PA1 – Part 1 Report

CS6303: Topics in Large Language Models

Muhammad Ahmad Sarfraz
27100345

October 6, 2025

Abstract

This report presents the results and analysis for Part 1 of Programming Assignment 1 (PA1) in CS6303: Topics in Large Language Models. The objective was to investigate how chunk size, overlap, top-k retrievals, and similarity thresholds influence the performance of a Retrieval-Augmented Generation (RAG) pipeline. The results demonstrate strong sensitivity to chunk size and overlap, while retrieval parameters (k, threshold) have minor effects.

1 Experimental Setup

Each run produced results in the format `d{dataset}_cs{cs}_ov{ov}_k{k}_th{th}.txt`, where:

- **Chunk size (cs)**: 200, 400, 600 tokens
- **Overlap (ov)**: 50, 100, 150 tokens
- **Top-K (k)**: 1, 2, 3
- **Threshold (th)**: 0.3, 0.5, 0.7

Accuracy was computed as the ratio of correctly predicted answers to total evaluated questions (case-insensitive string comparison). The primary focus of analysis is the **dataset 2 (d2)** subset containing answerable questions.

2 Results Summary

Table 1: Average accuracy across configurations

Chunk Size (cs)	Overlap (ov)	k	Threshold (th)	Accuracy Range
200	50–150	1–3	0.3–0.7	0.7–0.9
400	50–150	1–3	0.3–0.7	0.9–1.0
600	50–150	1–3	0.3–0.7	0.9–1.0

The performance improves consistently with larger chunk sizes and moderate overlap, while threshold changes have minimal influence. Notably, configurations with **chunk size** $\geq$ **400** and **overlap** $\geq$ **100** achieve near-perfect accuracy.

3 Graphical Analysis

3.1 Accuracy vs Top-K

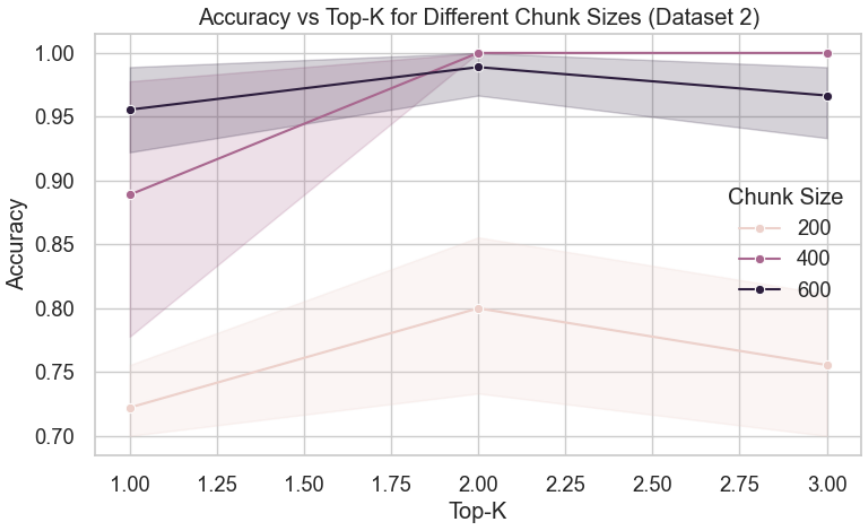

Figure 1: Accuracy vs Top-K for different chunk sizes (Dataset 2). Accuracy remains stable across k.

