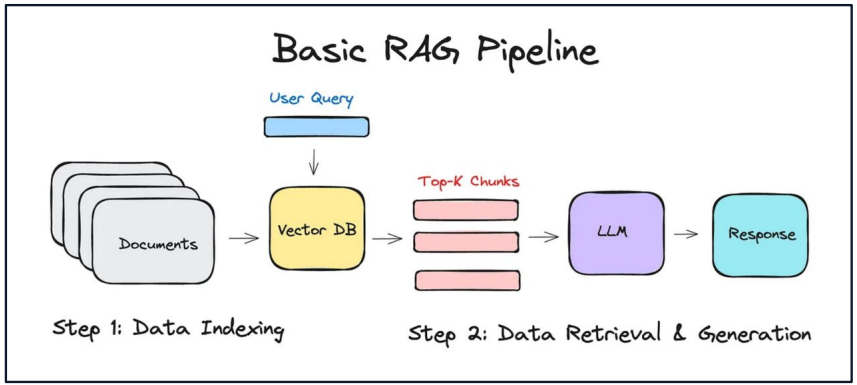


RAG System that expands Eligibility Criteria for Clinical Trial

Table of contents

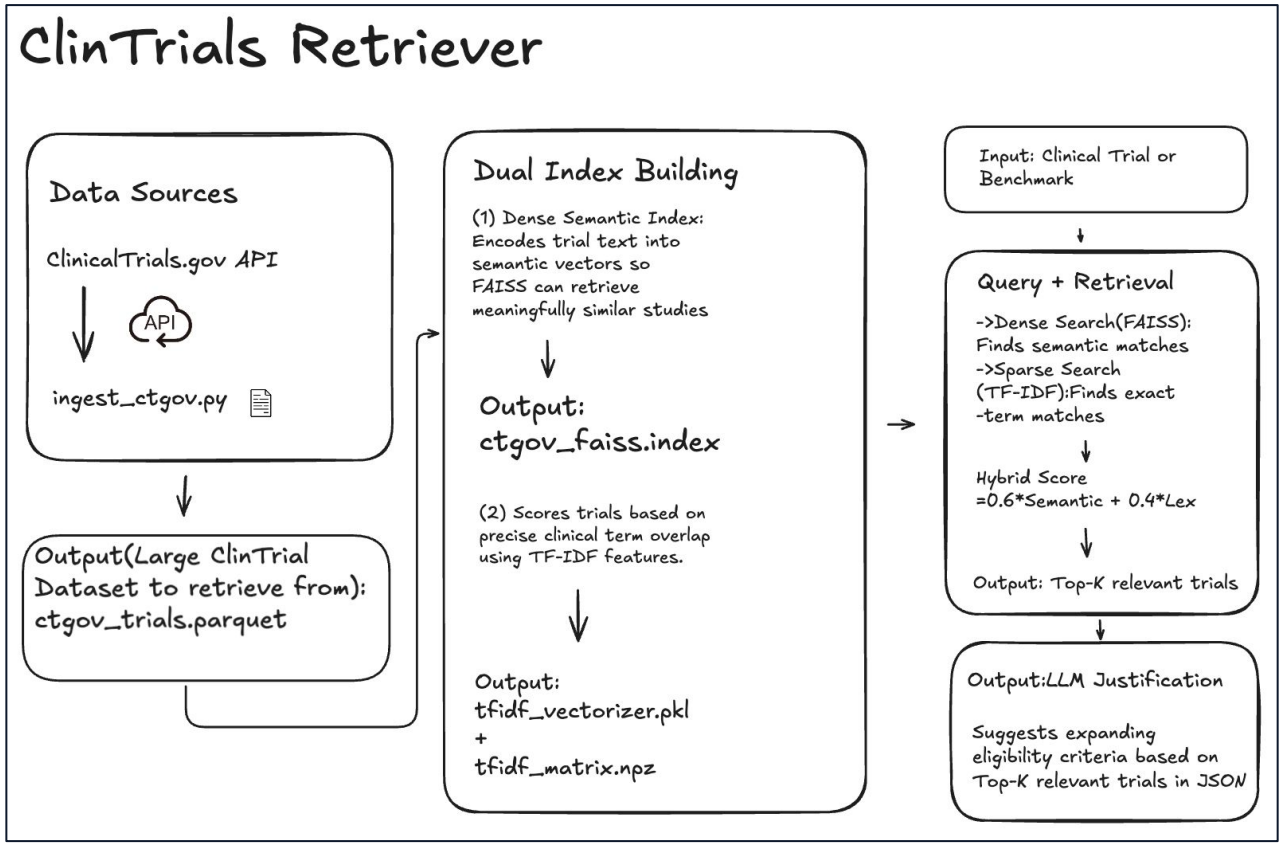
01 ClinTrial Retriever

Retrieving Relevant Clinical Trials from *ClinicalTrials.gov* that can be used as evidence for increasing eligibility criteria

