RAG Options

What the thing is: Choosing a RAG in today’s market

Options

* Cohere
* OpenAI
* Azure Search
* Google AI
* LangChain
* Llama Index
* Databricks

Summarise

* Background Sentence
* Who each tech is created for (App Builders | Data Engineers | DSs)
* Cost model (scale to zero)
* Abstraction

Strengths/Weaknesses

* Ease of Use (right abstraction etc)
* Operational Cost
* Trust
* Model Coupling
* Advanced Features (Hybrid)

Conclusion

Introduction

Wow what a difference a day makes!! When Vectara was founded the Retrieval Augmented Generation (RAG) was a very lonely space. But from the explosion of LLMs in 2022 have also seen the number of solutions which propose a RAG element to also increase. With this document, we aim to look at who are the major players in this space and do a feature comparison.

The Players

Like the introduction of the tributes in “The Hunger Games” here are the contenders. Amr would have been great representing District 12!!



Vectara

Founded in 2019 which has been designed from the ground up to be a focused RAG solution for application builders with easy APIs. The “Snowflake” of the category – purpose built to allow organisations to think about their IP in terms of Corpus with advanced features like ACLs and multi-corpus searches.

Cohere

Also founded in 2019 with multiple researchers from “Google Brain”, one of which was the author of the famous “Attention is All you Need” paper which introduced the transformer architecture and lay the groundwork for other heavy weights below. Partnered with Oracle to bring their technology into Oracle’s Fusion Cloud.

OpenAI:

They definitely have great brand recognition, though the name is kind of Ironic. Or as Elon Musk put it:

*OpenAI was created as an open source (which is why I named it “Open” AI), non-profit company to serve as a counterweight to Google, but now it has become a closed source, maximum-profit company effectively controlled by Microsoft.*

*Not what I intended at all.*

<https://twitter.com/elonmusk/status/1626516035863212034>

All that being said, they are the leaders in terms of General LLMs and have made moves towards RAG with their Assistant feature.

Azure Search

This is a “build it yourself” RAG system – achievable but you own the most of the steps and leveraging Open AI’s features to their Cognitive Search. I mean, at least they didn’t completely rip off one of their Open Source partners tech stacks and rebadge it as their own this time.

Google AI

TBD

LangChain

TBD

Llama Index

TBD

Databricks

I would be remiss not to also mention Databricks as a former employer. They have upped their game to allow Data Engineering and Data Science teams to collaborate and bring the benefits of LLMs to the Lakehouse.

Our Analysis

We’ll be primarily looking at how well these companies position themselves as a ready-to-go LLM solutions for Application Builders.

Comparison Criteria

The follow criteria are subjective comparisons between the organisations position with relation to serving a front-facing application operational workload. We will compare them against the following criteria:

* Library, PaaS or SaaS: SaaS scoring the best for application builders.
* For PaaS/SaaS, what do they handle (each gets a point):
  + Parsing
  + Encoding
  + Vector Storage
  + Retrieval
  + Prompt Engineering
  + LLM Execution
* Abstraction: How aligned to “Don’t make me think” are they for an Enterprise. We regard Documents/Corpus as the right level for most entities rather than Embeddings.
* Team Costs
  + Solution complexity
  + How many cooks to get started?
  + Operational costs to keep running?
* Trust
  + Who needs to perform the investigation – your legal/data scientists?
  + Who owns the responsibility?
* Advanced Rag Features
  + Filter Attributes?
  + Multi-Corpus Queries?
  + ACLs on the Corpus?

Scoring

Category

Through the lens of “SaaS specifically for serving RAG operationally”, Vectara is the only ready-to-go system for serving these workloads. The rest are either “platforms that can be used for” or “libraries” rather than a capability. No surprise, Vectara scored highest here.

