Hierarchical Attention Tree (HAT) Architecture

O(log n) search

Query

Session Centroid

Session Level

embed(q)

Query Vector

Doc 1

Doc 2

Doc 3

Document Level

Top-k Results

C1

C2

C3

C4

C5

Retrieved Chunks

Chunk Level