02 BiomedLit Retriever

Retrieving Relevant Biomedical Literature from *PubMed* that can be used as evidence for increasing eligibility criteria

01 ClinTrial Retriever Design

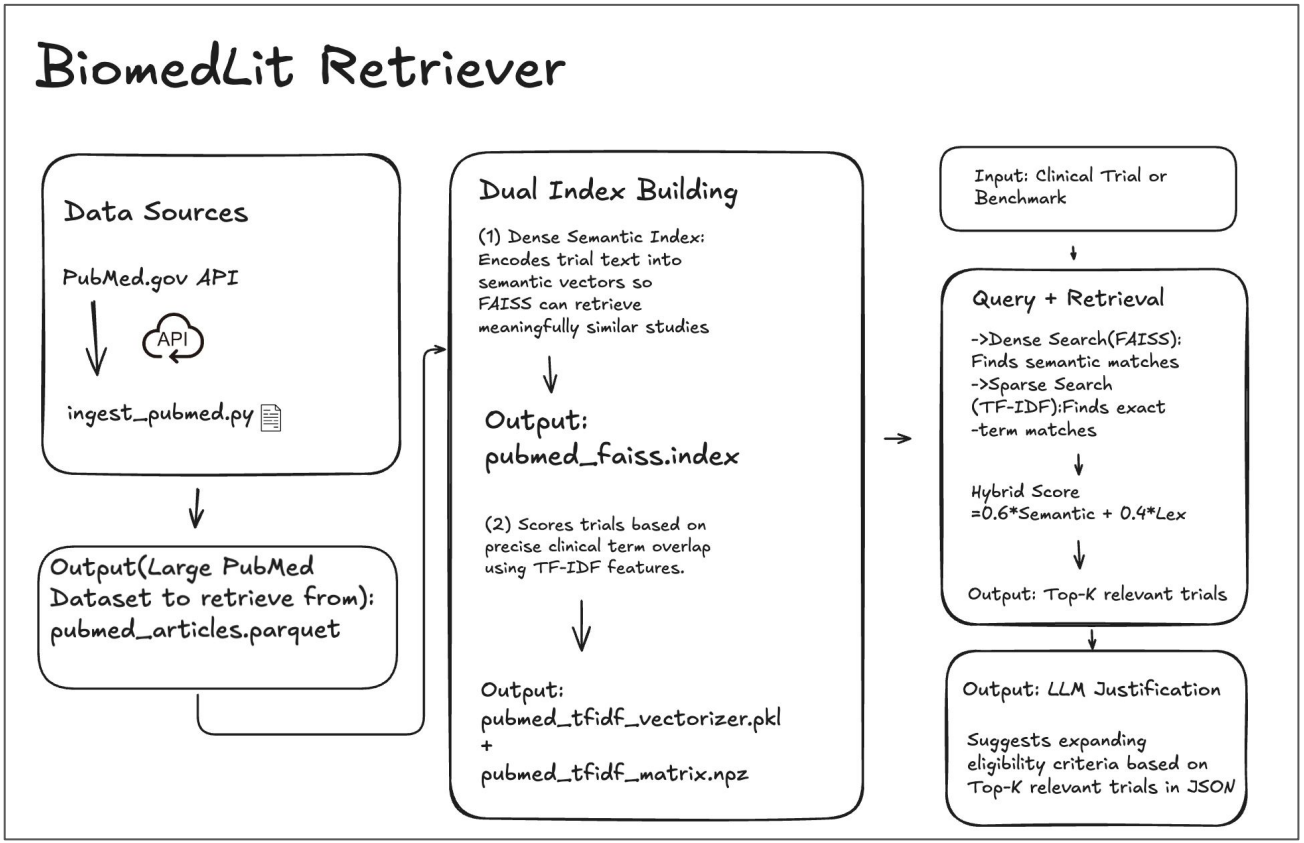


Dual Indexing Low Level Details:

