Adaptiq vs GPT

Comprehensive Performance Analysis

Generated on: 2025-08-13 17:24:04

Benchmark Configuration

```
BENCHMARK CONFIGURATION
================================================

LLM/VLM Model : GPT-4.1
Image Model   : black-forest-labs/flux-1.1-pro
Test Run ID   : 20250813_100138
Generated     : 2025-08-13 17:24:04

SAMPLE SIZES
================================================
Total Results      : 199 (AdaptiQ + GPT pairs)
Image Pairs        : 99 pairs
Cost/Latency Data  : N = 99 pairs (199 individual results)
CLIP Quality Data  : N = 93 pairs

DATA QUALITY NOTES
================================================
Missing CLIP scores: 6 pairs
Reason: API errors or invalid image formats

Notes on sample sizes:
- Cost/Latency/Token data captured for 100 image pairs.
- CLIP quality scores were successfully computed for 93 pairs due to
  intermittent API errors or invalid image formats.
- All per-pair comparisons are limited to the 93 pairs with complete data.
```

Performance Summary & Key Insights

Performance Statistics

```
Adaptiq:
  Avg Latency: 11.85±2.69s
  Avg Tokens: 7459±457
  Avg Cost: $0.0086
  Total Cost: $0.8578

GPT:
  Avg Latency: 13.94±3.33s
  Avg Tokens: 8347±1278
  Avg Cost: $0.0099
  Total Cost: $0.9826
```

Key Metrics Comparison

```
ADAPTIQ vs GPT (Δ = AdaptiQ - GPT):

Latency: -2.09s (-15.0%)
Cost: $-0.0013 (-13.6%)
Tokens: -887 (-10.6%)

Interpretation:
Negative cost/latency = AdaptiQ better
Positive quality = AdaptiQ better
```

Image Quality Summary

```
CLIP Score Winners (93 images):
  Adaptiq Wins: 31 (33.3%)
  GPT Wins: 41 (44.1%)
  Ties: 21 (22.6%)

Average CLIP Scores:
  Adaptiq: 91.010
  GPT: 91.182
```

