

Advanced RAG Optimization

Build & Ship Production-ready RAG





Aravind Parameswaran

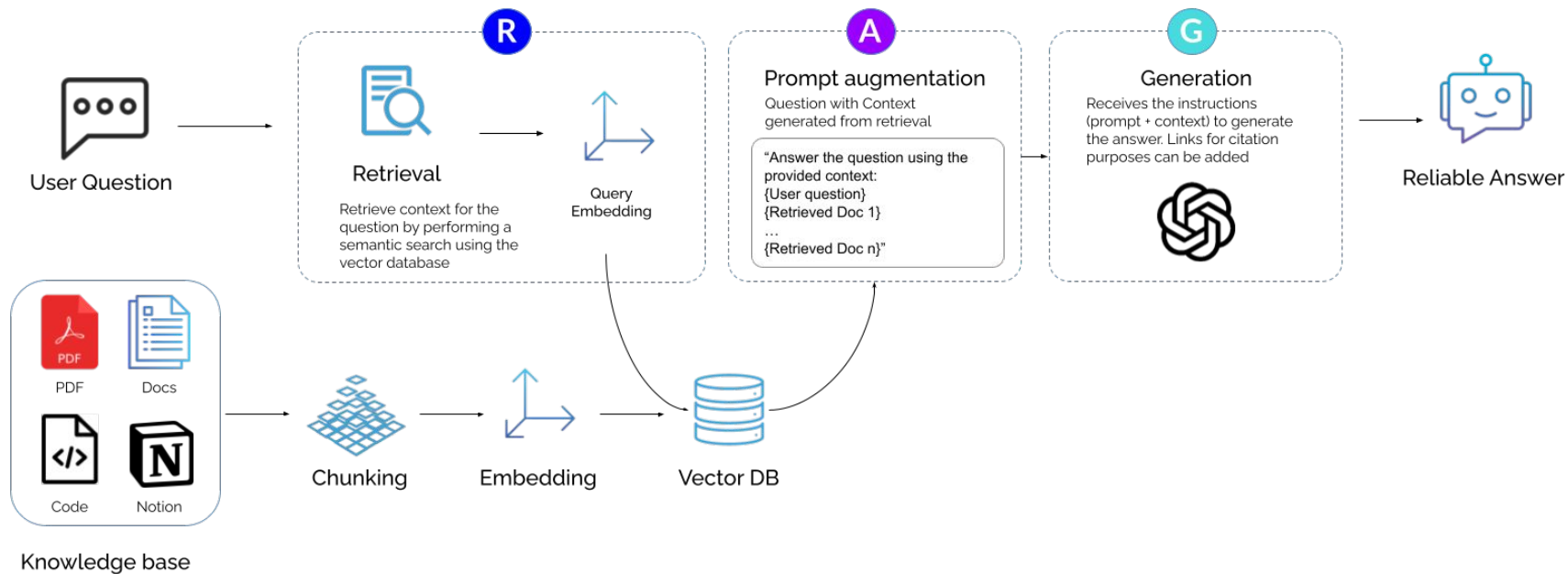
Co-Founder



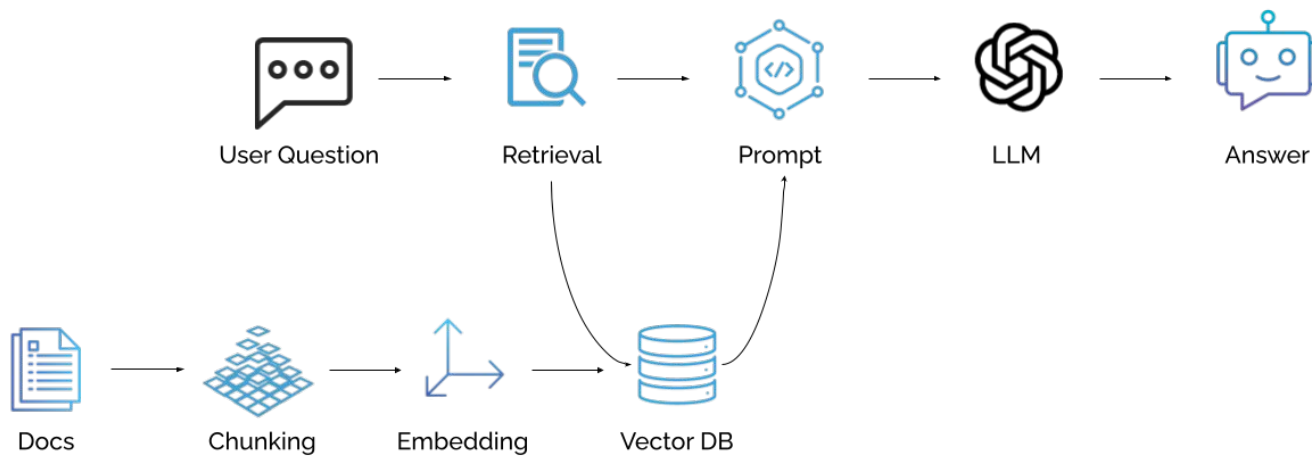
aravind@krux.ai



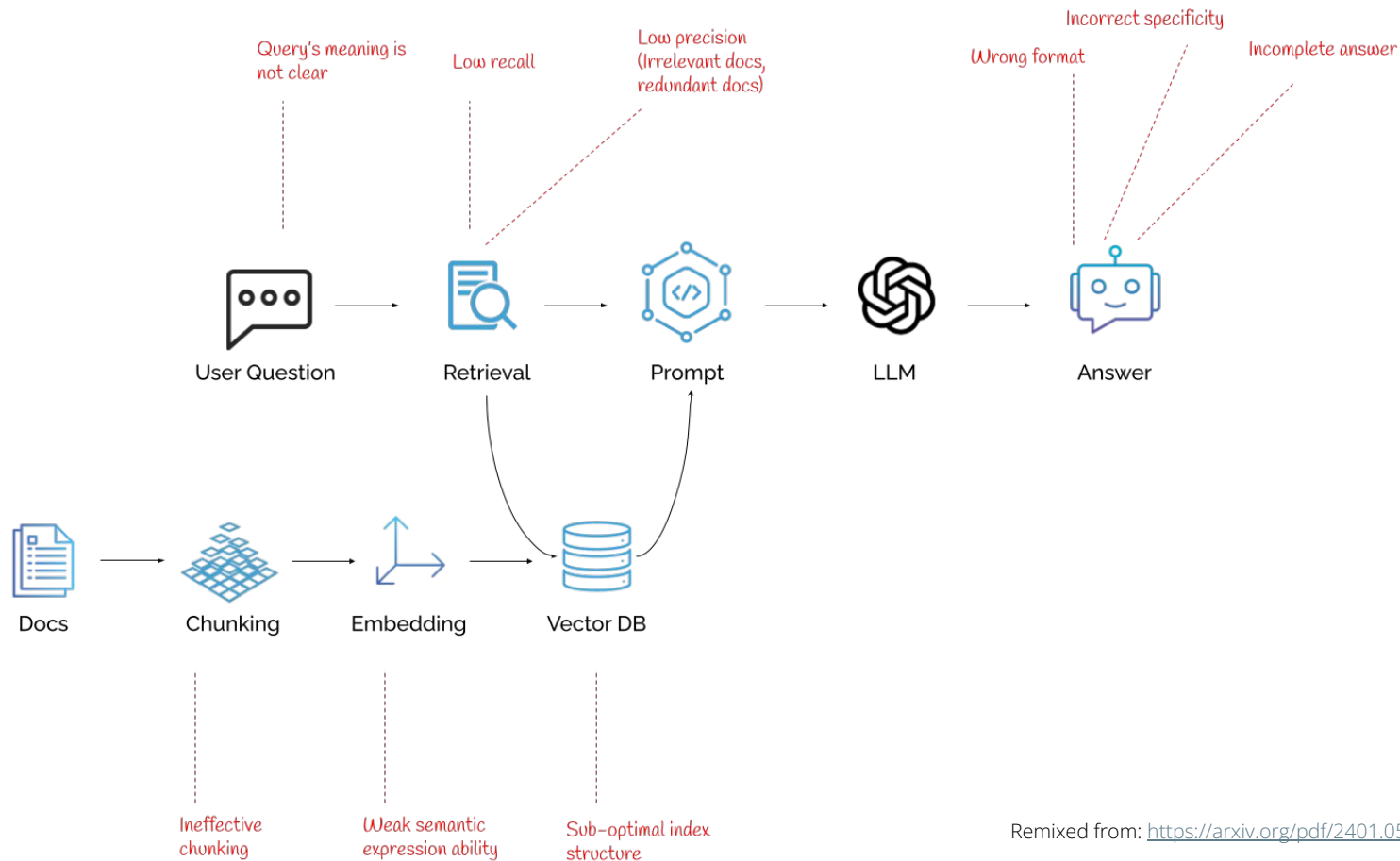
RETRIEVAL AUGMENTED GENERATION



NAIVE RAG



FAILURE POINTS IN RAG



RAG OPTIMIZATION TECHNIQUES

Pre-Retrieval

Improving the quality and retrievability of the information

TECHNIQUES

- Information density
- Optimize chunking
- Hierarchical index
- Retrieval symmetry
- Deduplication
- Context enrichment
- Query Transformation

Retrieval

Improving search performance & retrieval results

TECHNIQUES

- Query Routing
- Metadata Filtering
- Recursive Retrieval
- Ensemble / Fusion retrieval
- Late interaction models
- Combine graph + vector search

Post-Retrieval

Improving the relevance & information density of the retrieved results

TECHNIQUES

- Re-ranking
- Contextual compression
- Corrective RAG

Generation

Improving the the LLM call(s) that generate the final user response

TECHNIQUES

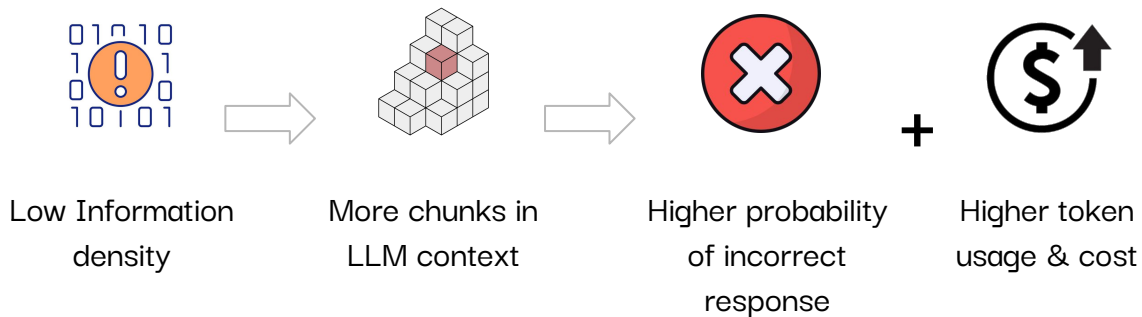
- Optimize Top "k" chunks to user in prompt
- Prompt optimization (CoT, CoN, etc.)
- Self-RAG
- Fine-tuning

PRE-RETRIEVAL OPTIMIZATION

Improving Information Density

Issues to address:

- Low information density
- Irrelevant information and/or noise
- Information duplication



PRE-RETRIEVAL OPTIMIZATION

Improving Information Density

- Raw HTML: ~55,000 tokens
- Stripped HTML: 1,500 tokens
- GPT-4 processed HTML: 330 tokens



Raw HTML

```
<html lang="en-US" xml:lang="en-US"><head class="at-element-marker">

  <script async="" src="https://cdn.branch.io/branch-latest.min.js"></script><script id="launch" data-launcher="success">
    document.addEventListener("at-library-loaded", (function(e) {
      document.getElementById('launch').setAttribute('data-launcher', 'success')
    })
  );
</script>

<meta charset="UTF-8">
<title>High Interest IRA Savings Account | Example Bank Select Savings IRA</title>

<meta name="description" content="Learn about Example Bank's Preferred Savings IRA, a high interest IRA savings account with higher rates when you bundle your Example accounts. Visit a Example Bank to open an account now.">
<meta name="template" content="example-generic-page-template">
<meta name="viewport" content="width=device-width, initial-scale=1">
  <link href="https://www.example.com/us/en/personal-banking/ira/savings" rel="canonical">
```