FAISS (Facebook AI Similarity Search) *works by converting each text into a dense numerical embedding and storing these in a vector index, allowing it to quickly retrieve semantically similar items using efficient nearest-neighbor search in high-dimensional space.*

TF-IDF (Term Frequency–Inverse Document Frequency) *works by representing each document as a weighted bag of words, scoring terms higher if they appear frequently in a document but rarely across others, thus identifying precise keyword matches that capture lexical relevance.*

02 BiomedLit Retriever Design



Dual Indexing Low Level Details:

FAISS (Facebook AI Similarity Search) *works by converting each text into a dense numerical embedding and storing these in a vector index, allowing it to quickly retrieve semantically similar items using efficient nearest-neighbor search in high-dimensional space.*

TF-IDF (Term Frequency–Inverse Document Frequency) *works by representing each document as a weighted bag of words, scoring terms higher if they appear frequently in a document but rarely across others, thus identifying precise keyword matches that capture lexical relevance.*

01 ClinTrial Retriever Example

Input: Phase 3 clinical trial evaluating metformin in adults with type 2 diabetes. Excludes patients with mild renal impairment (eGFR < 60) and those using insulin.

↓

Relevant Clinical Trials

nct_id	official_title	conditions	eligibility_criteria
NCT01020123	A 4-month, Randomized, Double-blind, Placebo- ...	Type II Diabetes Mellitus	Inclusion Criteria:\n\n* female of non-childbe...
NCT00111631	A Randomized, Double-blind Study of the Effect...	Diabetes Mellitus Type 2	Inclusion Criteria:\n\n* adult patients 18-75 ...
NCT04295031	AMEMET Study - Observational Multicentric Clin...	Diabetes Mellitus, Type 2	Inclusion Criteria:\n\n* adult outpatients (≥6...
NCT00660907	A 52-Week International, Multi-centre, Randomi...	Type 2 Diabetes	Inclusion Criteria:\n\n* Type 2 Diabetes\n* Tr...
NCT04854512	Randomized, Multicenter, Double-blind, Paralle...	Diabetes Mellitus, Type 2	Inclusion Criteria:\n\n1. Is male or female an...
NCT01951183	A MULTI-CENTER, DOUBLE-BLIND, RANDOMIZED, PLAC...	Diabetes Mellitus Type 2	Inclusion Criteria:\n\n* Adults between the ag...
NCT01353391	Metformin in Women With Type 2 Diabetes in Pre...	Type 2 Diabetes	Inclusion Criteria:\n\n1. Women who are betwee...
NCT00767351	Intraindividual Variation of Serum Metformin i...	Diabetes Mellitus Type 2, Metformin, Pharmacok...	Inclusion Criteria:\n\n* Patients above the ag...

↓

LLM Output

Based on the provided information, here is the analysis of the eligibility criteria for the target trial evaluating metformin in adults with type 2 diabetes, focusing on the exclusion of patients with mild renal impairment (eGFR < 60) and those using insulin.

1. Identify Overly Strict Criteria
The exclusion of patients with mild renal impairment (eGFR < 60) and those using insulin may be considered overly strict. Many patients with type 2 diabetes have varying degrees of renal function and may require insulin therapy as part of their management.

2. Suggest How Much to Widen
- **Renal Impairment**: Consider including patients with eGFR ≥ 45 mL/min. This would allow for a broader range of patients with mild to moderate renal impairment to participate.
- **Insulin Use**: Allow patients who are on stable doses of insulin (e.g., for at least 3 months prior to enrollment) to be included in the trial.

3. Justify with NCT Citations and Reasoning
- **Renal Impairment**:
  - In NCT01020123, patients with an eGFR < 60 mL/min were excluded, which aligns with the target trial's criteria. However, NCT04854512 allows inclusion of patients with eGFR ≥ 30 mL/min, indicating that patients with mild renal impairment can still be safely included in trials evaluating metformin (NCT04854512). This suggests that widening the criteria to eGFR ≥ 45 mL/min could be safe and beneficial.

- **Insulin Use**:
  - NCT00111631 and NCT00660907 both exclude patients using insulin, which may limit the applicability of findings to real-world scenarios where insulin is often used in conjunction with metformin. Allowing stable insulin users could enhance the trial's relevance and provide insights into the efficacy of metformin in a broader patient population.