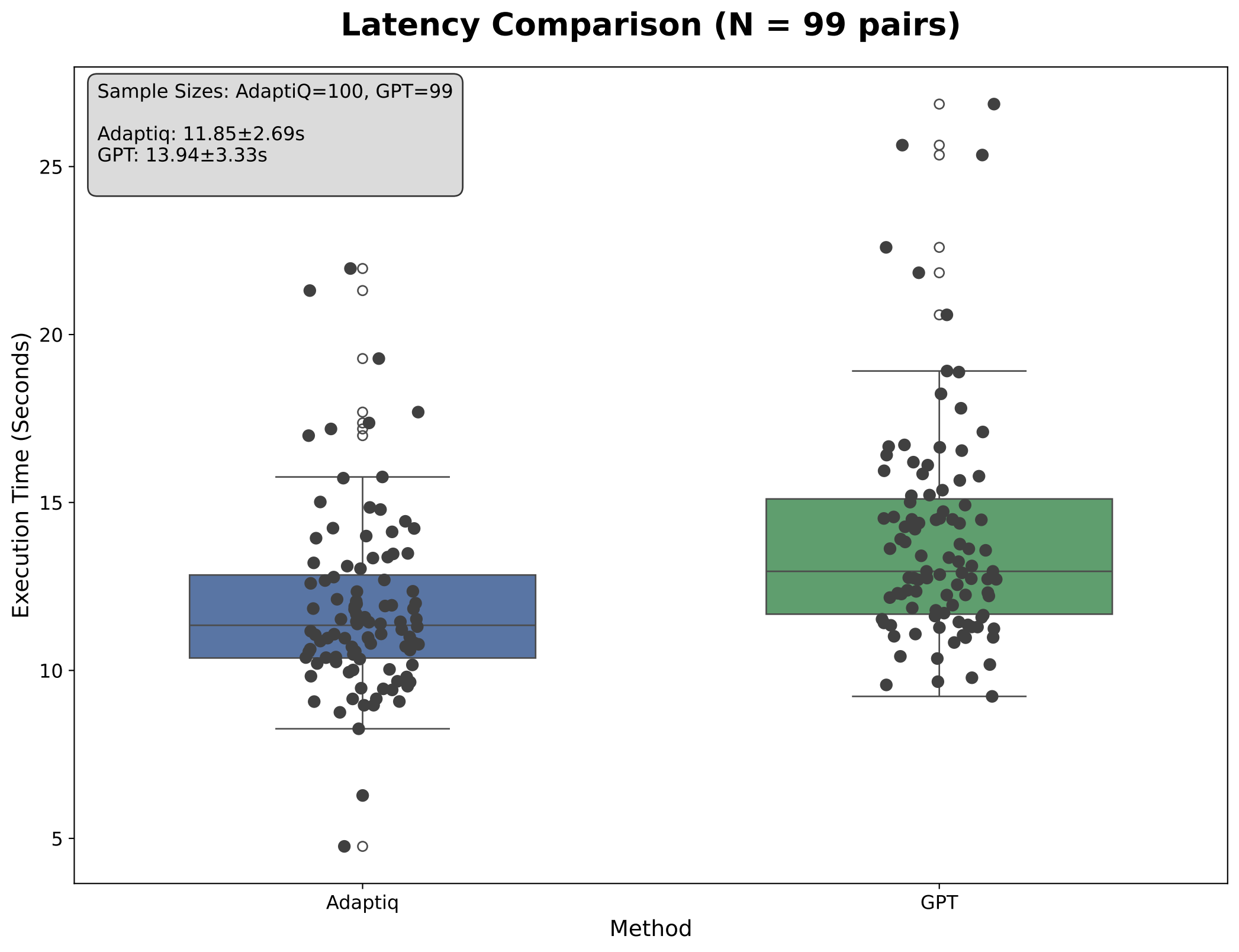
Key Insights

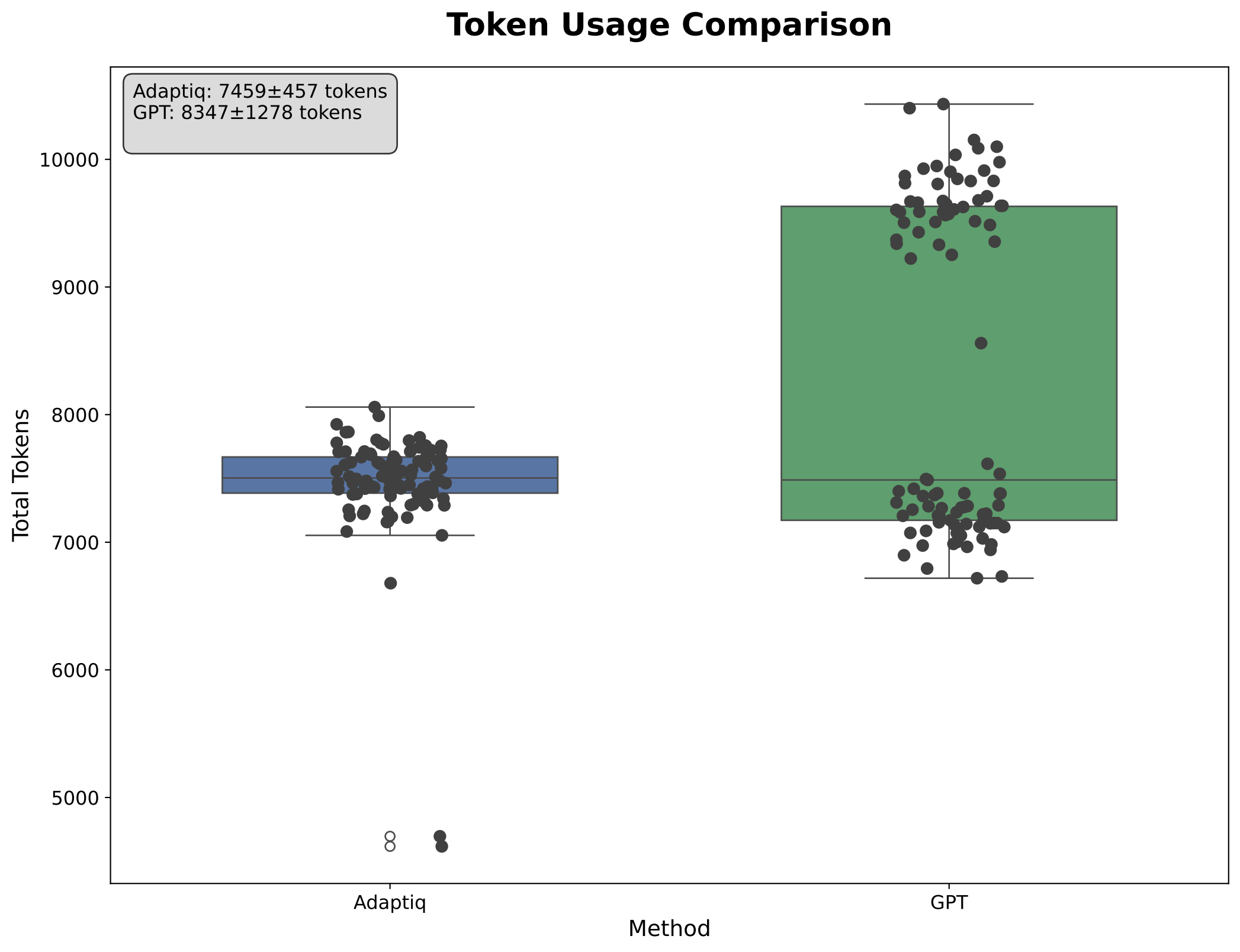
KEY FINDINGS:

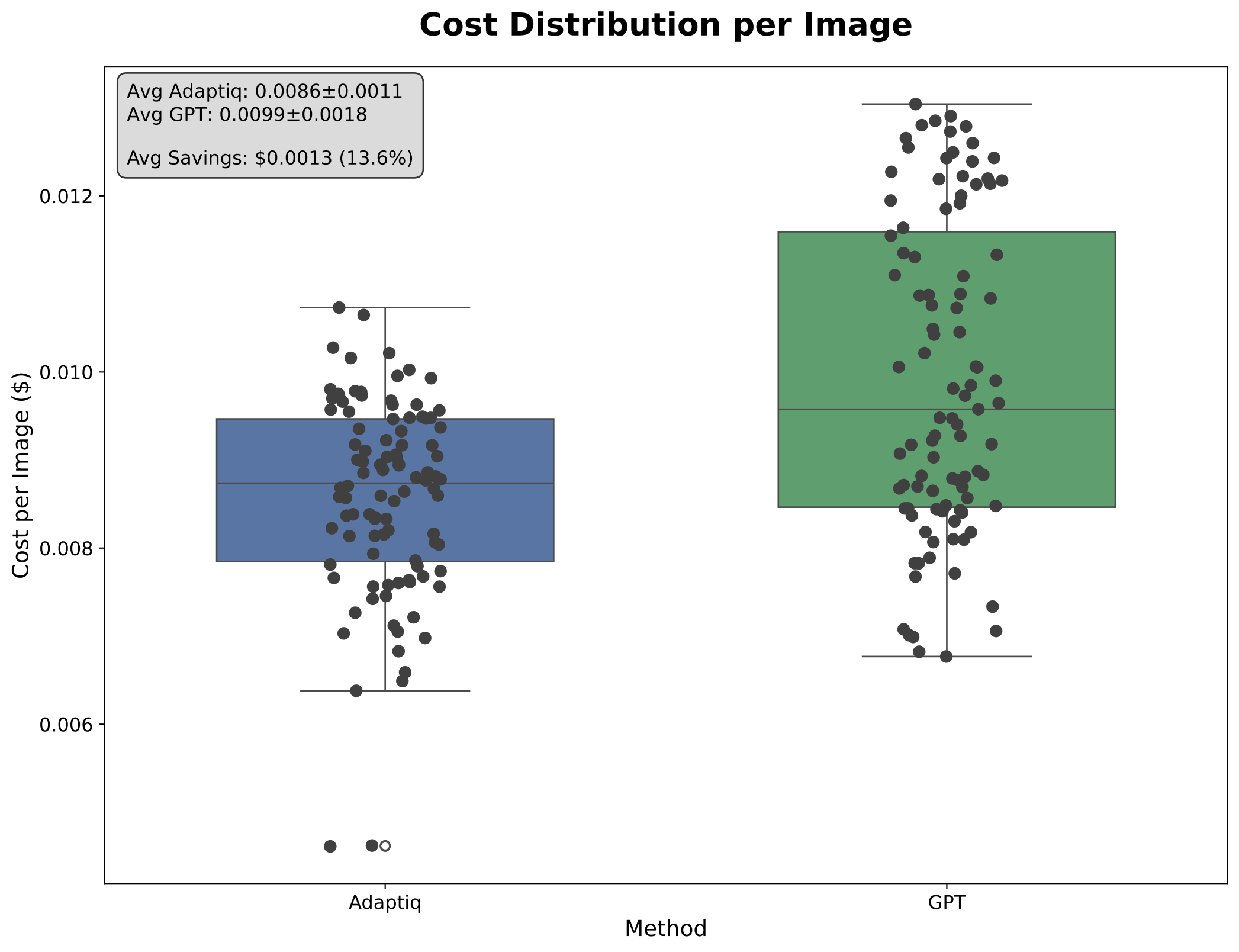
 Adaptiq is 12.7% more cost-effective.

 GPT produces higher quality images more often.