HTML Stripped of CSS/JS/HTML Tags

High Interest Retirement Account Savings | Example Bank Select Savings Retirement Account Example Savings Retirement Account 1

Increase your earnings by maintaining a higher balance and connecting a qualifying Example account 2

\$10,000 Initial deposit requirement to open Rate enhancement Superior rates when you connect a qualifying Example account 2

No monthly charges No monthly maintenance charges Secure Initiate in a physical location Compare Benefits you receive Boost your earnings by maintaining a higher balance and connecting a qualifying account. 2

Retirement savings An Example Savings Retirement Account could be an optimal choice for those seeking a superior interest rate with the flexibility to add funds anytime \$10,000 initial deposit Permits additional contributions location You are examining info for _____ Modify your location location icon NA,NA Region

GPT-4 Fact Extraction

GPT-4 Fact-Extracted Content

Example Bank offers a high-yield retirement savings account titled the Example Savings Retirement Account. This account enables you to enhance your earnings by keeping a larger balance and connecting a qualifying Example account. To initiate this account, a minimum of \$10,000 is required. Additionally, this account is exempt from monthly maintenance fees.

The Example Savings Retirement Account is an excellent option for those seeking a competitive interest rate alongside the ability to contribute funds at any moment. Your interest rate escalates as your balance increases, aiding in maximizing your savings. An extra rate increase is available when you link a qualifying Example Bank mortgage, home equity loan, credit card, or an operational personal or small business checking account.

This account promotes the acceleration of your retirement savings with its tiered interest rates. It also provides tax advantages. For Traditional IRAs, earnings accumulate tax-deferred, and upon withdrawal, earnings are taxed as income. Contributions might be tax-deductible. For Roth IRAs, earnings accumulate tax-free, and earnings withdrawn are not subject to income tax if they meet the criteria for a qualified distribution.



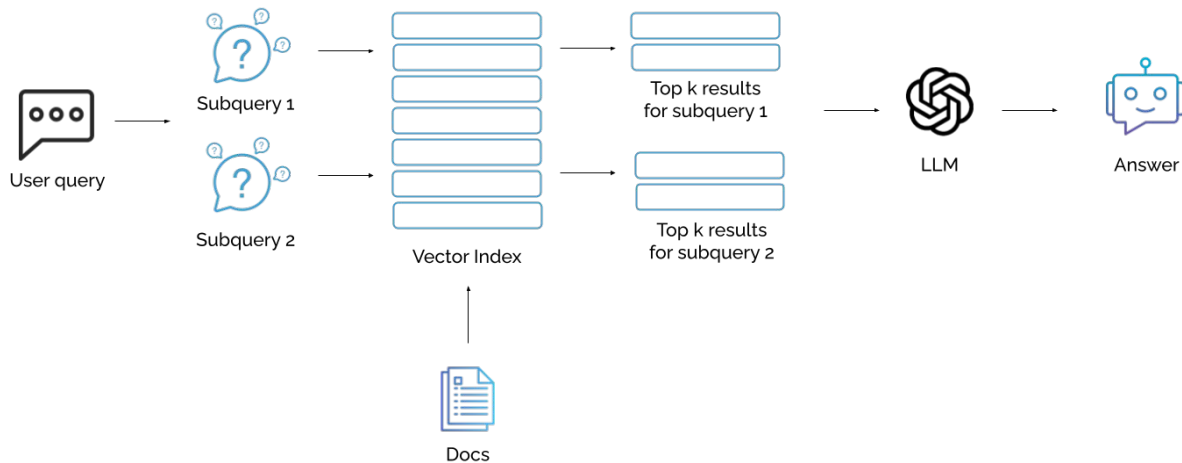
Caveat: Risk of Information Loss

PRE-RETRIEVAL OPTIMIZATION

Query Transformation - Query Expansion, Query Reformulation

Issues to address:

- Complex queries
- Ambiguous or poorly worded queries



PRE-RETRIEVAL OPTIMIZATION

Query Transformation - Optimize search query with conversation history as input

Customer: "What are the interest rates for your CDs?"

Assistant: "Our interest rates are XYZ."

Customer: "Which credit card is good for travel?"

Assistant: "The XYZ credit card is good for travel for ABC reasons"

Customer: "Tell me more about the interest rate for that"

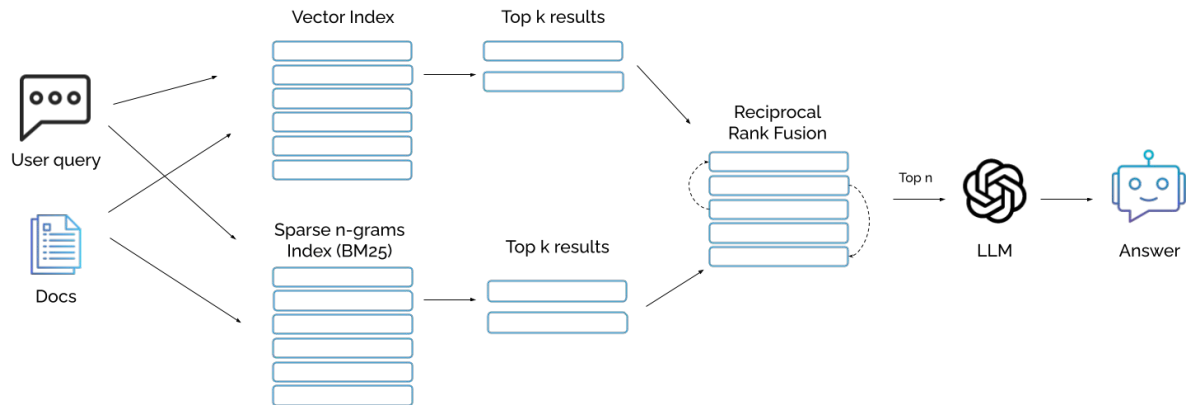
Prompt = 'You are examining a conversation between a customer of Example bank and an Example bank chatbot. A documentation lookup of Example bank's policies, products, or services is necessary for the chatbot to respond to the customer. Please construct a search query that will be used to retrieve the relevant documentation that can be used to respond to the user.'

