REFORM vs Baseline RAG: Detailed Performance Comparison

Dataset: Natural Questions (10 samples) | Model: Llama-3.1-8B-Instruct | Retrieval: Wiki18 FAISS (Top-8)

Metric	Baseline RAG	REFORM RAG	Change	Interpretation
Performance Metrics				
F1 Score	0.3250 (32.5%)	0.2486 (24.86%)	**-23.5%**	❌ Accuracy decreased
Token Count	17,038 tokens	1,571 tokens	**-90.8%**	✅ Massive reduction
Compression Rate	1.0×	10.81×	**+981%**	✅ 10.8× compression
Latency Breakdown				
Retrieve Time	25.36s (58.0%)	31.49s (68.2%)	**+24.2%**	❌ Slower retrieval
Refine Time	0.00s (0.0%)	13.94s (30.2%)	**+∞**	❌ New overhead
Generate Time	4.98s (42.0%)	0.73s (1.6%)	**-85.3%**	✅ Much faster
Total Latency	30.33s	46.16s	**+52.2%**	❌ Slower overall
Answer Quality Distribution				
Perfect (F1=1.0)	3/10 (30%)	2/10 (20%)	**-33%**	❌ Fewer perfect answers
Partial (0<F1<1)	1/10 (10%)	3/10 (30%)	**+200%**	⚠ More partial matches
Failed (F1=0)	6/10 (60%)	5/10 (50%)		

Key Findings:
• REFORM achieves 10.8× compression but sacrifices 23.5% F1 score
• Generate time reduced by 85.3% (4.98s → 0.73s)
• Total latency increased by 52.2% due to Refine overhead (13.94s)
• Trade-off: Token efficiency ↑ vs Accuracy ↓ and Speed ↓

Positive Change | Negative Change | Baseline Values | REFORM Values