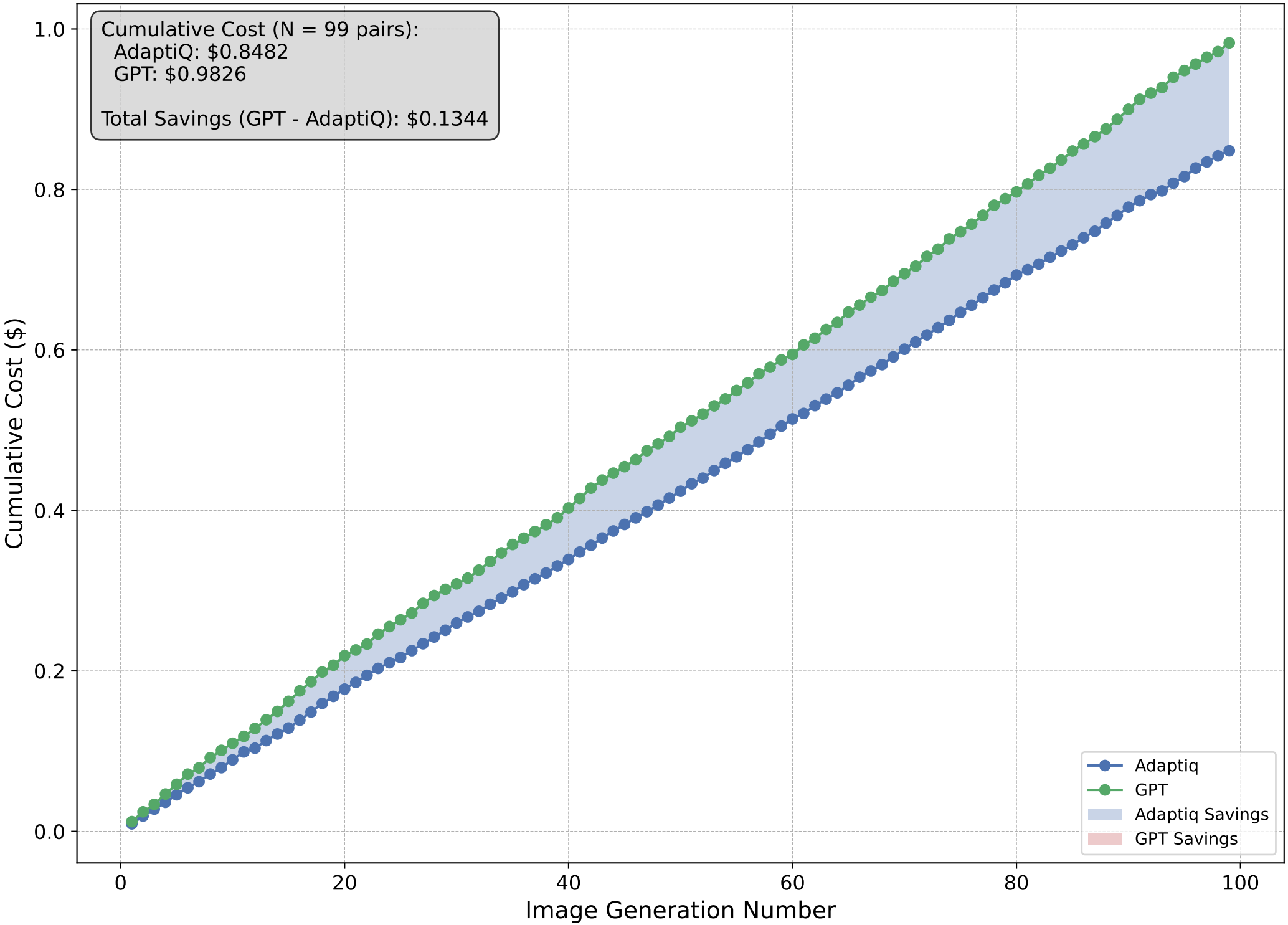
 See individual pages for detailed analysis.

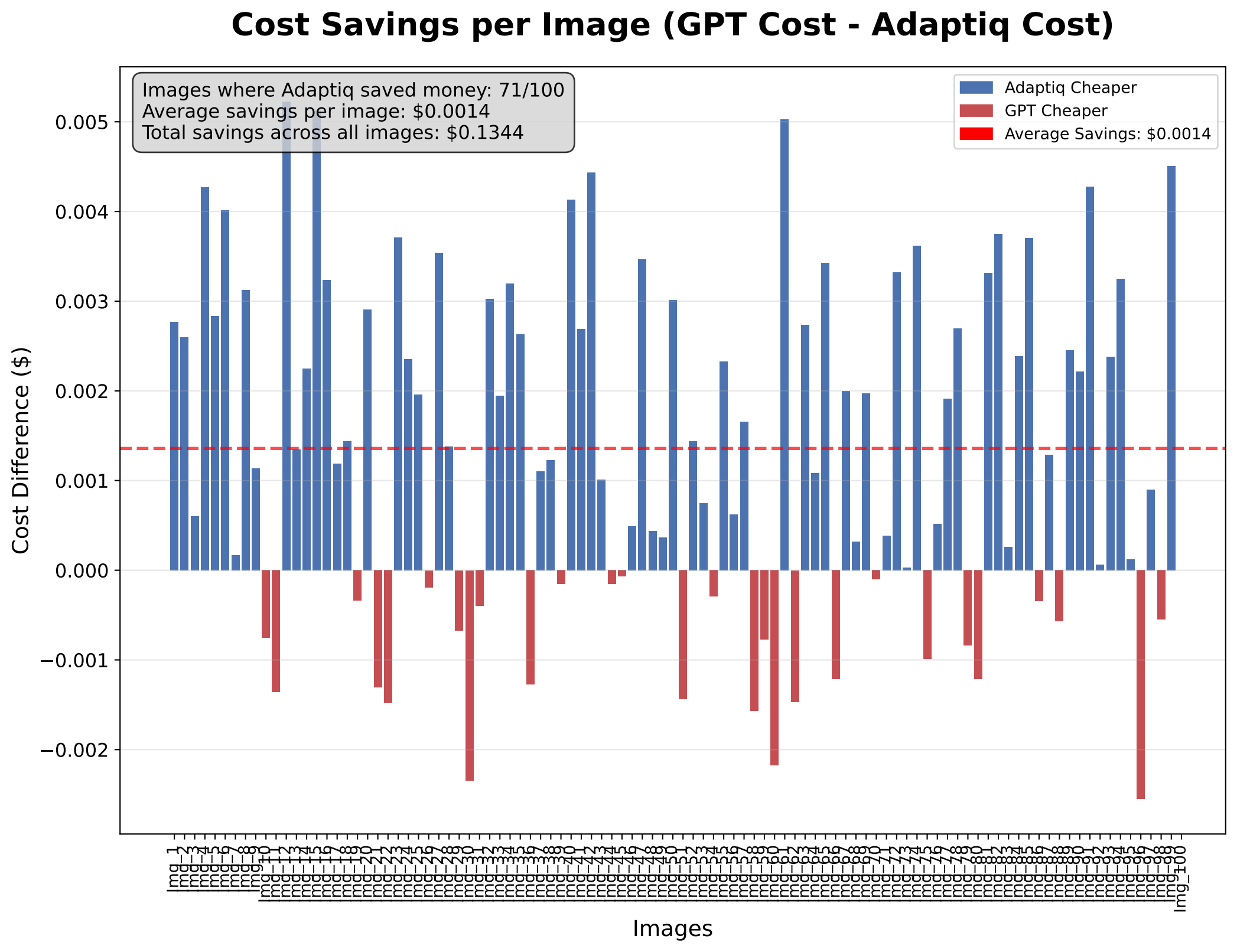

Latency Comparison (N = 99 pairs)
Sample Sizes: AdaptiQ=100, GPT=99
Adaptiq: 11.85±2.69s
GPT: 13.94±3.33s
Execution Time (Seconds)
25
20
15
10
5
Adaptiq
GPT
Method

