What we have:

Ada-embedding

768 tolkens

Want to optimize quality of output from RAG

Define quality?

Training Data – what is the golden dataset?

Levers to explore:

* Training data
  + SQuAD – preexisting training data
* North Star Metric (how do we measure how good the output is?)
  + BERT Score
  + BERT vectors - measure distance between answers
  + Cosine Similarity
  + Flesch-Kincaid reading level aka readability (lib: textstat)
  + Check for grammatical errors ~ probably not
* Top K (aka what top results are being used from the vector database?)
  + Currently its set to 3
* Change embedding?
  + Currently its Ada
* Chunking length of input text? (Does it depend on the original document length?)
  + Type:
    - Technical documents should be shorter
    - Story / fictional: longer is better
    - Code: chunk by function
    - Legal: By sections
    - News: By paragraph or article if short enough)
  + Length:
    - Short documents: short chunks
    - Long documents: longer might be better
  + Retrieval Method:
    - Embedding: shorter chunks \*\* we are here
    - Keyword: larger chunks
  + General Warning:
    - HAVE TO BE CAREFUL LONGER CHUNKS cause INFORMATIONAL OVERLOAD
    - If a chunk is truncated it will cause significant information loss
  + Guidelines:
    - Between 500-2000 tokens
    - Try to use natural breaks (paragraph / sentence / page?)
* Length of overlap between text is ideal?
  + Guidelines:
    - 20-50 tokens