

# Pragmatic AI-powered Search

Keeping it Simple, not Stupid


[opensourceconnections.com](https://opensourceconnections.com)

---



*Charlie Hull, Search Solutions, 22nd November 2023*

# Who am I?

- Started in search in 1999
- Ran  , UK consultancy for 17 years  
The Open Source Search Specialists
- Co-author of 'Searching the Enterprise'
- Founding member of The Search Network
- Project lead, manager, trainer, speaker, blogger
- Past lives included electronic engineering ....  
& circus performer

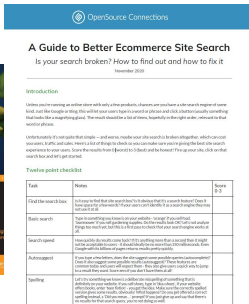
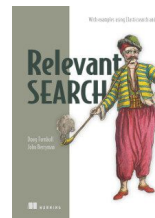


# Who are OpenSource Connections ?

- A [team of search engine specialists](#) with deep knowledge & decades of experience
- Our approach is data-driven, scientific and focused on business needs
- We help [organizations in the USA and EU](#) build powerful, scalable, accurate and relevant search applications by **empowering their teams to succeed** with a mix of [tools](#), [processes](#), [training](#) and [consultancy](#)

[www.opensourceconnections.com](http://www.opensourceconnections.com)

- We write books & reports



- We host leading search events...



[www.haystackconf.com](http://www.haystackconf.com)

...and present at many more

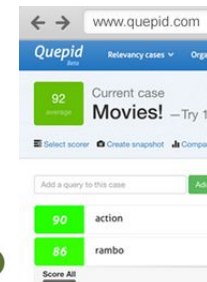


- We build open source, free tools for tuning search engines

```
1 {
2   "query": {
3     "query": {
4       "matching_query": {
5         "query": "notebook"
6       },
7       "query_fields": [ "title^3.0", "brand^2.1", "shortSummary",
8         "rewriters": [ "common_rules" ]
9     }
10  }
11 }
```



Elasticsearch Learning to Rank GO





**OTTO**

**BRACK.CH**

**POND5**

**RUBIX**



**uspto**

# Pragmatic AI-powered Search

Keeping it Simple, not Stupid

---

1. Why everything's changed
2. What this means for search
3. Plus ça change...
4. Pragmatic solutions
5. The importance of measurement
6. Takeaways

# Pragmatic AI-powered Search

Keeping it Simple, not Stupid

- 
1. Why everything's changed
  2. What this means for search
  3. Plus ça change...
  4. Pragmatic solutions
  5. The importance of measurement
  6. Takeaways

*What do I mean by Pragmatic?*

- *I'm an engineer. I want to build real things, not pie in the sky*
- *I work in marketing, but I don't like buzzwords and false promises*
- *I care about making it easier for search teams to achieve their goals*
- *I like process, data and 'open' ways of working*

# Caveat

---

Everything is changing so quickly I'm sure to have missed some new developments!

# Why everything's changed

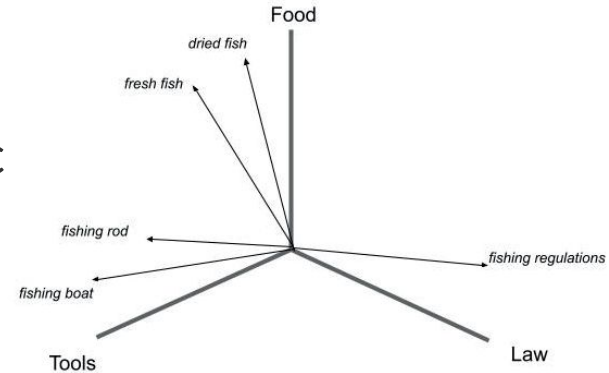


AI, LLMs, GPT, buzzword soup and an 'open' side order

# What was hard is now easy (apparently)

---

- We can capture semantics with embeddings
  - Turns content (words, images, paragraphs) into multidimensional magic numbers - **vectors**
  - Uses a machine learning model
  - Places each piece of content into a **vector space**
  - Store embeddings in a **vector database**



