Aman Singhal

New York University

Introduction

In this project, we aim to extract context for RAG from pdf documents. Directions include exploring:

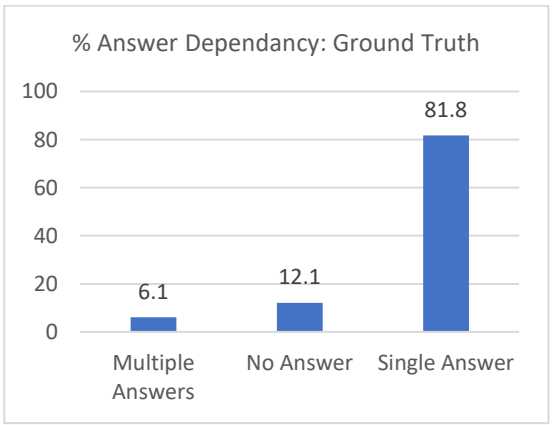
- Parsing/chunking: e.g. correcting over-chunking
- 1-level & 2-level search: retrieve/retrieve and re-rank
- Evaluation metrics e.g. Recall@k, MRR
- Zero/Single/Multiple retrieved contexts

Models

We consider FAISS FlatIndex due to our small corpus size and experiment with embedders:

1. TF-IDF/BM25 (level-1) and DRAGON+ (re-ranker)
2. DRAGON+
3. TF-IDF/BM25
4. Sentence Transformer

Labelling



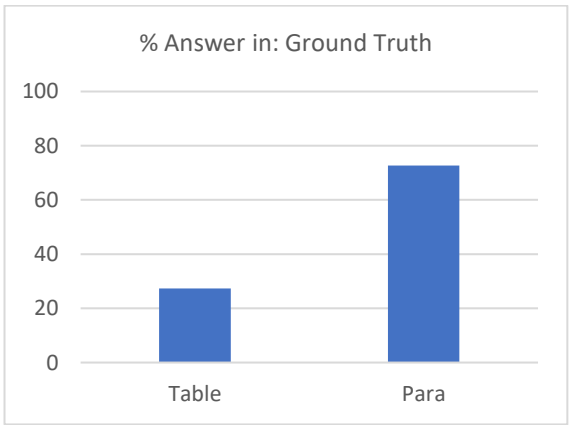


Fig. Questions are manually labelled with corresponding answer contexts and tagged as table or paragraph-based questions.

Mob: +15513446364

Email: as14661@nyu.edu

Metrics

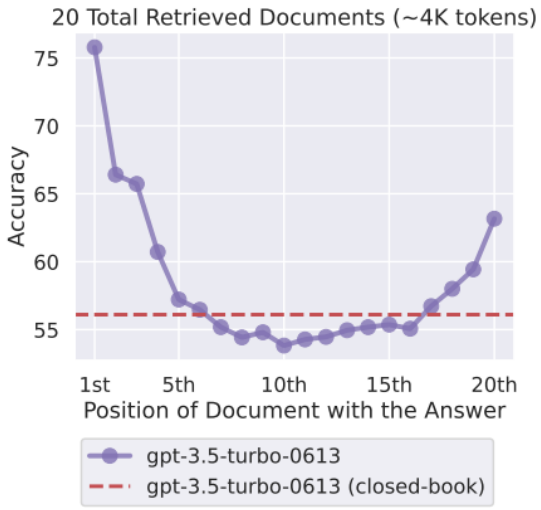


Fig. "RAG: Lost in the middle": Most likely chunks should be at the start of a context. With Recall@10: we capture all positives, providing relevant context to an LLM; and with MRR: position more relevant contexts earlier