RETRIEVAL OPTIMIZATION

Ensemble / Fusion Retrieval

- Combine multiple retrievers
- For eg: keyword-based search (BM25) + vector search
- Use Reciprocal Rank Fusion to merge results

$$\text{RRF Score} = \sum_{i=1}^N \left(\frac{\text{weight}_i}{\text{rank}_i + c} \right)$$

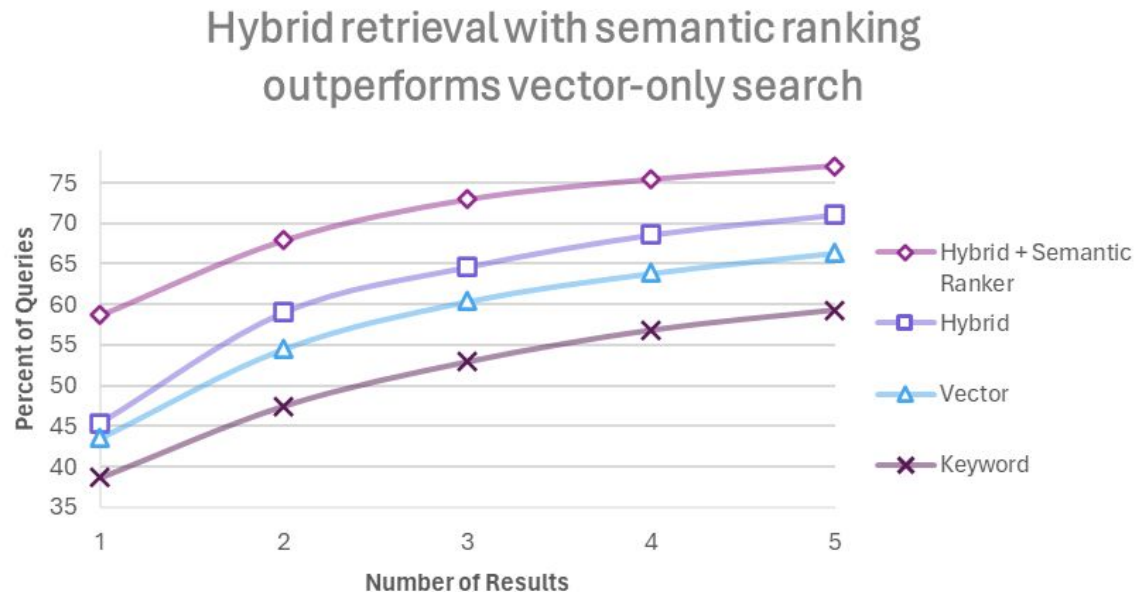


RETRIEVAL OPTIMIZATION

Ensemble / Fusion Retrieval

- Combine multiple retrievers
- For eg: keyword-based search (BM25) + vector search
- Use Reciprocal Rank Fusion to merge results

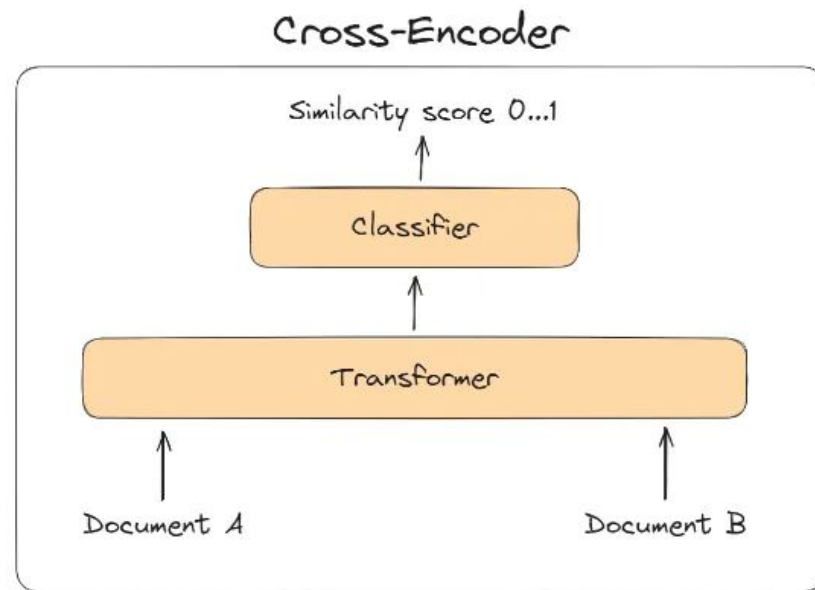
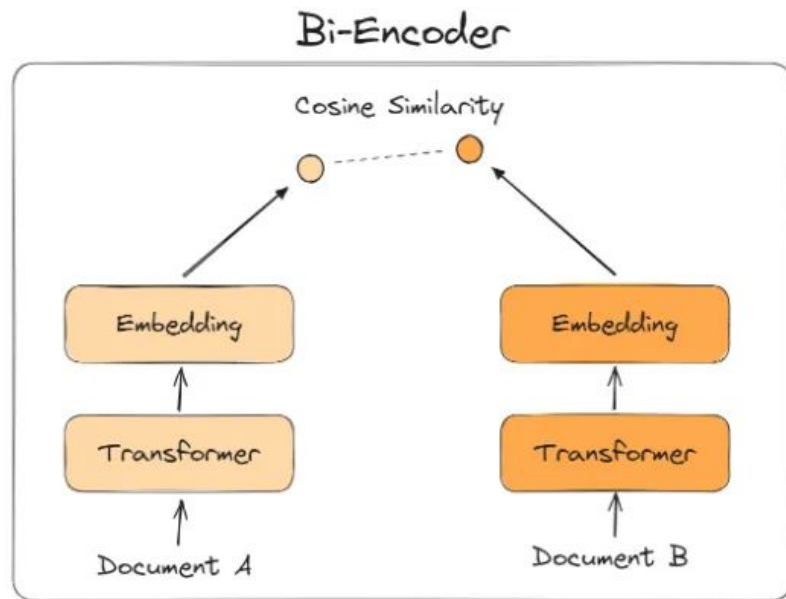
$$\text{RRF Score} = \sum_{i=1}^N \left(\frac{\text{weight}_i}{\text{rank}_i + c} \right)$$