<https://simonwillison.net/2023/Oct/23/embeddings/>

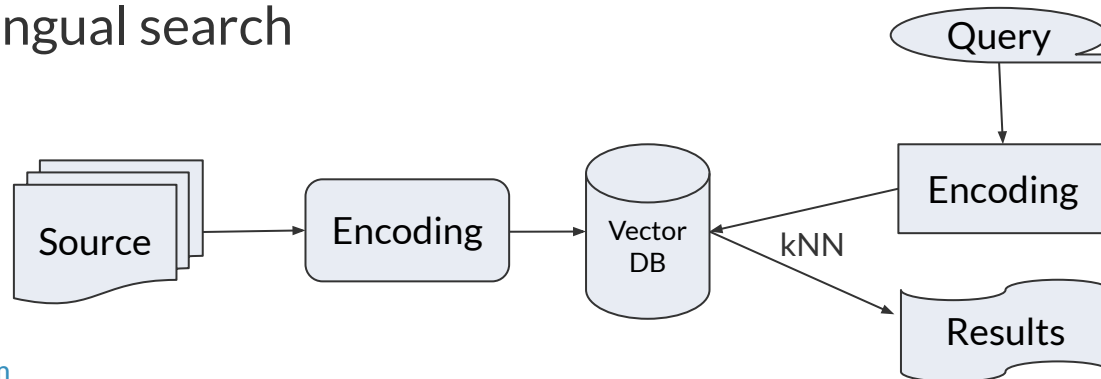
<https://vickiboykis.com/what-are-embeddings/>

<https://ig.ft.com/generative-ai/>

# What was hard is now easy (apparently)

---

- We can capture semantics with embeddings
- Vector databases can be used for search (also similarity, clustering...)
  - Increased recall, no need for exact word matching
  - Multimodal, e.g. search for images with text
  - Multilingual search



# What was hard is now easy (apparently)

---

- We can capture semantics with embeddings
- Vector databases can be used for search (also similarity, clustering...)
- Machines can now write plausible text (and poems, software, ...)
  - Language models are great at completion!
  - Summarization from sources
  - NOTE: plausible does not mean good or correct

# What was hard is now easy (apparently)

---

- We can capture semantics with embeddings
- Vector databases can be used for search (also similarity, clustering...)
- Machines can now write plausible text (and poems, software, ...)
- True conversations can be had with a computer
  - Models can have some short-term memory
  - You can chat to your PDFs!
  - Prompt engineering lets one craft answers and carry out tasks

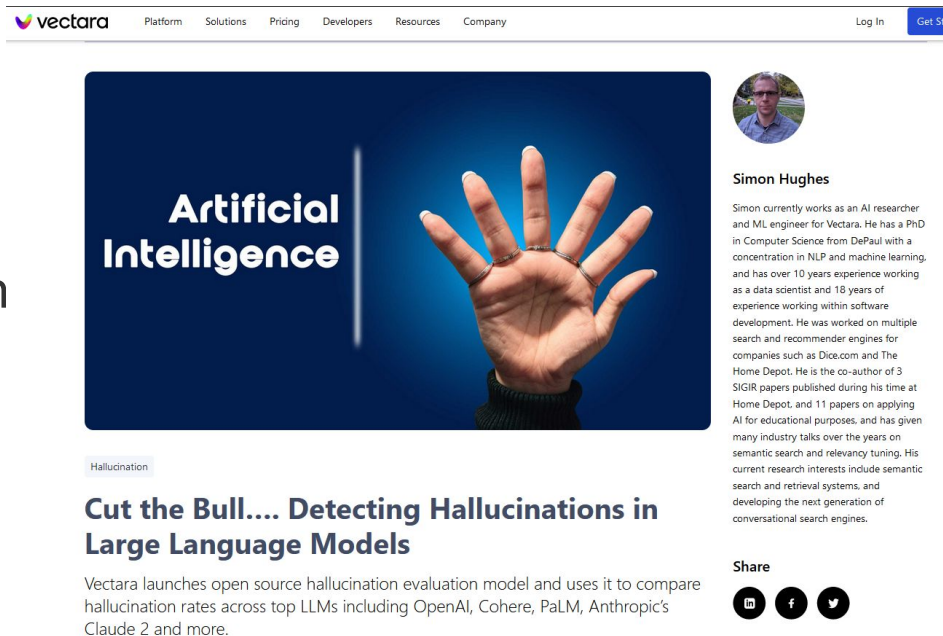
# What was hard is now easy (apparently)

