RAG System Performance Report

October 9, 2025

# Executive Summary

This report presents the performance analysis of our Retrieval-Augmented Generation (RAG) system over the past quarter. Key findings include improved response accuracy, reduced latency, and enhanced scalability.

# Performance Metrics

## Latency Analysis

Average response time: 150ms (target: 200ms)  
P95 latency: 280ms  
P99 latency: 450ms

## Accuracy Metrics

Document retrieval accuracy: 92%  
Response relevance score: 4.3/5.0  
User satisfaction: 88%

# Library Comparison

|  |  |  |  |
| --- | --- | --- | --- |
| Library | Speed | Accuracy | Cost |
| Unstructured | Fast | Good | Free |
| Azure DI | Moderate | Excellent | Paid |
| PyPDF2 | Very Fast | Fair | Free |

# Recommendations

1. Based on our analysis, we recommend:
2. Implementing hybrid extraction strategy
3. Scaling worker pods during peak hours
4. Adding semantic caching layer