4. Output JSON
```json
{
  "criteria": [
    {
      "criterion": "eGFR < 60 mL/min",
      "recommendation": "Include patients with eGFR ≥ 45 mL/min",
      "magnitude": "Widen by 15 mL/min",
      "justification": "Patients with mild renal impairment can be safely included based on evidence from NCT04854512.",
      "citations": ["NCT04854512"]
    },
    {
      "criterion": "Exclusion of insulin users",
      "recommendation": "Allow stable insulin users (on stable doses for at least 3 months)",
      "magnitude": "Remove exclusion for insulin use",
      "justification": "Including stable insulin users enhances trial relevance and reflects real-world treatment scenarios.",
      "citations": ["NCT00111631", "NCT00660907"]
    }
  ]
}
```

This JSON output provides a structured approach to widening the eligibility criteria based on evidence from existing trials, ensuring that the trial can capture a more diverse patient population while maintaining safety and efficacy standards.

The ClinTrial Retriever *takes a trial description as input, searches ClinicalTrials.gov for the most similar studies, and displays their key eligibility criteria for comparison. The LLM then analyzes these retrieved trials to justify how the input trial's criteria could be safely broadened, generating a structured JSON with evidence-based recommendations.*

02 BiomedLit Retriever Example

Input: Phase 3 breast cancer trial excluding mild diabetics; age 18–65; ECOG 0–1.

↓

Relevant Literature

```
Retrieved 8 articles:

1. PMID 41008932
   Title: Breast Immunology Network: Toward a Multidisciplinary and Integrated Model for Breast Cancer Care in...
   Journal: Cancers
   Authors: Botticelli, Andrea; Brignoli, Ovidio; Caruso, Francesco; Curigliano, Giuseppe; D...

2. PMID 41088941
   Title: Breaking Barriers in Breast Cancer: Multi-Targeted Therapeutic Insights.
   Journal: Mini reviews in medicinal chemistry
   Authors: Unni, Apsara; Rajagopal, Kalirajan; Shevate, Krishna; Byran, Gowramma

3. PMID 41081997
   Title: The Management of ER-Low, HER2-Negative Breast Cancer in the United Kingdom: A Survey of Current Cli...
   Journal: Oncology and therapy
   Authors: Sit, Ka Chai; Berner, Alison M; McCahill, Conor; Atkinson, Sarah; Phillips, Meli...

4. PMID 41097632
   Title: Decline of PD-L1 Immunoreactivity with Storage Duration in Formalin-Fixed Paraffin-Embedded Breast C...
   Journal: Cancers
   Authors: Yanagihara, Keiko; Nagata, Koji; Yamakawa, Tamami; Kato, Sena; Tamura, Miki; Yos...

5. PMID 41072918
   Title: Tailoring Treatment to Cancer Risk and Patient Preference: The 2025 St Gallen International Breast C...
   Journal: Annals of oncology : official journal of the European Society for Medical Oncology
   Authors: Burstein, H J; Curigliano, G; Gnant, M; Loibl, S; Regan, M M; Loi, S; Denkert, C...

6. PMID 41098738
   Title: Breaking the heterogeneity barrier: a robust prognostic signature for survival stratification and im...
   Journal: Frontiers in immunology
   Authors: Shen, Haixing; Zheng, Qing; Wang, Zhenyu; Zheng, Daitian; Gong, Zhenqi; Wang, Hu...

7. PMID 41133751
   Title: [Development of a predictive model for estimating healthcare costs in women with non-metastatic brea...
   Journal: Revista espanola de salud publica
   Authors: Rodríguez-Abellón, María Jesús; Acea-Nebril, Benigno; Balboa-Barreiro, Vanesa

8. PMID 41007771
   Title: The Role of Immunotherapy and Radiation Therapy in the Treatment of Breast Cancer.
   Journal: Biomedicines
   Authors: Panse, Drishti; Hsieh, Kristin; Arons, Danielle; Runnels, Juliana; Wassel, Monic...
```

The BiomedLit Retriever *takes a trial description as input, searches PubMed.gov for the most similar literature, and displays them. The LLM then analyzes these retrieved trials to justify how the input trial's criteria could be safely broadened, generating a structured JSON with evidence-based recommendations.*

