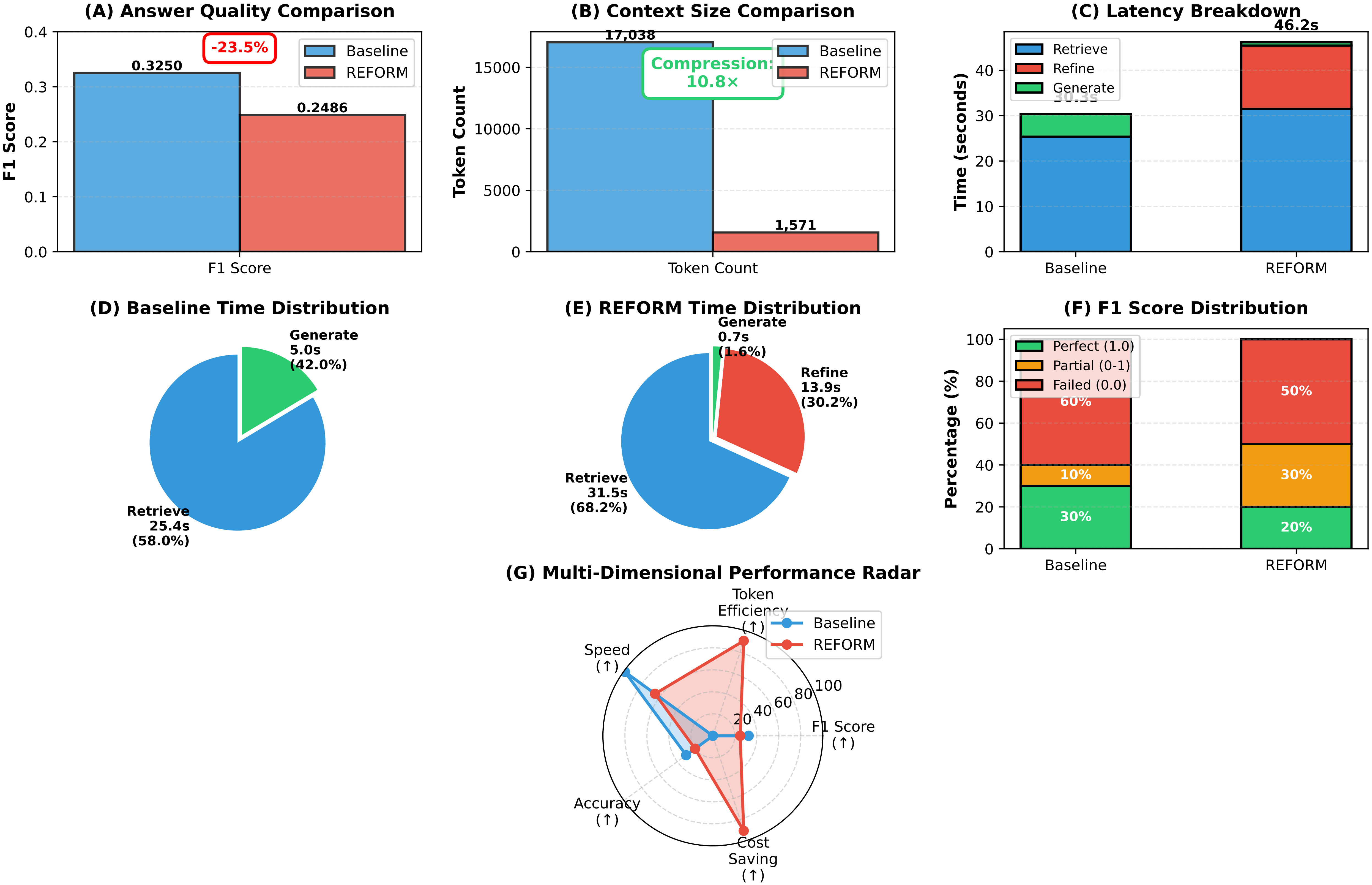
REFORM vs Baseline RAG: Comprehensive Performance Analysis
Dataset: Natural Questions (10 samples) | Model: Llama-3.1-8B-Instruct | Retrieval: Wiki18 FAISS (Top-8)


(A) Answer Quality Comparison
F1 Score
Baseline
REFORM
F1 Score
(B) Context Size Comparison
Token Count
Compression:
10.8×
Baseline
REFORM
Token Count
(C) Latency Breakdown
Time (seconds)
46.2s
30.3s
Retrieve
Refine
Generate
Baseline
REFORM
(D) Baseline Time Distribution
Generate
5.0s
(42.0%)
Retrieve
25.4s
(58.0%)
(E) REFORM Time Distribution
Generate
0.7s
(1.6%)
Refine
13.9s
(30.2%)
Retrieve
31.5s
(68.2%)
(F) F1 Score Distribution
Percentage (%)
Perfect (1.0)
Partial (0-1)
Failed (0.0)
Baseline
REFORM
(G) Multi-Dimensional Performance Radar
Token
Efficiency
(↑)
Speed
(↑)
F1 Score
(↑)
Accuracy
(↑)
Cost
Saving
(↑)
Baseline
REFORM