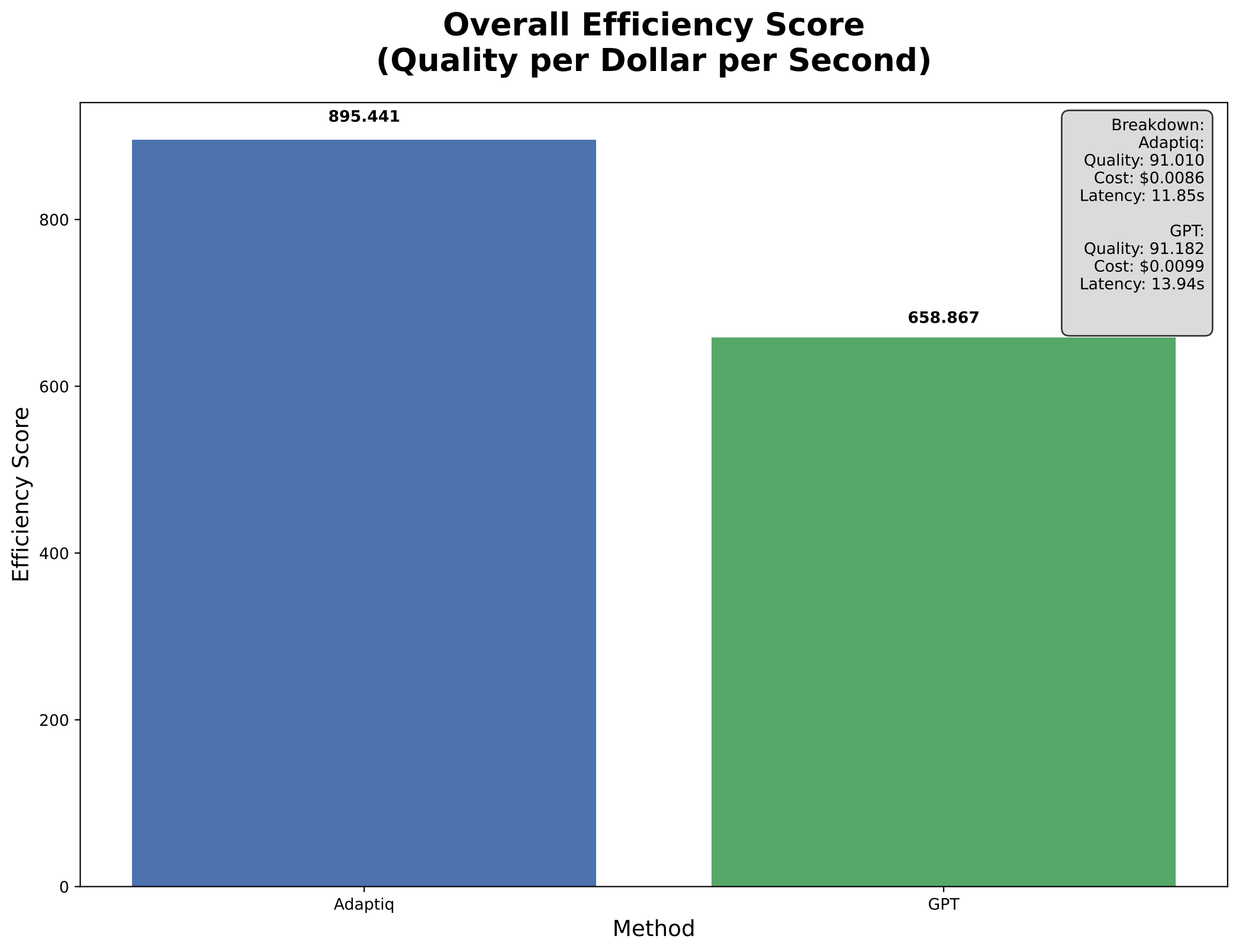

Token Usage Comparison
Adaptiq: 7459±457 tokens
GPT: 8347±1278 tokens
10000
9000
8000
7000
6000
5000
Total Tokens
Adaptiq
GPT
Method

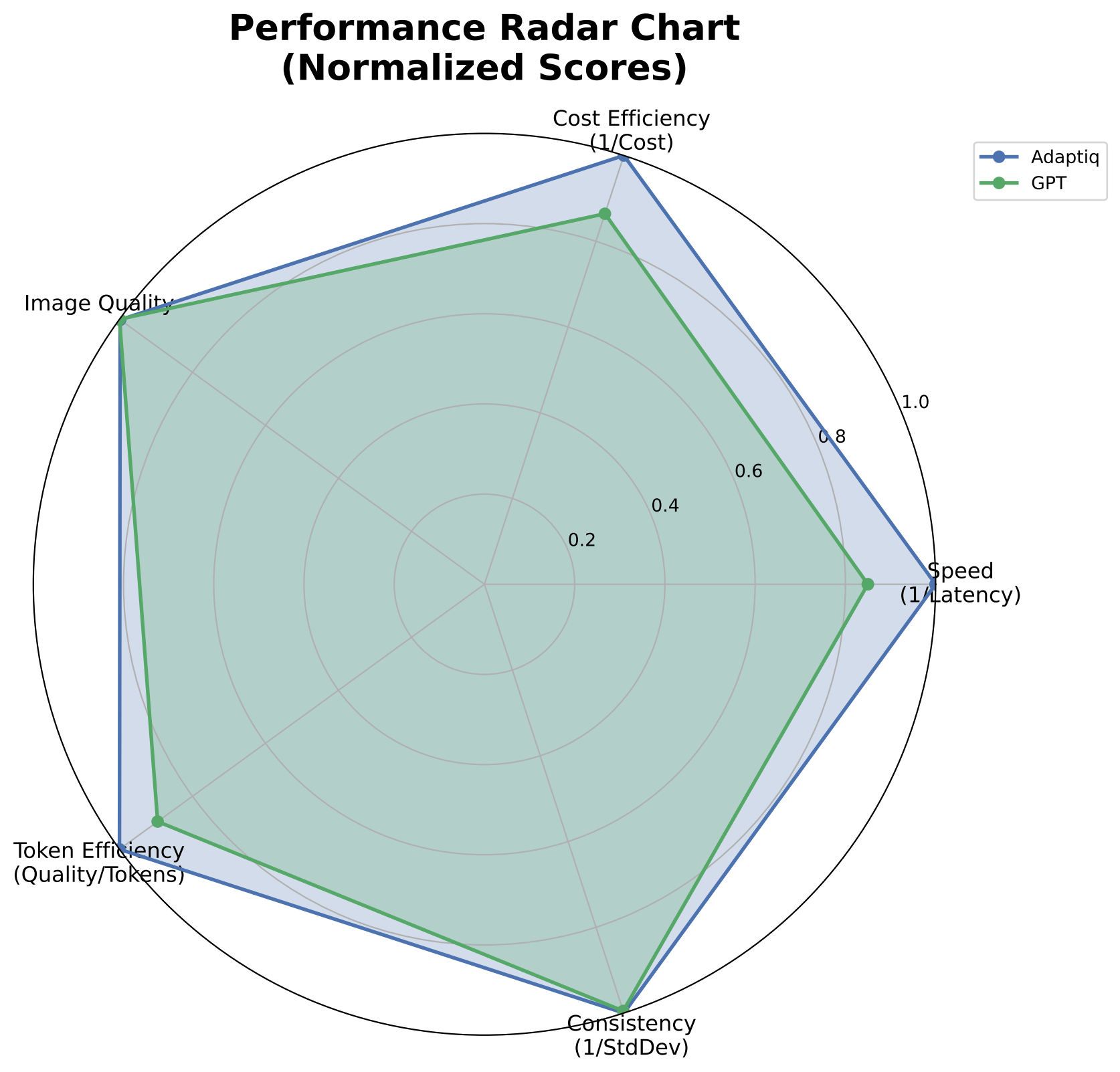

Cost Distribution per Image
Avg Adaptiq: 0.0086±0.0011
Avg GPT: 0.0099±0.0018
Avg Savings: $0.0013 (13.6%)
0.012
0.010
0.008
0.006
Cost per Image ($)
Adaptiq
GPT
Method