Percentage of queries where high-quality chunks are found in the top 1 to 5 results

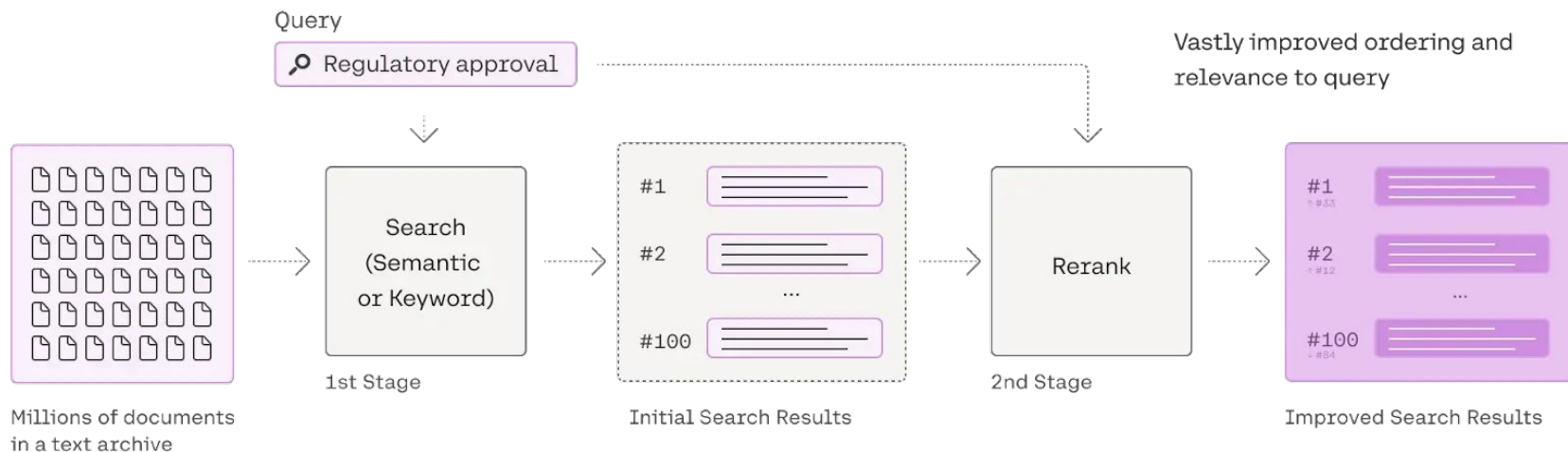
POST-RETRIEVAL OPTIMIZATION

Cross-encoder Re-ranking



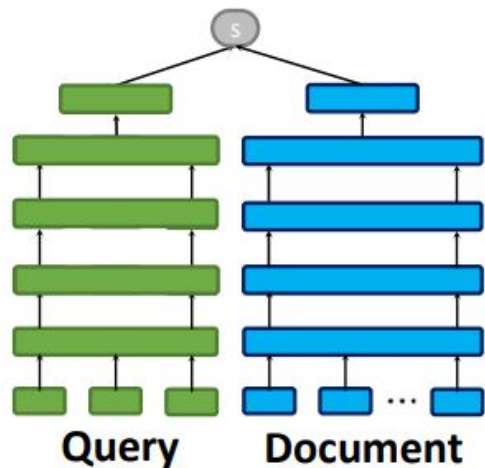
POST-RETRIEVAL OPTIMIZATION

Cross-encoder Re-ranking

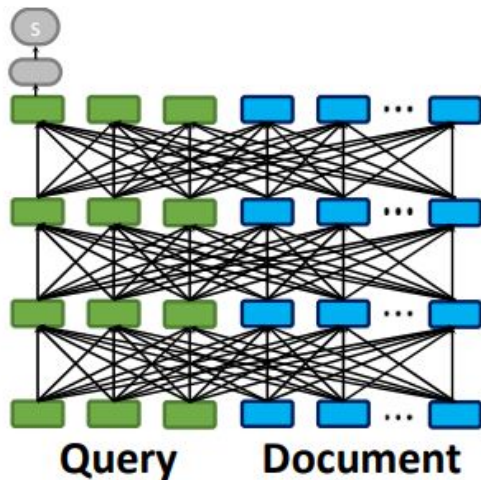


POST-RETRIEVAL OPTIMIZATION

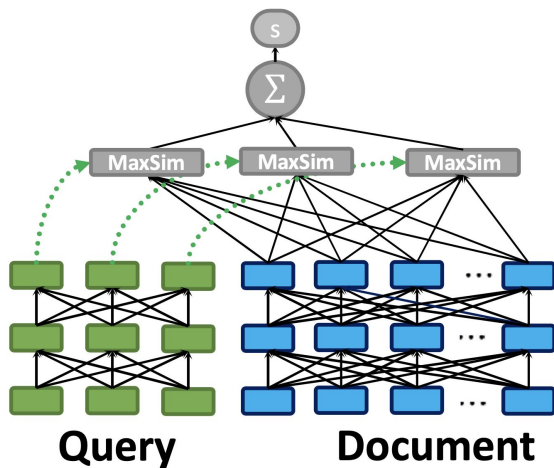
ColBERT model