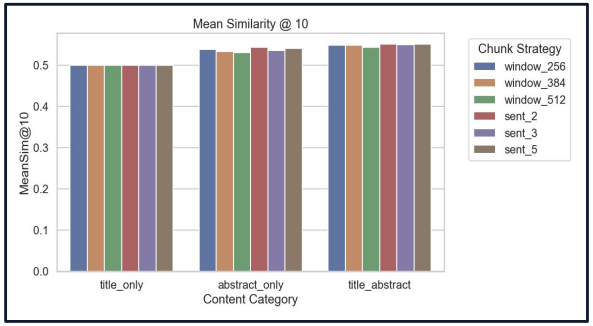
01 ClinTrial Retriever Performance on Benchmark Across Different Chunking Strategies

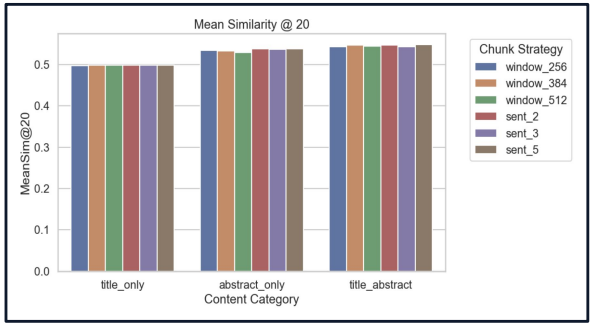
Evaluation Metrics

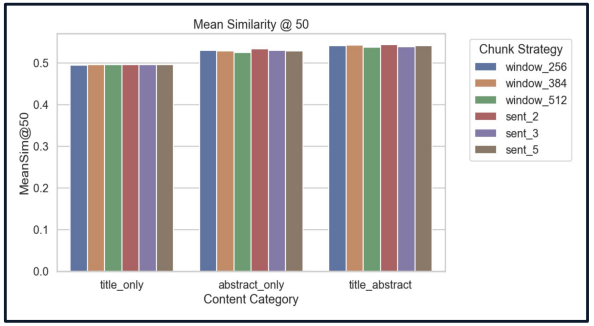
- **MeanSim@K:** measures the average cosine similarity between a query's embedding and its top-K retrieved results, showing how semantically close the retrieved trials are to the query in embedding space.
- **Recall@K:** measures the proportion of retrieved results among the top K whose similarity to the query exceeds a defined threshold, indicating how many of the most relevant or semantically matching trials the retriever successfully captured.

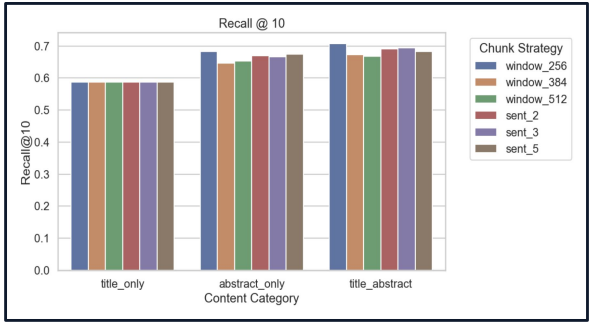
Chunking:

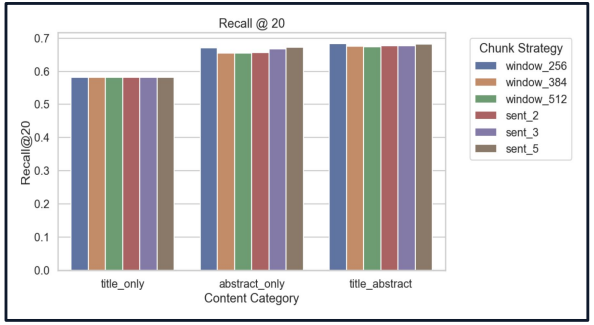
- **Window-based chunking** splits text into fixed-size overlapping character windows (e.g., 256, 384, 512 characters) to preserve continuity and context across boundaries.
- **Sentence-based chunking** divides text into groups of a set number of sentences (e.g., 2, 3, 5) to create semantically coherent chunks aligned with natural language structure.