Results & Analysis

1. Correcting over-chunking

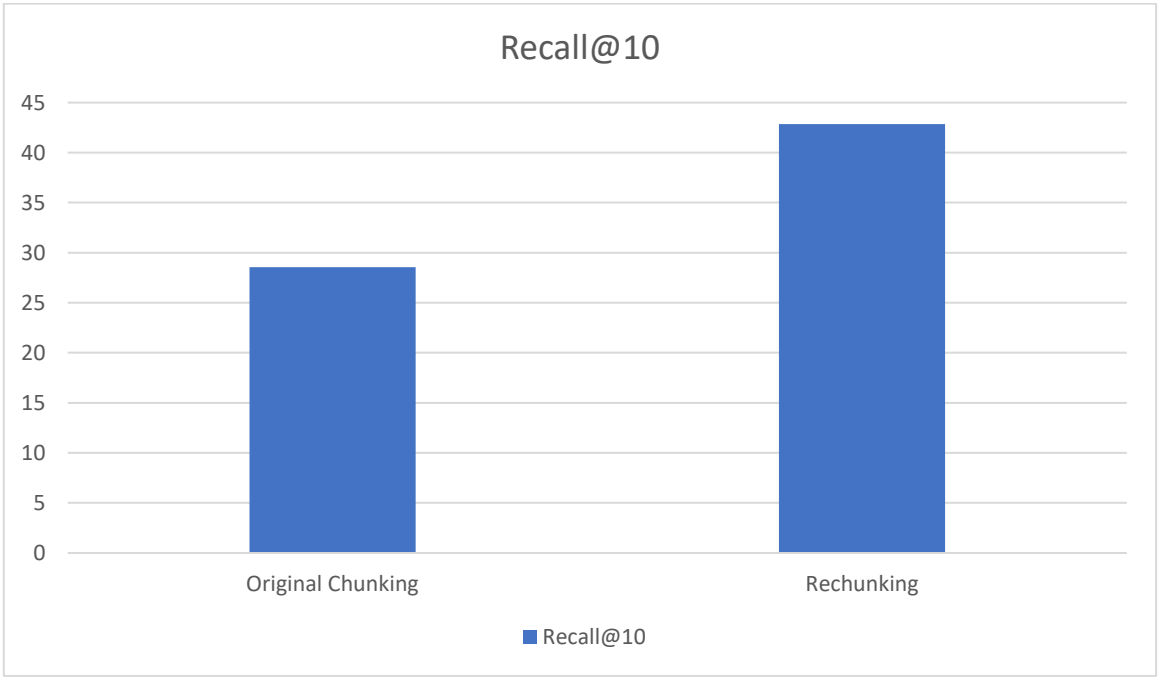


Fig. Rechunking reduces drop in recall due to incomplete chunks. Revised recall matching criteria is fuzzy (ground truth, retrieved context) > 0.9 and len(retrieved context) > 0.95*len(ground truth) - account for cases when ground truth ⊆ retrieved context