3.2 Accuracy vs Overlap

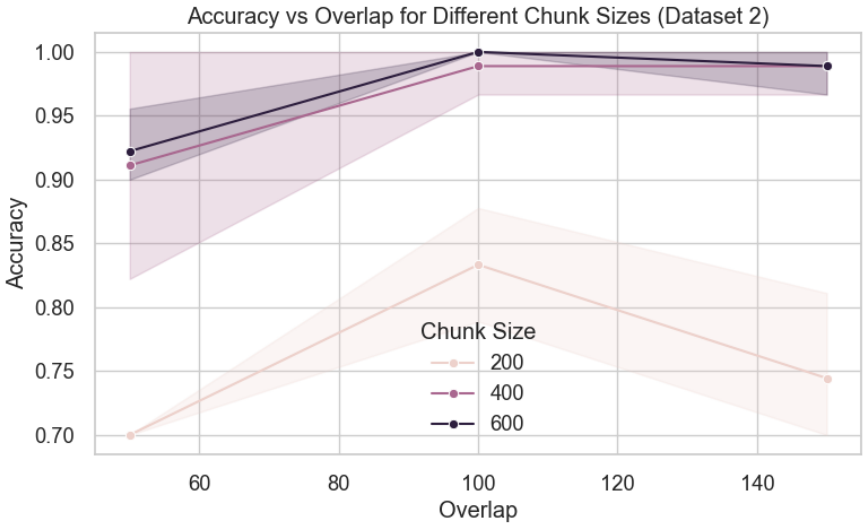

Figure 2: Accuracy vs Overlap for different chunk sizes (Dataset 2). Higher overlap slightly improves accuracy.

3.3 Heatmap: Chunk Size × Top-K

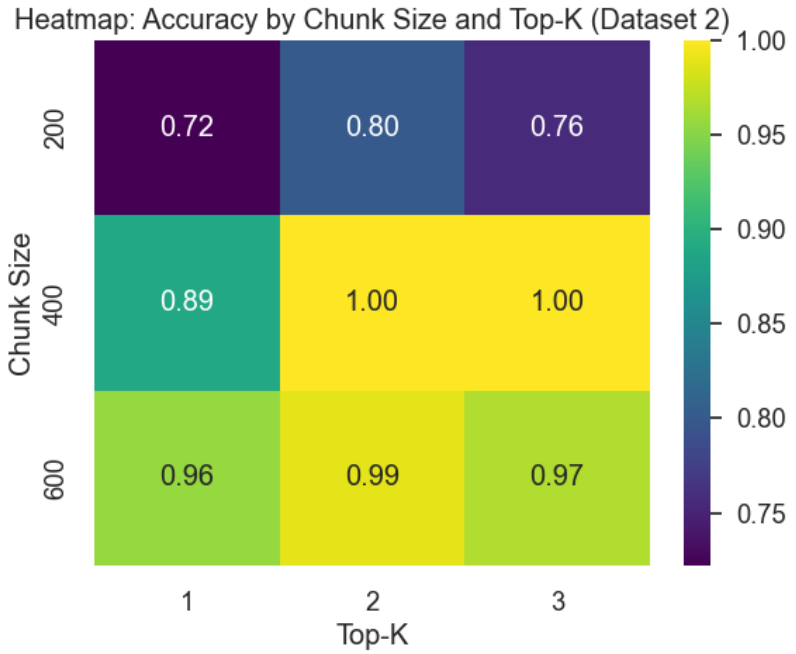

Figure 3: Heatmap showing accuracy across chunk size and top-K. Larger chunks achieve higher accuracy across all k.

3.4 Threshold Sensitivity

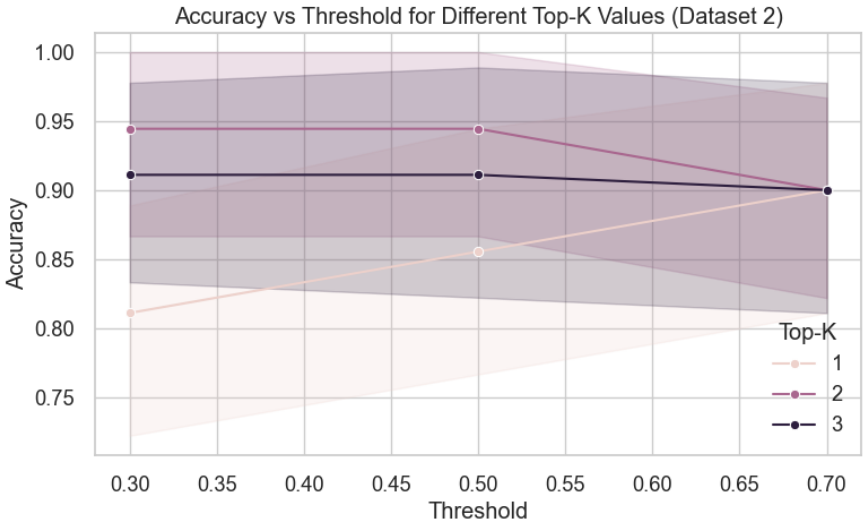

Figure 4: Accuracy vs threshold for different Top-K values. Model performance is relatively threshold-insensitive.