Completeness

In the realm of “Don’t make me think”, how much of each platform is done for our users? Again Vectara is platform that “just works” whereas others require many of the additional steps to be both done and wired in. Again, Vecata scores well here. For example, Cohere is mature but you have to “do-it-yourself” to accomplish the major parts of RAG (parsing, encoding, vector storage and search): <https://docs.cohere.com/docs/retrieval-augmented-generation-rag#step-2-fetching-relevant-documents>

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Vendor | Parsing | Encoding | Vector Storage | Retrieval | Prompt Engineering | LLM Execution | Score |
| Vectara | Yes | Yes | Yes | Yes | Yes | Yes | 10.0 |
| Cohere | No | No | No | No | Yes | Yes | 3.3 |
| OpenAI | Yes | Yes | Yes | Partial | Yes | Yes | 9.2 |
| Azure Search\* | No | No | Yes | Partial | Partial | Partial | 4.2 |
| Google AI |  |  |  |  |  |  |  |
| LangChain |  |  |  |  |  |  |  |
| Llama Index |  |  |  |  |  |  |  |
| Databricks | No | No | Yes | Partial | Partial | Partial | 4.2 |

Note: Partial scores relate either to situations where you “need to build/maintain it” or “you can by integrating other parts of the stack, e.g. “Azure Search + Cosmos”. Open AI assistants do feature retrieval but only by vectors, there is no hybrid search or multi-corpus search.

Abstraction

In this fast moving vertical, the comparison is like looking at Databricks or Snowflake as a Data Warehouse (circa 2021). You could achieve Data Warehousing with Databricks in 2021 but there was a lot of banging, tears and a fair few sharp corners to get something equivalent to Snowflake. Databricks bridged that gap with Serverless and Unity Catalog, but kudos where kudos is warranted, Snowflake is still the optimum level of abstraction for many Data Warehouse folk coming from 20 years on legacy platforms. Vectara’s goal is to be the Snowflake of RAG, being the equivalent for Application Builders.

Team Costs

How many cooks does it take to get your RAG pipeline working … and to continue working. This measure is subjective but based on my experiences working with technical teams across my career.

Trust

The scoring metric here was applied in relation to how much you trust the Vendor to provide reliable responses for both the retrieval and summarisation phases of the RAG pipeline. If the onus is entirely on the client to perform their own evaluation, this scores poorly amongst “application builders” who just want something that works … and works well. Though Vectara scored well (I mean, yes we’re a bit biased) we still didn’t put a 10 as the market vertical is still too new – maybe after another year of great results and expanded use of Boomerang we’ll self-report a 9.

Advanced RAG Features

What additional features doe the vendor support specifically for RAG?

Result Summary

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Vendor | Category | Completeness | Abstraction | Team Costs | Trust | Adv Features |
| Vectara | SaaS (10) | 10.0 | Corpus (10) | Low (10) | 8 | 10 |
| Cohere | PaaS(5) | 3.3 | Low Level (6) | Mid (5) | 3 | 1 |
| OpenAI | SaaS(10) | 9.2 | Assistant (7) | Mid (5) | 6 | 3 |
| Azure Search | PaaS(5) | 4.2 | 3 | High (2) | 5 | 2 |
| Google AI |  |  |  |  |  |  |
| LangChain |  |  |  |  |  |  |
| Llama Index |  |  |  |  |  |  |
| Databricks | PaaS(5) | 4.2 | Low Level (4) | High (2) | 5 | 3 |

Databricks does a little better here with lineage from Unity Catalog lineage and Model Monitoring – two of the best features I was sad to leave behind. Open AI, whilst it has an assistant, is not aligned to a corpus abstraction and lacks expected maturity which also affects other categories.

Final Scores (recalculate with missing):

1. First: Vectara (58)
2. Second: Open AI (40.2)
3. Third: Cohere (23.3)
4. Fourth: Databricks (23.2)
5. Fifth: Azure Search (21.2)

Conclusion

I mean, we’re tooting our own horn and came out first? Who would have thought. But seriously though: Vectara’s RAG features are on another level and the focused delivery of a platform for application builders by application builders has resulted in an incredibly easy path to Generative AI solutions.