Mob: +15513446364 Email: as14661@nyu.edu

2. Level-1 retrieval: Comparing dense and sparse embedders

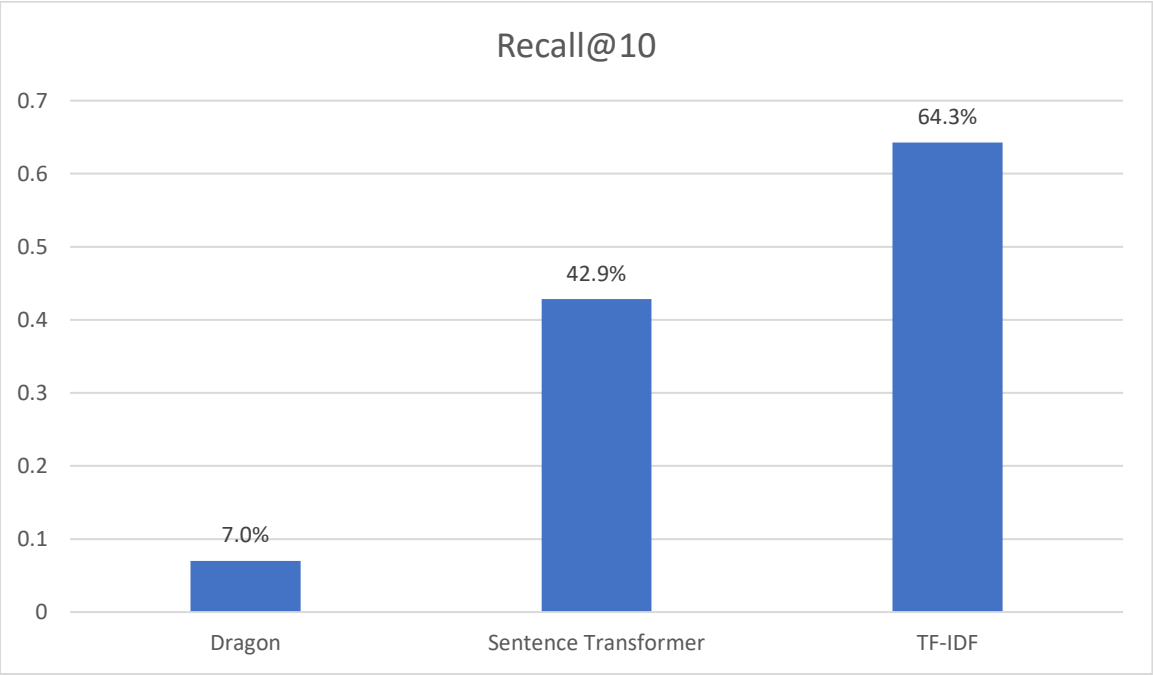
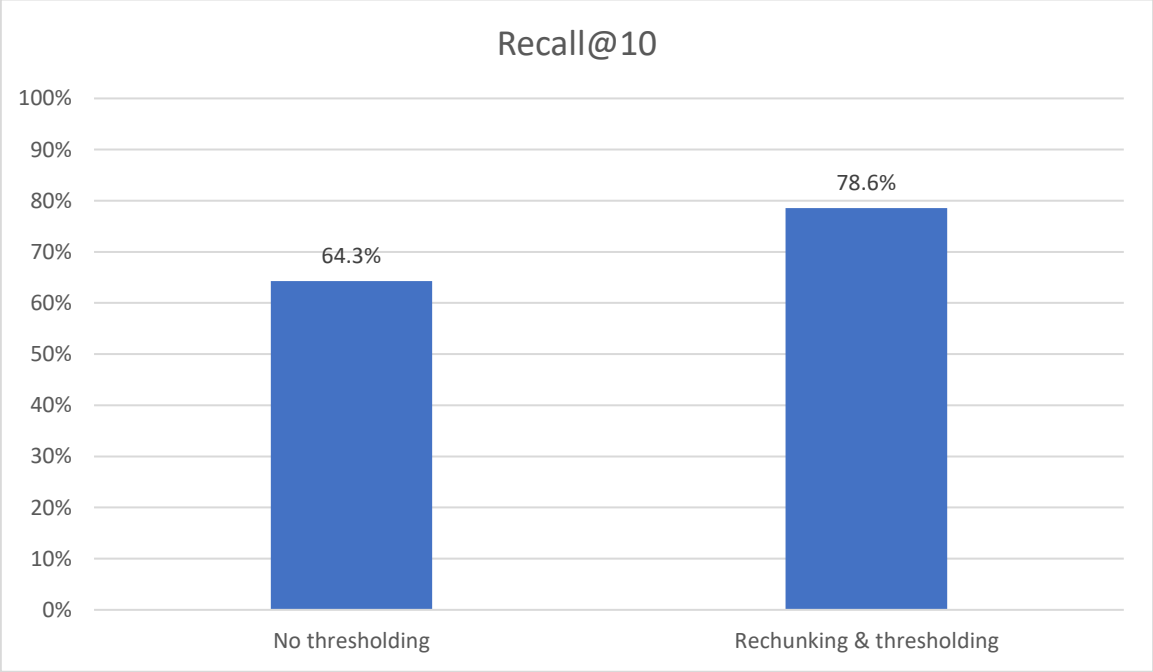


Fig. Dense vs sparse indexing. Dense retrieval discourages n-gram matches; all given questions require n-gram match

3. No answer threshold: Reduce downstream LLM hallucination as sometimes answers are not present



Mob: +15513446364 Email: as14661@nyu.edu

Fig. Thresholding sets retrieved contexts to None, as answer context may not present. Threshold is median level-1 similarity between retrieved chunk and query, when ground truth is NULL set