Bi-encoder



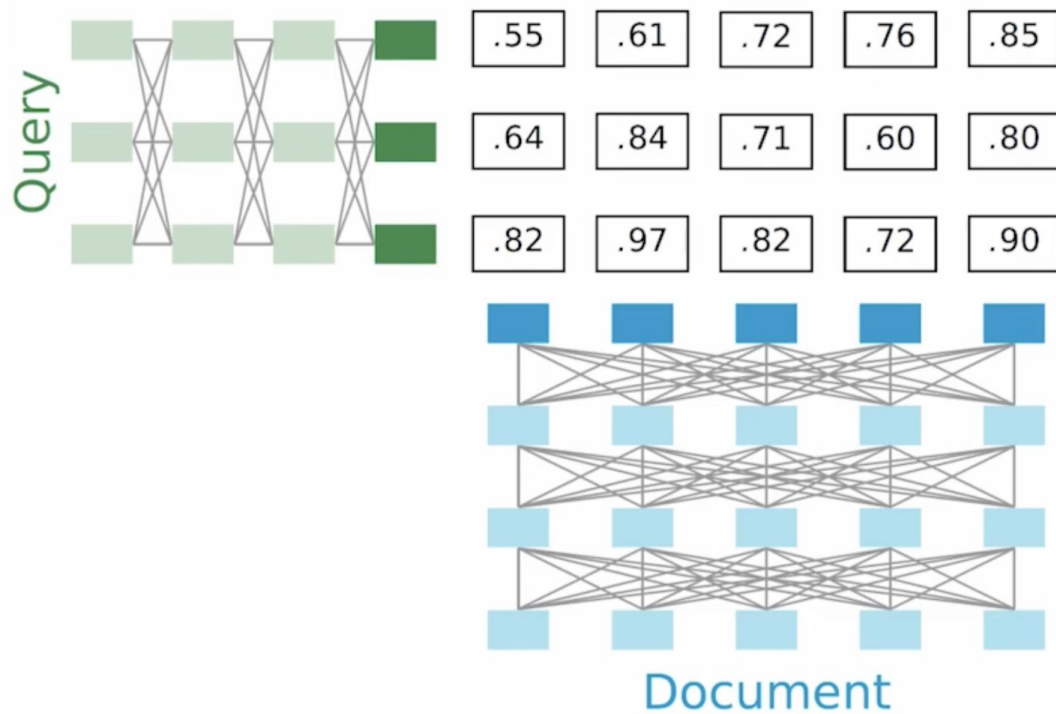
Cross-encoder



ColBERT

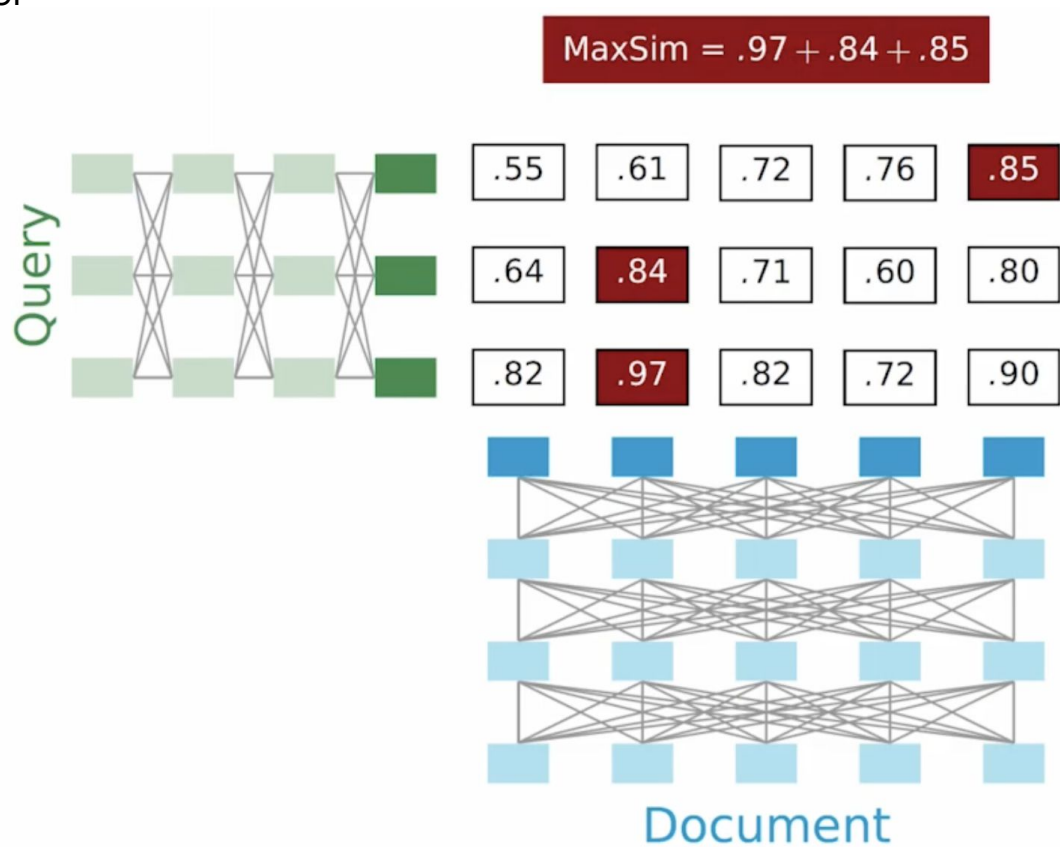
POST-RETRIEVAL OPTIMIZATION

ColBERT model



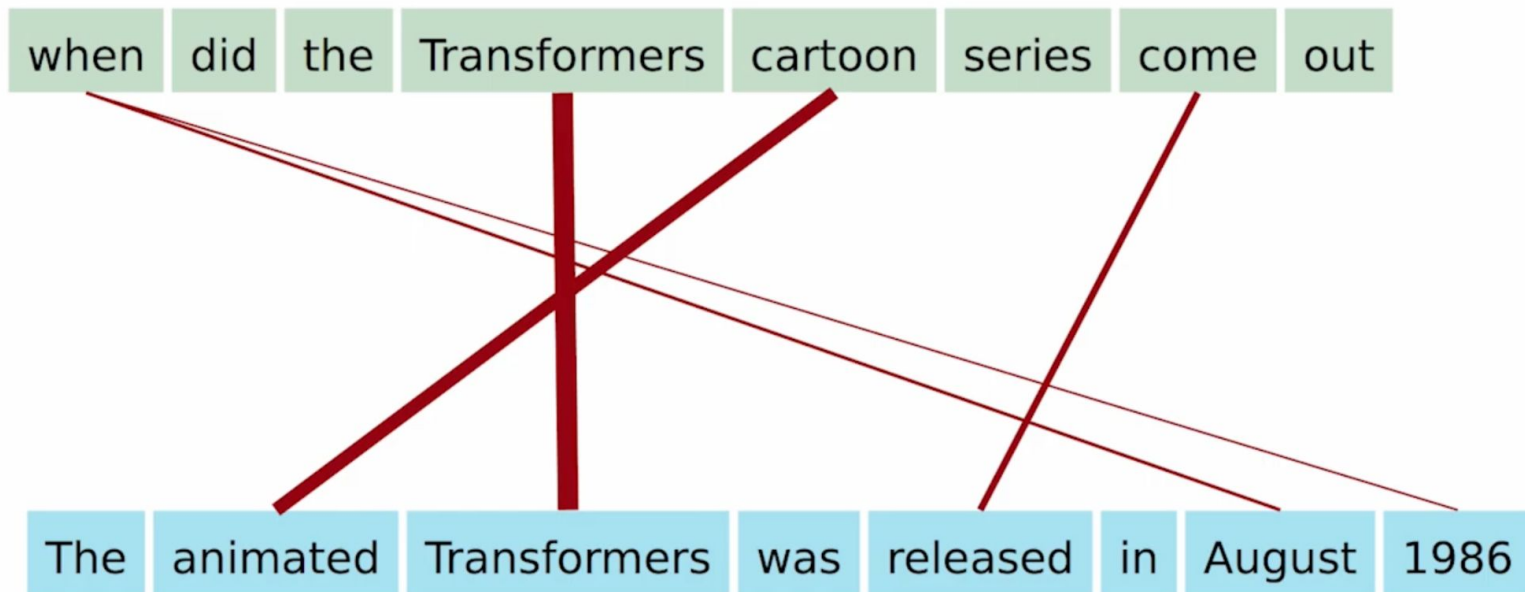
POST-RETRIEVAL OPTIMIZATION

ColBERT model



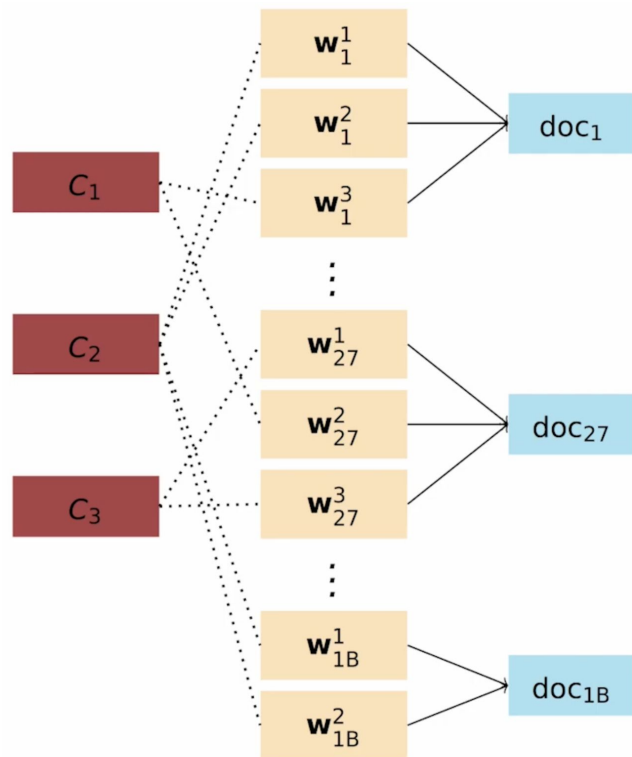
POST-RETRIEVAL OPTIMIZATION

ColBERT model



POST-RETRIEVAL OPTIMIZATION

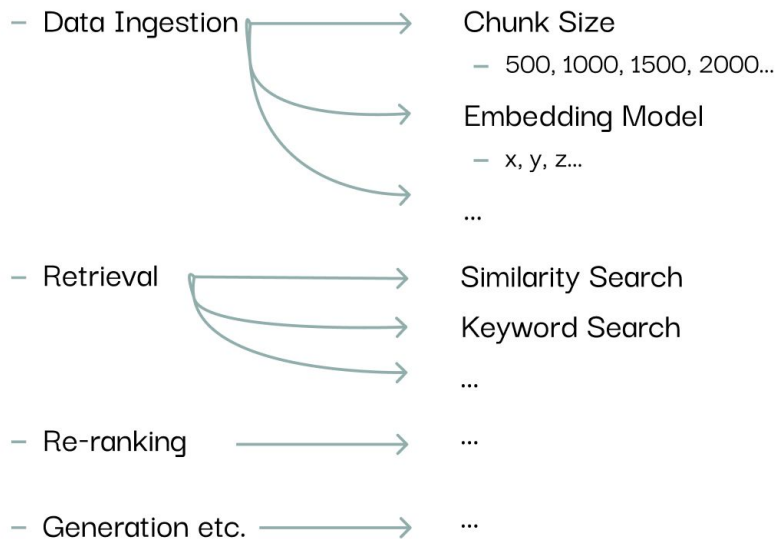