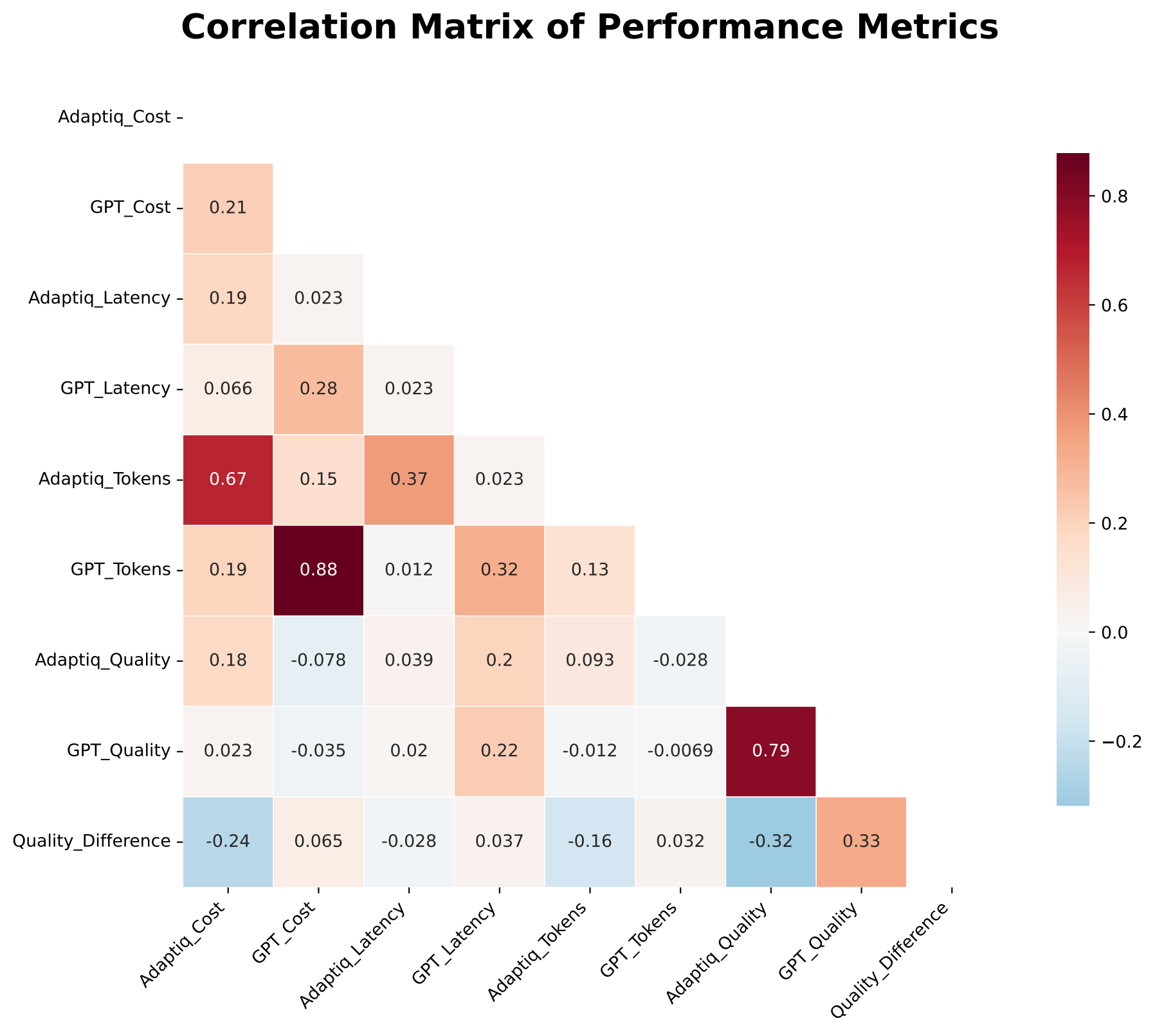
Cumulative Cost Over Time (N = 99 pairs)




Cost Savings per Image (GPT Cost - Adaptiq Cost)
Images where Adaptiq saved money: 71/100
Average savings per image: $0.0014
Total savings across all images: $0.1344
Adaptiq Cheaper
GPT Cheaper
Average Savings: $0.0014
Cost Difference ($)
0.005
0.004
0.003
0.002
0.001
0.000
−0.001
−0.002
Images

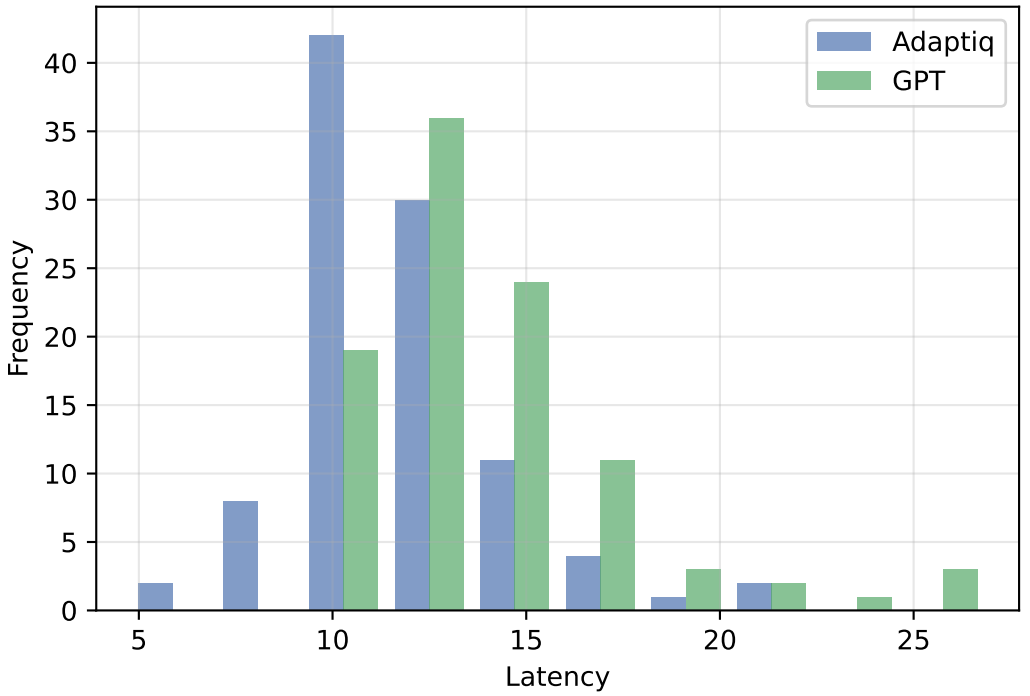

Overall Efficiency Score
(Quality per Dollar per Second)
895.441
658.867
Breakdown:
Adaptiq:
Quality: 91.010
Cost: $0.0086
Latency: 11.85s
GPT:
Quality: 91.182
Cost: $0.0099
Latency: 13.94s
Efficiency Score
0
200
400
600
800
Adaptiq
GPT
Method


Performance Radar Chart
(Normalized Scores)
Cost Efficiency
(1/Cost)
Image Quality
Speed
(1/Latency)
Token Efficiency
(Quality/Tokens)
Consistency
(1/StdDev)
0.2
0.4
0.6
0.8
1.0
Adaptiq
GPT