---

- We can capture semantics with embeddings
- Vector databases can be used for search (also similarity, clustering...)
- Machines can now write plausible text (and poems, software, ...)
- True conversations can be had with a computer
- General Purpose AI is coming real soon now!

# It's not all good...

- Hallucination
  - LLMs don't "know" when they're wrong
  - Very dangerous in some settings
  - Some are trying to measure this



The screenshot shows a blog post on the Vectara website. The header includes the Vectara logo and navigation links for Platform, Solutions, Pricing, Developers, Resources, and Company. A user profile for Simon Hughes is visible on the right. The main content area features a large image of a hand with the text 'Artificial Intelligence' overlaid. Below the image is the article title 'Cut the Bull... Detecting Hallucinations in Large Language Models' and a short introductory paragraph.

**Artificial Intelligence**

**Cut the Bull... Detecting Hallucinations in Large Language Models**

Vectara launches open source hallucination evaluation model and uses it to compare hallucination rates across top LLMs including OpenAI, Cohere, PaLM, Anthropic's Claude 2 and more.

<https://vectara.com/cut-the-bull-detecting-hallucinations-in-large-language-models/>

# Supermarket AI meal planner app suggests recipe that would create chlorine gas

**Pak 'n' Save's Savey Meal-bot cheerfully created unappealing recipes when customers experimented with non-grocery household items**



📷 An app launched by a New Zealand supermarket that produces AI-generated recipes for leftovers has recommended cooks try 'bleach-infused rice surprise' among other things.  
Photograph: Jacobs Stock Photography Ltd/Getty Images

A **New Zealand** supermarket experimenting with using AI to generate meal

## Supermarket AI n suggests recipe th chlorine gas

**Pak 'n' Save's Savey Meal-bot chee recipes when customers experim household items**



📷 An app launched by a New Zealand super leftovers has recommended cooks try 'blea Photograph: Jacobs Stock Photography Ltd

A **New Zealand** supermarket exper

# Lawyers have real bad day in court after citing fake cases made up by ChatGPT

Lawyers fined \$5K and lose case after using AI chatbot "gibberish" in filings.

JON BRODKIN - 6/23/2023, 1:32 PM





5 months ago



What tools can be used to collect human judgements in search?



APP 5 months ago



Collecting human judgments for search relevance evaluation is an essential step in improving search quality. Several tools can be used to gather human judgments for search results:

1. Relevance Assessment Tool (RAT): RAT is an open-source tool to collect human judgments for search results. It provides a web interface and exports the collected data for further analysis. You can find the tool here:  
[https://github.com/o19s/relevance\\_assessment\\_tool](https://github.com/o19s/relevance_assessment_tool)

**THIS IS A LIE!!!!!!**

2. Quepid: Quepid is a search relevance evaluation tool that helps you collect human judgments, test different search algorithms, and track search relevance improvements over time. It supports multiple search engines, including Solr and Elasticsearch (or OpenSearch). Quepid offers a user-friendly interface for assessors to rate search results and visualize the impact of algorithm changes. You can find the tool here: <https://quepid.com/>

3. CrowdFlower (now Figure Eight): Figure Eight (formerly CrowdFlower) is a data annotation platform that helps you collect human judgments for search results by leveraging a large crowd of annotators. You can create custom tasks for