ColBERTv2 - Centroid based re-ranking



OVERALL...

Creating an Optimized RAG is really hard...

- Lots of moving parts with multiple parameters
- So many techniques & strategies
- Takes several months of trial & error



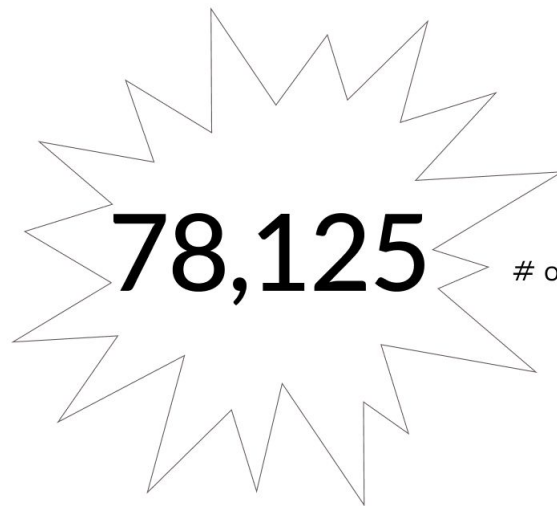
- 5 chunking methods,
- 5 chunk sizes,
- 5 embedding models,
- 5 retrievers,
- 5 re-rankers/compressors
- 5 prompts
- 5 LLMs

=

How many RAG
configurations?