4. TF-IDF/BM25 vs 2-level TF-IDF/BM25 & DRAGON+

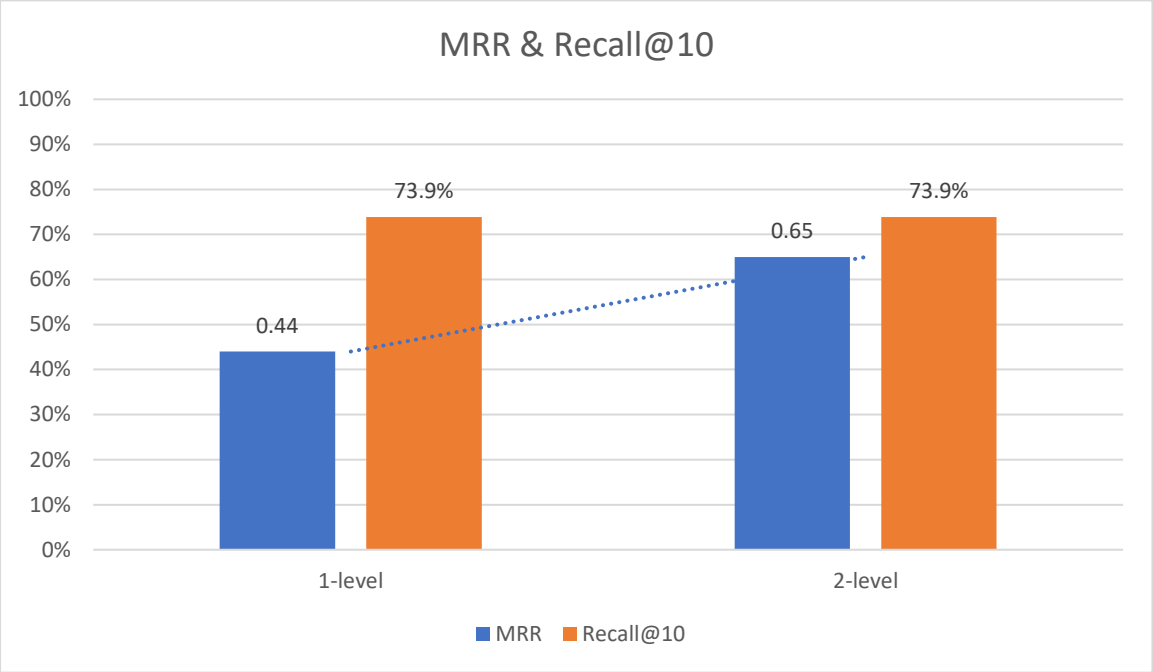
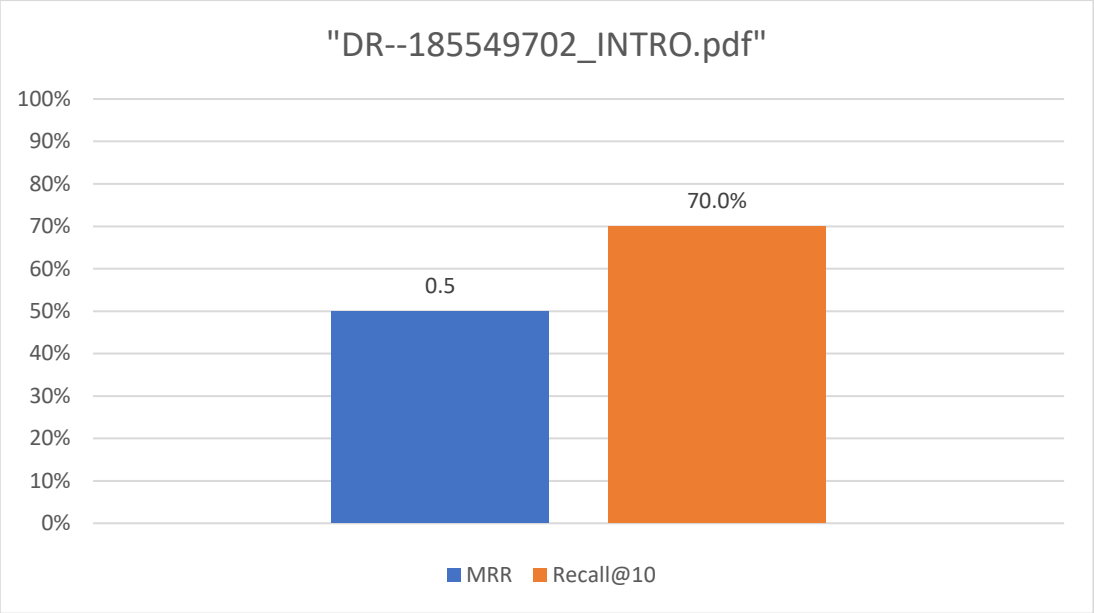


Fig. 2-level retrieval improves MRR while Recall@10 remains same. 2-level comprises TF-IDF sparse retrieval & Dragon dense re-ranking

Mob: +15513446364 Email: as14661@nyu.edu

5. Result on DR--185549702_INTRO



Future Work & Error Analysis

1. Chunking: In-house document segmentation/Parent Retriever (https://python.langchain.com/docs/modules/data_connection/retrievers/parent_document_retriever/)

2. Retrieval: SPLADE and (https://huggingface.co/spaces/mteb/leaderboard)

References

[1] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, Percy Liang; Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 2024; 12 157–173. doi: https://doi.org/10.1162/tacl_a_00638

[2] How to Train Your DRAGON: Diverse Augmentation Towards Generalizable Dense Retrieval SC Lin*, A Asai, M Li, B Oguz, J Lin, Y Mehdad, W Yih, X Chen* arXiv preprint arXiv:2302.07452

Mob: +15513446364 Email: as14661@nyu.edu

Mob: +15513446364 Email: as14661@nyu.edu