# It's not all good...

---

- Hallucination
- Bias
  - Training data contains biases
  - *“For example, if the training data primarily comprises data from Western countries, the AI may struggle to produce accurate and culturally relevant content for non-Western audiences. This omission perpetuates societal inequalities and prevents the AI system from being an inclusive and unbiased information source.”*

<https://www.forbes.com/sites/forbestechcouncil/2023/09/06/navigating-the-biases-in-llm-generative-ai-a-guide-to-responsible-implementation/>

# It's not all good...

---

- Hallucination
- Bias
- Existential fear
  - Generated by those who want to control the market?
  - (A)I'll be back!

# It's not all good...

---

- Hallucination
- Bias
- Existential fear
- Panicked regulation (do governments even understand AI?)

**WIRED**

LONG READS BUSINESS CULTURE GEAR SCIENCE SECURITY VIDEO

PETER GUEST

BUSINESS 23.10.2023 07:00 AM

## Britain's Big AI Summit Is a Doom-Obsessed Mess

UK prime minister Rishi Sunak's global summit on AI governance will focus on extreme scenarios of algorithms causing harm. Many British AI experts would rather he focus on near-term problems.

# It's not all good...

---

- Hallucination
- Bias
- Existential fear
- Panicked regulation
- License confusion
  - Meta's Llama2 is open source says Yann LeCun! Nope...
  - <https://www.linkedin.com/pulse/why-some-new-open-ish-ai-licenses-may-fly-enterprise-charlie-hull-umvxe>

# It's not all good...

---

- Hallucination
- Bias
- Existential fear
- Panicked regulation
- License confusion
- Crazy money
  - FOMO from VCs



Login

Search Q

TechCrunch+

## France's Mistral AI blows in with a \$113M seed round at a \$260M valuation to take on OpenAI

Ingrid Lunden @ingridlunden / 5:02 PM GMT+1 • June 13, 2023

 Comment

AI is well and truly off to the races: [a startup that is only four weeks old](#) has picked up a \$113 million round of seed funding to compete against OpenAI in the building, training and application of large language models and generative AI.

# It's not all good...

---

- Hallucination
- Bias
- Existential fear
- Panicked regulation
- License confusion
- Crazy money
- Huge, planet-destroying compute



<https://greensoftware.foundation/>

# What this means for search



...and search people!

Search is now easy!  
Hooray!  
We can all go home!

# Vector search isn't magic

---

- Exciting as it is, it can't do everything:
  - Exact searches (part numbers)
  - Better recall with the price of less precision (k in kNN)

# Vector search isn't magic

---

- Exciting as it is, it can't do everything
- A **hybrid** approach is worth considering for real-world applications
  - Blend results e.g. with RRF
  - Two-pass
  - Fallback (e.g. for zero results)
  - K in kNN issue