- 5 chunking methods,
- 5 chunk sizes,
- 5 embedding models,
- 5 retrievers,
- 5 re-rankers/compressors
- 5 prompts
- 5 LLMs

=



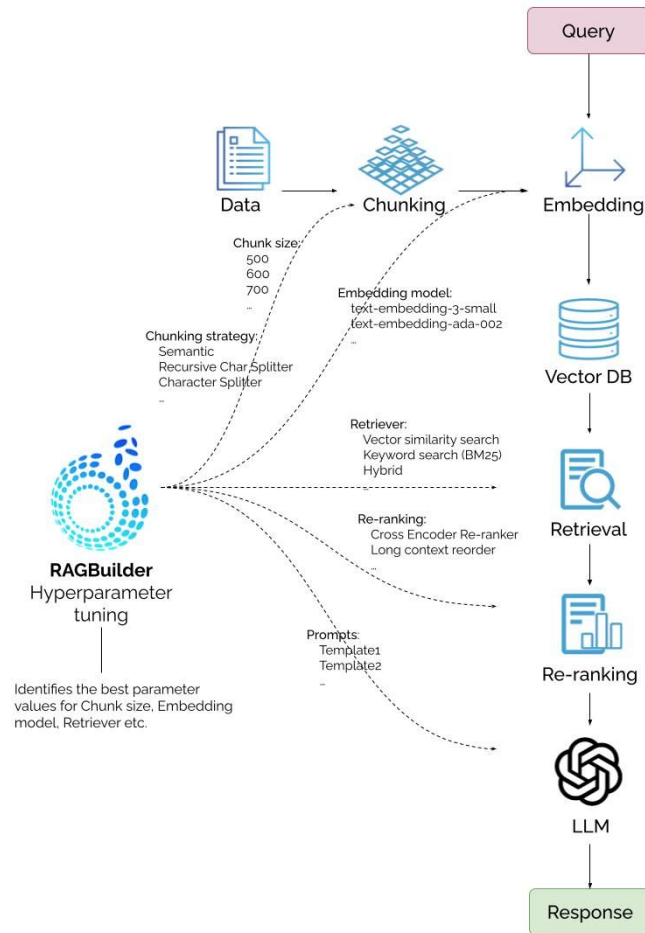
of RAG Configurations

5 mins per config => **271 days**



RAGBuilder

- ✓ Hyperparameter tuning
- ✓ SOTA predefined RAG templates
- ✓ Synthetic test data generation with auto re-use
- ✓ Build with the best models of your choice (open/closed)
- ✓ Knowledge base/ Graph RAG
- ✓ Bring your own RAG or component [coming soon]
- ✓ Cloud hosted solution [coming soon]

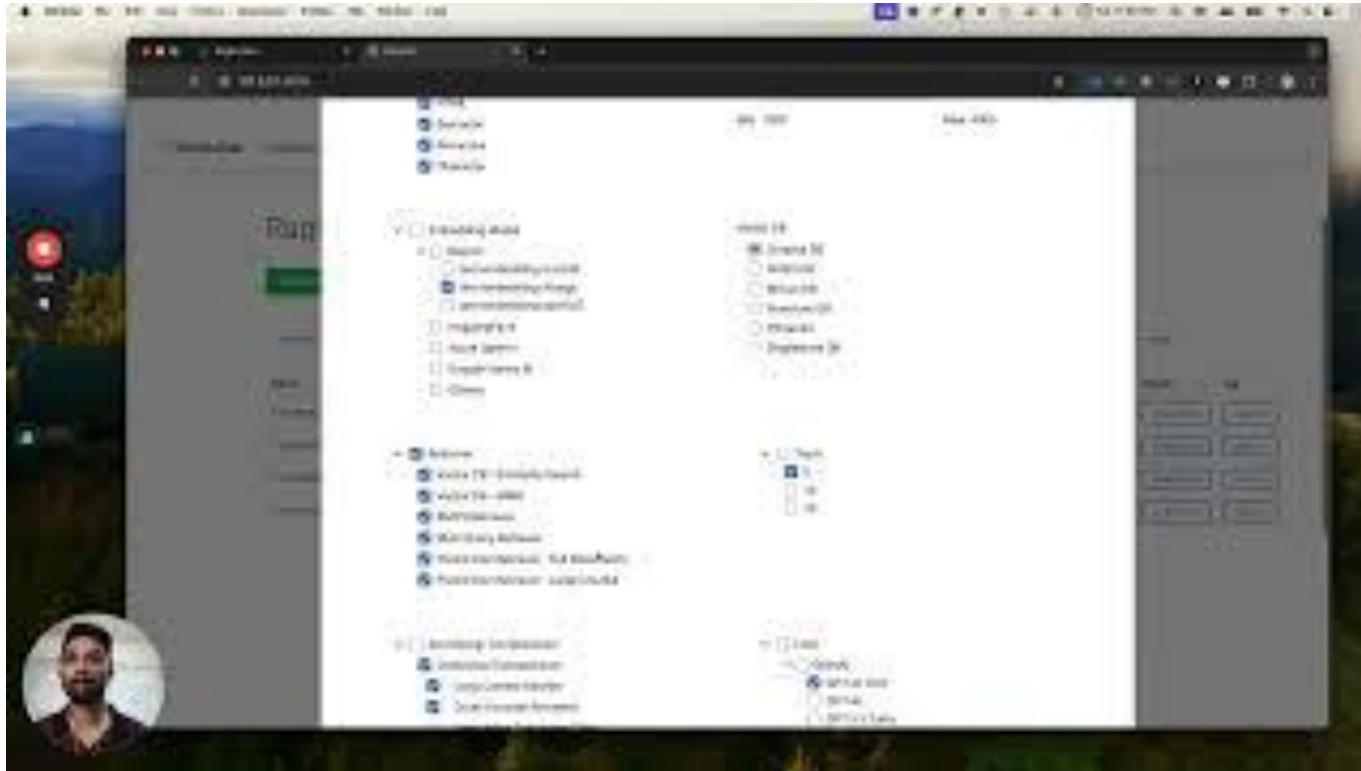


GITHUB TRENDING

#11 Repository Of The Day

Open source repo: <https://github.com/KruxAI/ragbuilder>

Demo time



<https://youtu.be/-vfBSNZuAPE>



Can you take this 1 min survey?