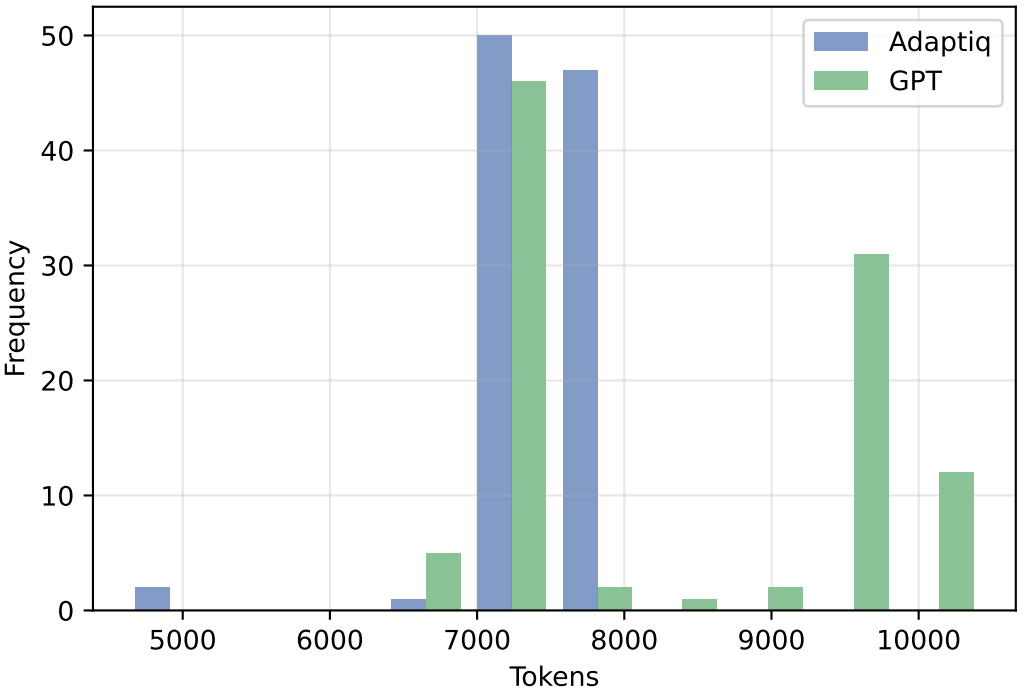

Correlation Matrix of Performance Metrics
Adaptiq_Cost
GPT_Cost
Adaptiq_Latency
GPT_Latency
Adaptiq_Tokens
GPT_Tokens
Adaptiq_Quality
GPT_Quality
Quality_Difference
0.21
0.19 0.023
0.066 0.28 0.023
0.67 0.15 0.37 0.023
0.19 0.88 0.012 0.32 0.13
0.18 -0.078 0.039 0.2 0.093 -0.028
0.023 -0.035 0.02 0.22 -0.012 -0.0069 0.79
-0.24 0.065 -0.028 0.037 -0.16 0.032 -0.32 0.33
0.8
0.6
0.4
0.2
0.0
−0.2

Distribution Comparison: Adaptiq vs GPT

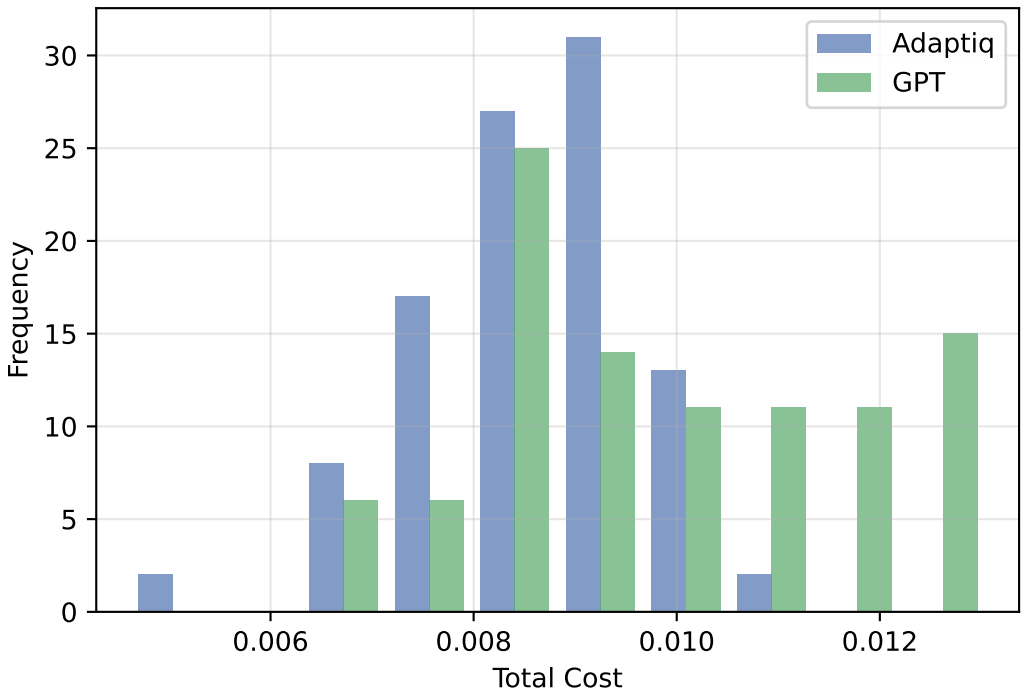
Latency Distribution (s)



Token Usage Distribution



Cost Distribution ($)