**Reciprocal Rank Fusion**

**How to Stop Worrying about Boosting**

**Philipp Krenn**

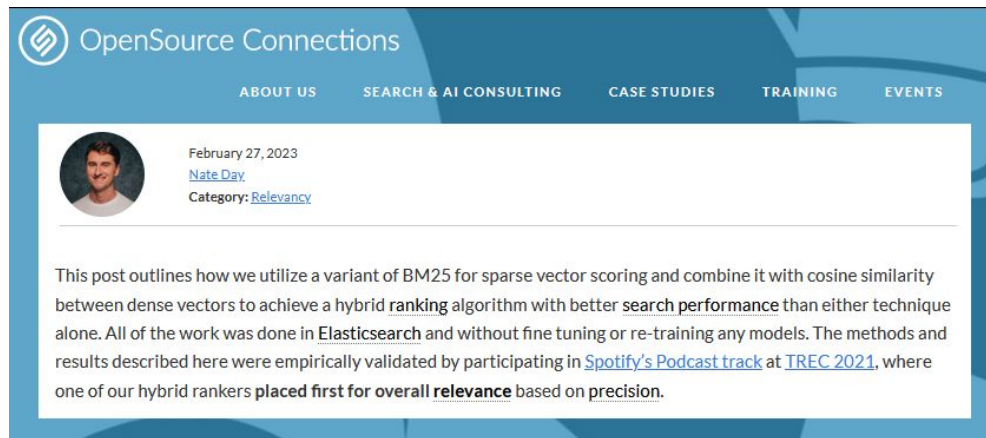
**@xeraa**

<https://haystackconf.com/eu2023/talk-2/>

# Vector search isn't magic





- Exciting as it is, it can't do everything
- A hybrid approach is worth considering for real-world applications
  - OSC “won”\* a TREC track using a hybrid approach in 2021:

\*sorry, Ellen



# Technology choices - everyone's adding vectors!

---

- Search engines   Elasticsearch 
- Open source (e.g. Lucene-based)
  - Not as performant as vector-first options? See <https://ann-benchmarks.com/>
  - You get all the standard stuff as well
- Vespa 
  - Do-everything, but complex and small community
  - Recently spun out of Yahoo
- Commercial
  - Algolia NeuralSearch - very early
  - Sinequa, Coveo etc. all adding vectors

# Technology choices - everyone's adding vectors!

- Search engines
- Vector-first databases
  - Weaviate, Qdrant, Pinecone, Milvus...
  - Some also have lexical (BM25) features
  - Often lack traditional search features like facets, spelling suggestion, language analysis
  - Creating some great education, lots of Devrel people out there



## Not All Vector Databases Are Made Equal

A detailed comparison of Milvus, Pinecone, Vespa, Weaviate, Vald, GSI and Qdrant



Dmitry Kan · Follow

Published in Towards Data Science · 7 min read · Oct 2, 2021

<https://towardsdatascience.com/milvus-pinecone-vespa-weaviate-vald-gsi-what-unites-these-buzz-words-and-what-makes-each-9c65a3bd0696>

# Technology choices - everyone's adding vectors!

---

- Search engines
- Vector-first databases
- Databases
  - Redis, Cassandra, even Oracle have added vector features
  - Open source `pgvector` lets you add vectors to Postgres
  - Not as performant as vector-first options? See <https://ann-benchmarks.com/>

pgvector/pgvector

Open-source vector similarity search for Postgres



14

Contributors

10

Issues

7k

Stars

275

Forks



# Technology choices - everyone's adding vectors!

- Search engines
- Vector-first databases
- Databases
- Frameworks
  - Jina.ai, LlamaIndex and more - toolkits for building some kinds of search
  - Lots of great examples & Python notebooks
  - More useful for prototypes than production

---

## Welcome to LlamaIndex !

LlamaIndex is a data framework for LLM-based applications to ingest, structure, and access private or domain-specific data. It's available in Python (these docs) and [Typescript](#).



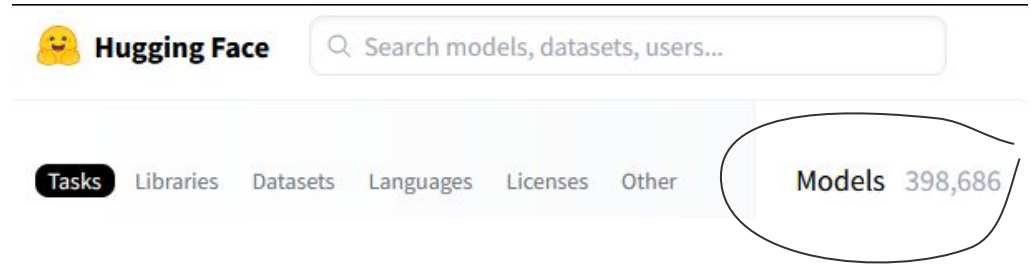
### Why LlamaIndex?

LLMs offer a natural language interface between humans and data. Widely available models come pre-trained on huge amounts of publicly available data like Wikipedia, mailing lists, textbooks, source code and more.