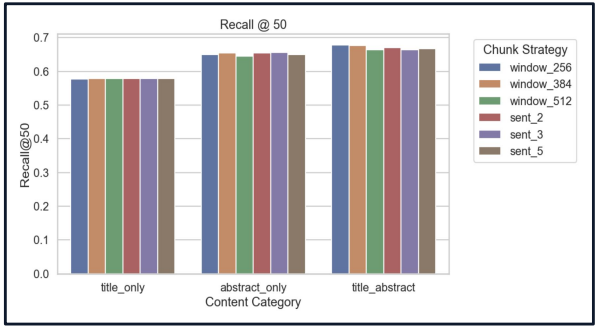




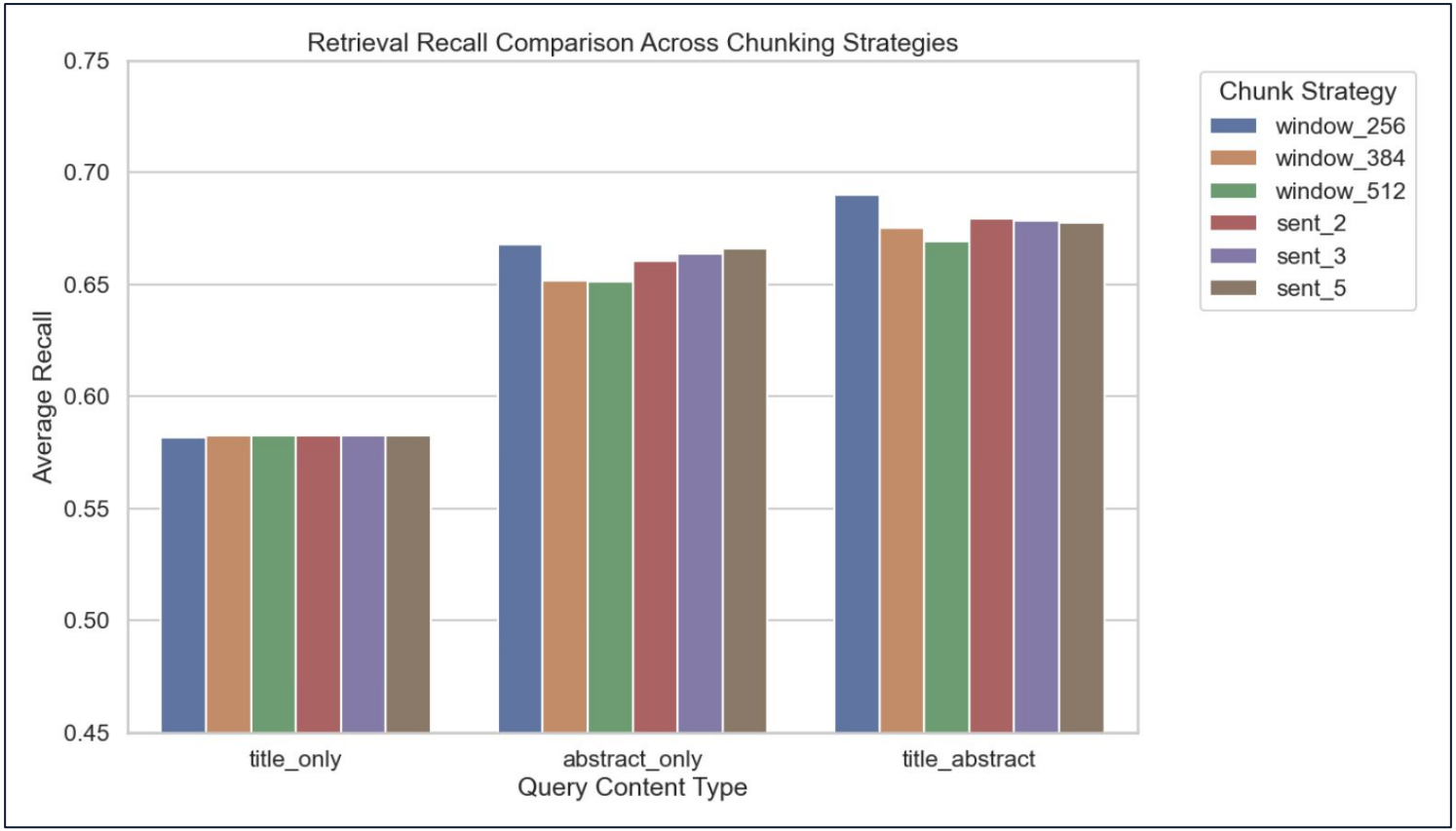








01 ClinTrial Retriever Performance on Benchmark Across Different Chunking Strategies(Average Eval)



Results represent the average performance across all evaluated Mean Similarity (MeanSim@K) and Recall@K metrics, providing an overall view of retrieval quality for each chunking strategy.

Questions?