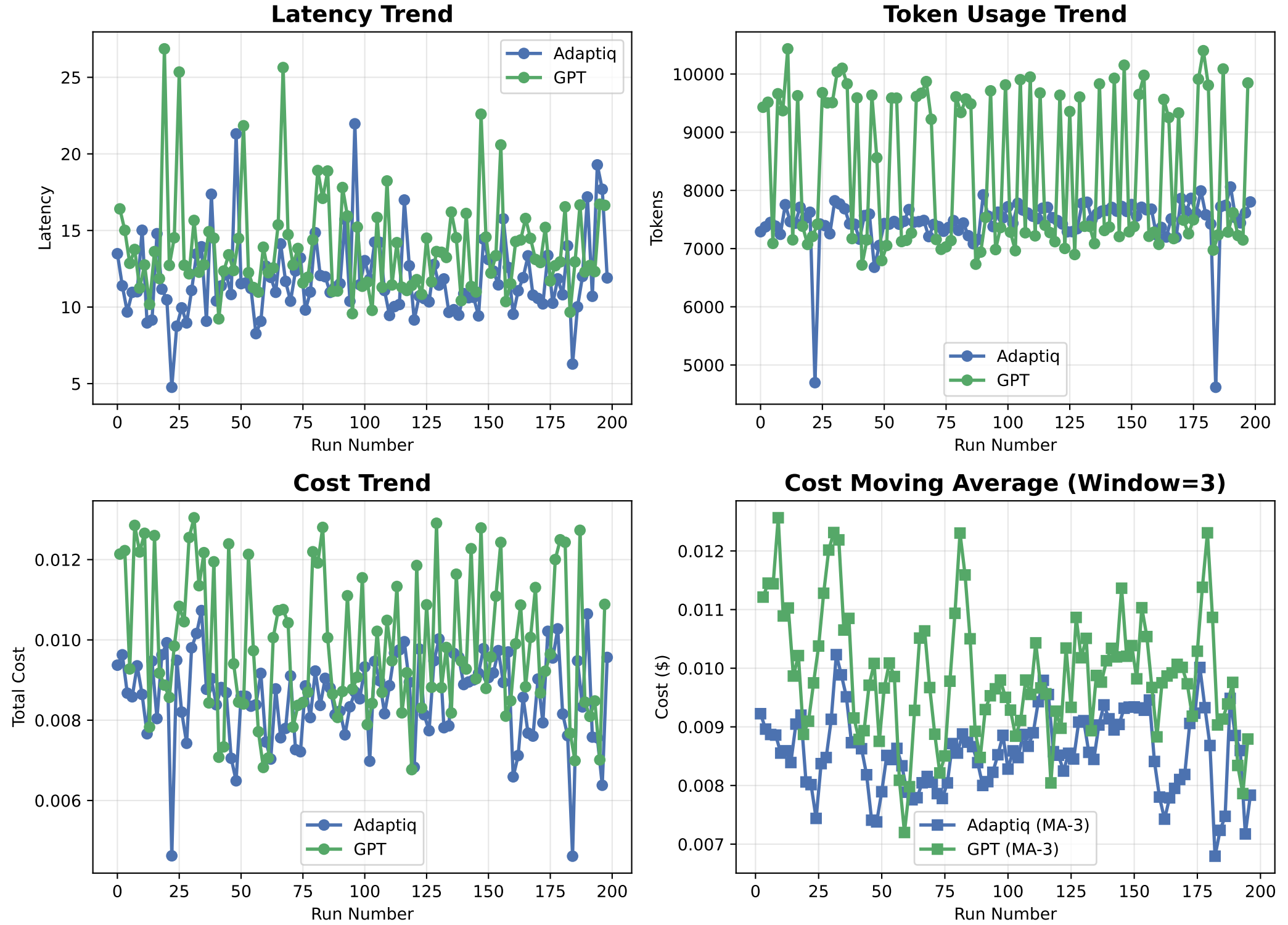
Statistical Tests

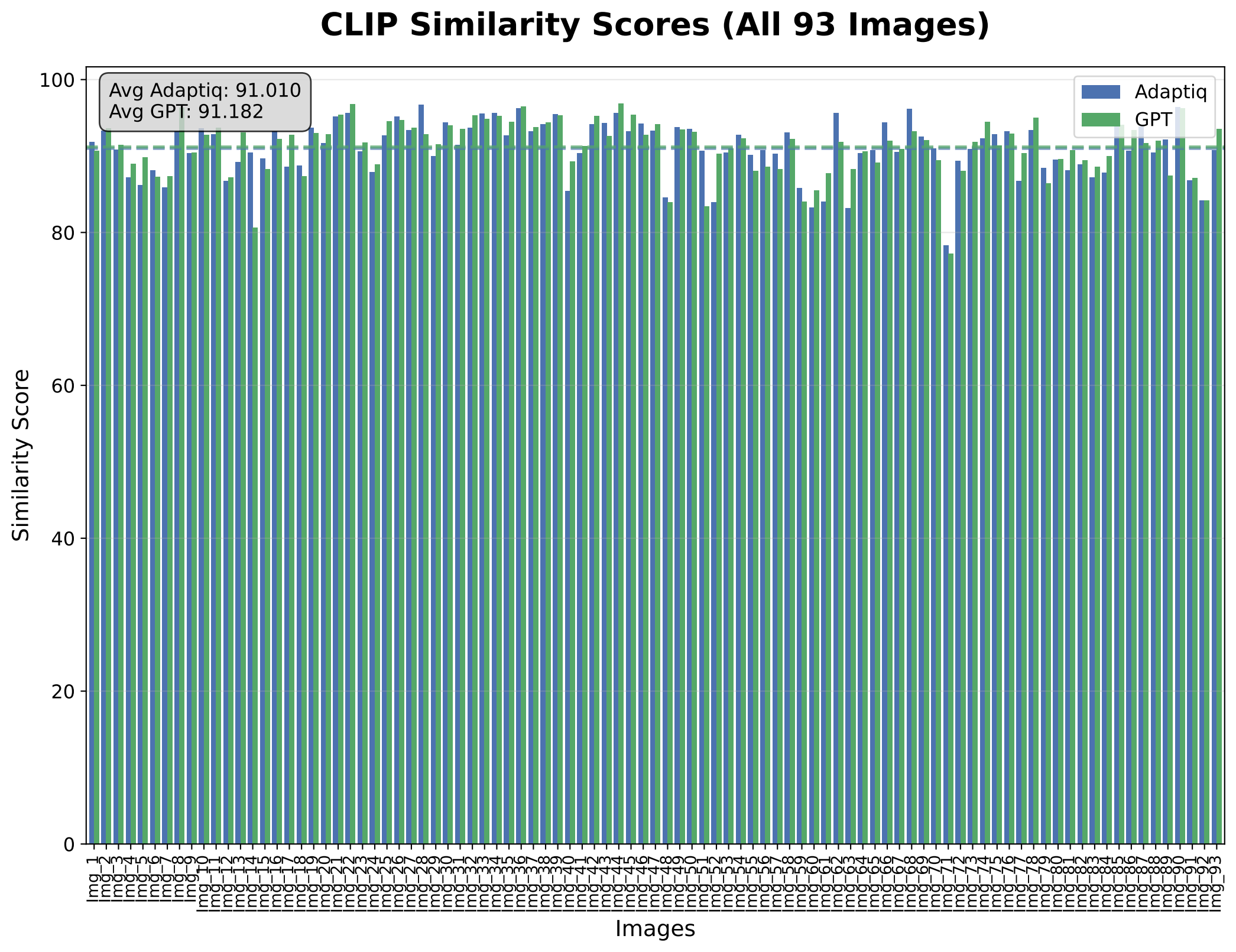
```
Mann-Whitney U Test Results:

latency: p=0.0000 ***
tokens: p=0.3656 ns
total_cost: p=0.0000 ***

* p<0.05, ** p<0.01, *** p<0.001
ns = not significant
```

Performance Trends Over Sequential Runs




CLIP Similarity Scores (All 93 Images)
Avg Adaptiq: 91.010
Avg GPT: 91.182
Adaptiq
GPT
Similarity Score
0
20
40
60
80
100
Images

