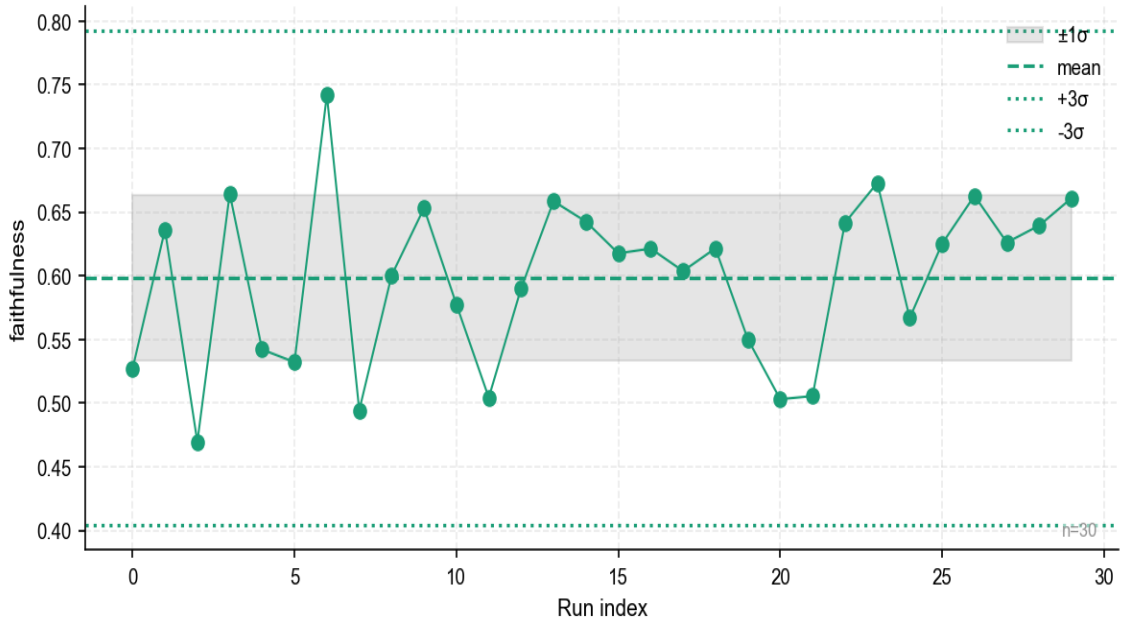


hist ndcg@k.png



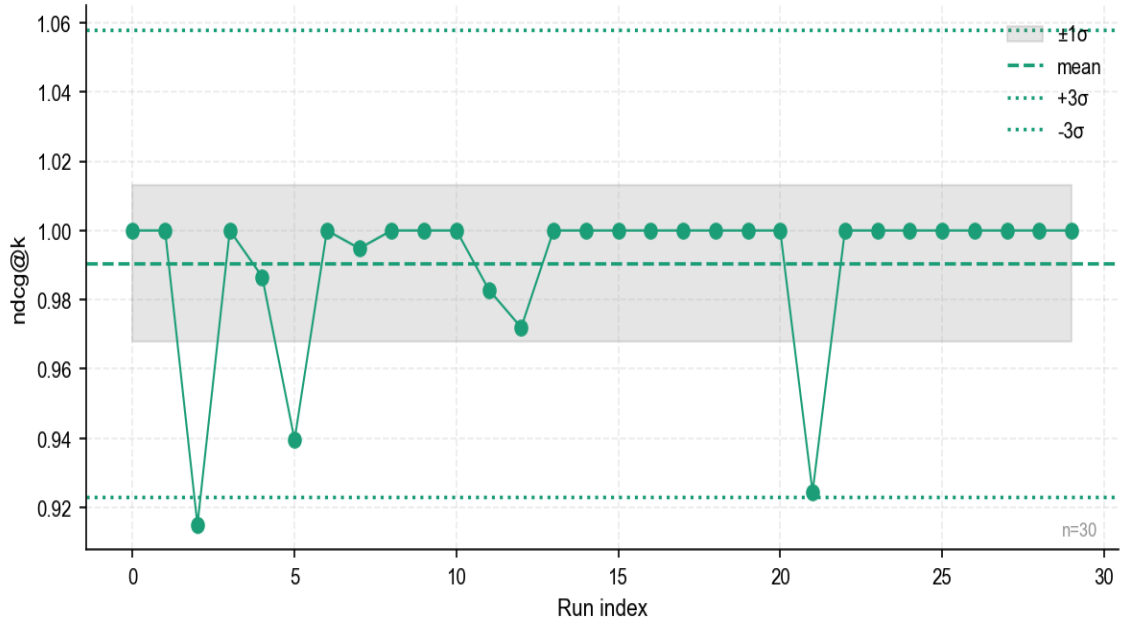
run order faithfulness.png

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

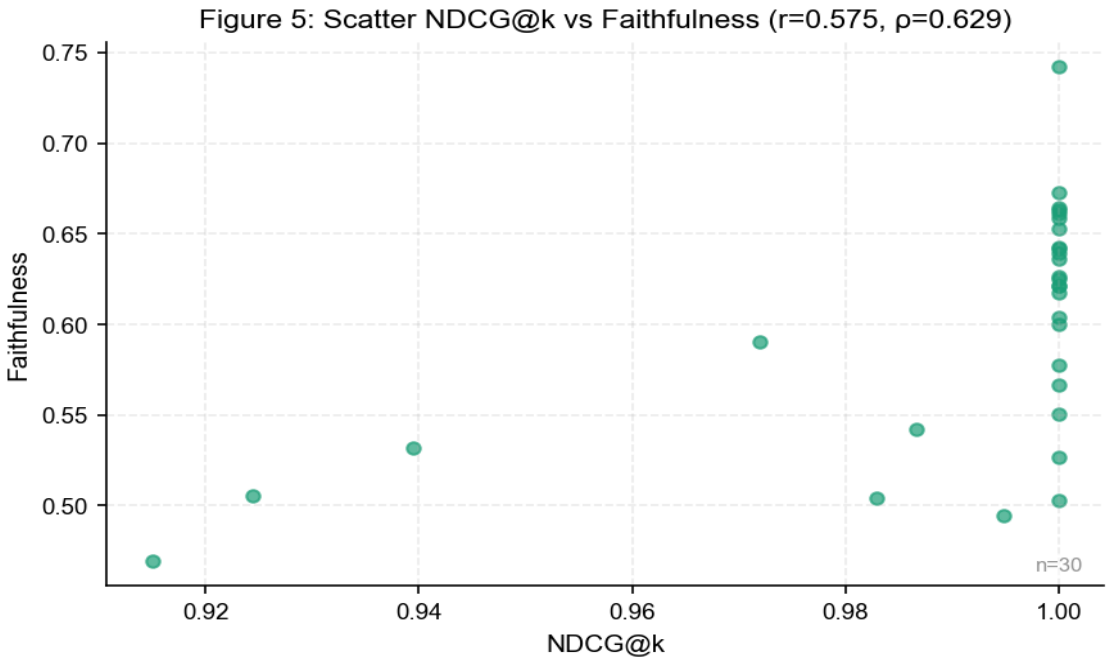


Figure 5: Scatter NDCG@k vs Faithfulness (r=0.575, ρ=0.629)
Faithfulness
0.75
0.70
0.65
0.60
0.55
0.50
0.92
0.94
0.96
0.98
1.00
NDCG@k
n=30

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 ±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.