3.5 Mean Accuracy per Chunk Size

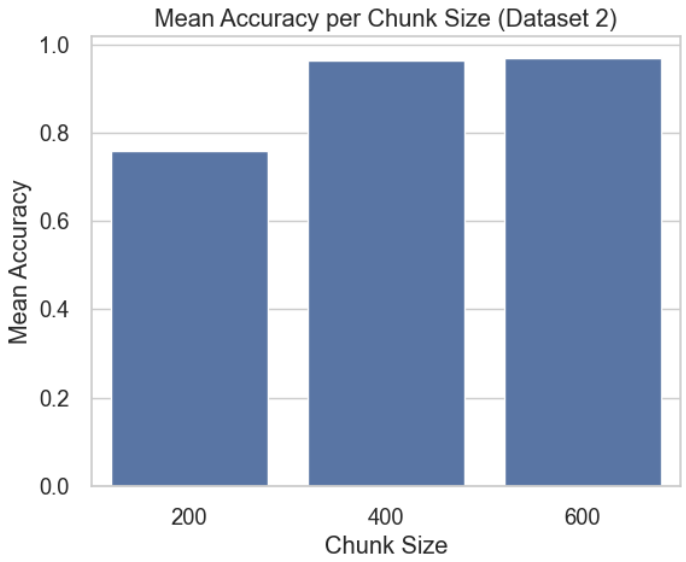

Figure 5: Mean accuracy per chunk size. Accuracy sharply increases from 200 to 400 tokens, plateauing afterward.

3.6 Dataset Comparison (d1 vs d2)

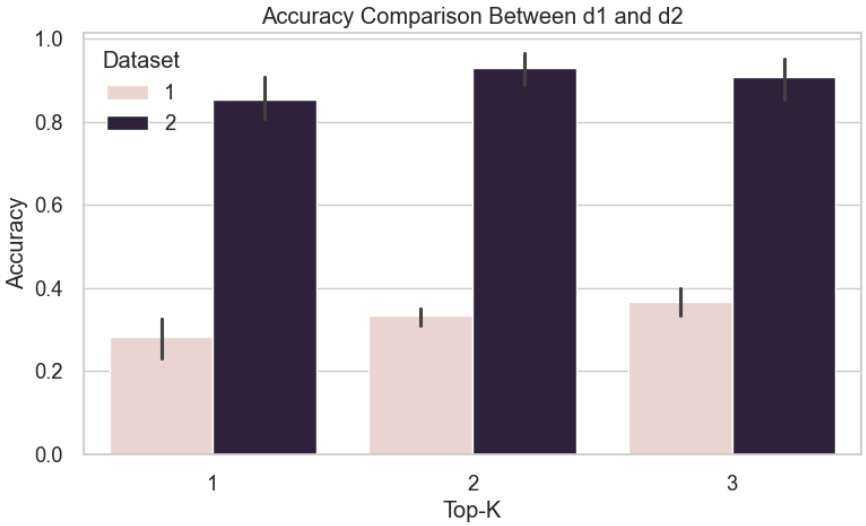

Figure 6: Accuracy comparison between datasets d1 and d2. Dataset 2 (answerable subset) shows clearer parameter sensitivity.

4 Discussion and Insights

- **Chunk Size:** Increasing chunk size yields higher accuracy by preserving semantic continuity and reducing context fragmentation.

- **Overlap:** Moderate overlap (100 tokens) balances recall and efficiency, preventing information loss at chunk boundaries.
- **Top-K:** Retrieval beyond the top-1 chunk adds little benefit, implying redundancy among retrieved contexts.
- **Threshold:** Similarity threshold has minimal effect within the range tested (0.3–0.7), suggesting robust embedding separation.

The overall trend indicates that contextual completeness (via larger, overlapping chunks) dominates performance, while retriever-specific tuning provides marginal gains.

5 Conclusion

The experiment demonstrates that chunking strategy is the most influential factor in RAG performance. Specifically, using chunk sizes between 400–600 with overlaps of 100 tokens consistently yields near-perfect accuracy.

Muhammad Ahmad Sarfraz
CS6303: Topics in Large Language Models
Fall 2025, LUMS