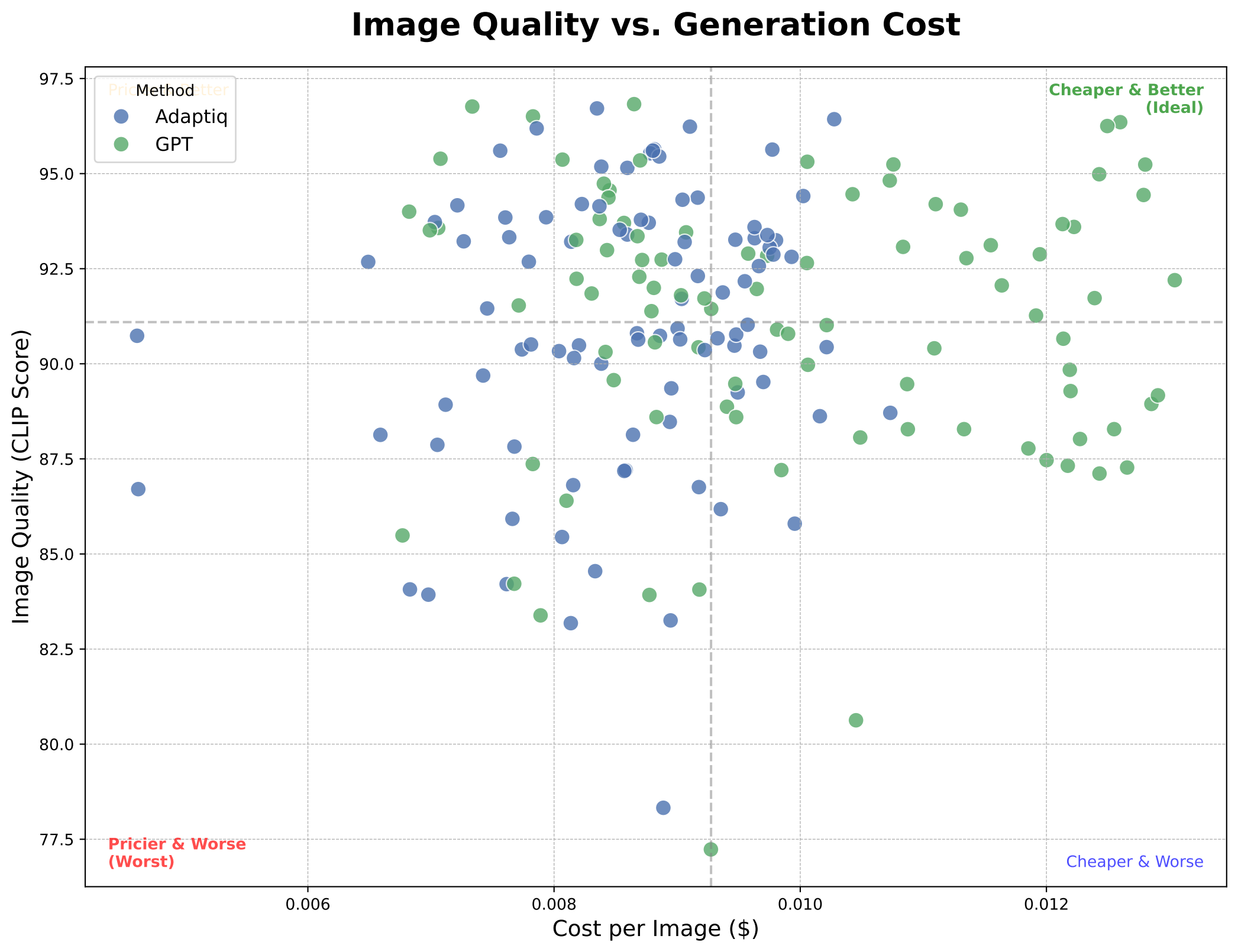
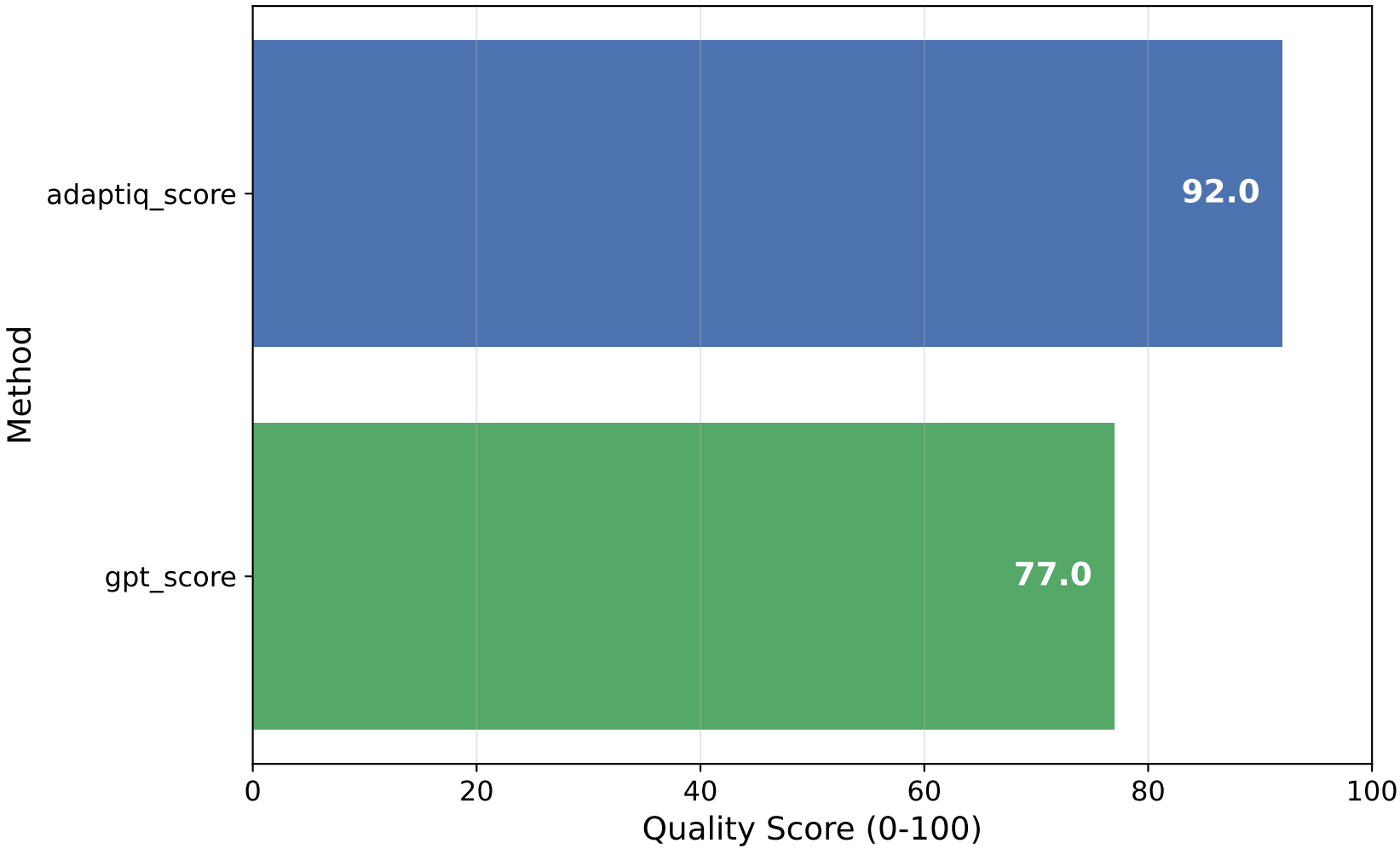


Image Quality vs. Generation Cost
Method
Adaptiq
GPT
Cheaper & Better
(Ideal)
Pricier & Worse
(Worst)
Cheaper & Worse
Image Quality (CLIP Score)
Cost per Image ($)
97.5
95.0
92.5
90.0
87.5
85.0
82.5
80.0
77.5
0.006
0.008
0.010
0.012

Prompt Engineering Quality Scores



Summary:
Adaptiq outperforms GPT method across all criteria due to its granular, modular workflow, explicit error handling, and rigid structural design. It is more precise, reliable, easier to debug, and more effective at consistently producing optimal, safe prompts. While GPT's approach covers the core requirements, it trades strictness for flexibility, which can reduce effectiveness and traceability in demanding or high-stakes workflows.