# Technology choices - everyone's adding vectors!

---

- Search engines
- Vector-first databases
- Databases
- Frameworks
- Models
  - Many choices, new ones every day trained in different ways
  - HuggingFace a great place for open source ones <https://huggingface.co>
  - Small, large and very large (bigger not always better)
  - Some build their own (e.g. Bloomberg) but this needs a lot of investment
  - You can fine-tune in-house but this will also need a ML team and time



# Technology choices - everyone's adding vectors!

- Search engines
- Vector-first databases
- Databases
- Frameworks
- Models
- APIs
  - Use these to generate embeddings, summarize etc.
  - OpenAI the leader also with ChatGPT and now GPTs you can create
  - Microsoft, Amazon etc. also offering many APIs
  - Lots of startups too



Credit: <https://twitter.com/peterkwwells>

Maybe we should have open models so we don't have to depend on an external company?

**The Guardian**

Microsoft hires Sam Altman to lead advanced AI research team after OpenAI ousting - business live

# LLM/GPT Risks

---

- Lack of knowledge
  - “A key challenge of these things is that they do not come with a manual! They come with a “Twitter influencer manual” instead, where lots of people online loudly boast about the things they can do with a very low accuracy rate, which is really frustrating.”  
<https://simonwillison.net/2023/Aug/3/weird-world-of-llms/>

# LLM/GPT Risks

---

- Lack of knowledge
- Explainability & user experience
  - Why did we get this answer? Because embeddings...
  - Hallucination

# LLM/GPT Risks

---

- Lack of knowledge
- Explainability & user experience
- Not trained on your data
  - Internal company terms/acronyms
  - General case vs. specific case
  - Fine-tuning is hard & expensive

# LLM/GPT Risks

---

- Lack of knowledge
- Explainability & user experience
- Not trained on your data
- Security & access control
  - Filtering is hard
  - Are you happy sending data to third parties?

# Plus ça change...

---

# It's a brave new world!

---

- If you're an experienced search person, this may all seem very new
  - Will AI replace search?
  - Will I be out of a job?

# It's a brave new world!

---

- If you're an experienced search person, this may all seem very new
- But....

▲ softwareDoug 47 minutes ago | prev | next [-]

Everyone I talk to who is building some vector db based thing sooner or later realizes they also care about the features of a **full-text search engine**.

They care about filtering, they care to some degree about direct lexical matches, they care about paging, getting groups / facet counts, etc.



# Things that are still important to understand


---

- Data quality
  - Even more important as we now might need it for model training
  - Can be enhanced or even synthetically generated with LLMs (if you trust them!)
  - Watch for bias

# Things that are still important to understand

---

- Data quality
- Indexing
  - Language analysis & other pre-processing (cafe vs. café)
  - Vector indexes are slow to build & hard to update

“In vector databases, specifically those powered by HNSW (or other layered graph algorithms), however, you have to [rebuild the entire index](#) from the ground up.”  Pinecone

<https://www.pinecone.io/blog/hnsw-not-enough/>

# Things that are still important to understand

---

- Data quality
- Indexing
- Testing
  - Harder now we can't see *why* something matched
  - Return to the user problem: if this helps lawyers, maybe only a lawyer can tell if it's any good
  - The old IR evaluation methods still apply
  - LLMs can help (with caveats) e.g GPT for rating search results

<https://opensourceconnections.com/blog/2023/06/13/g-for-rating-search-results/>

# Remember

---

- Search hasn't gone away
  - It isn't being 'replaced'
  - We're getting some great new tools

# Remember

---

- Search hasn't gone away
- Many of those using AI for search haven't done search before
  - Just read the forum/Slack of a vector search product!
  - Need for experience and education

# Remember

---

- Search hasn't gone away
- Many of those using AI for search haven't done search before
- We've been using ML in search for years
  - Learning to Rank

# Remember

---

- Search hasn't gone away
- Many of those using AI for search haven't done search before
- We've been using ML in search for years
- Traditional techniques are still the best for many use cases
  - Exact matching
  - Perhaps use AI for edge cases or when all else fails

# Remember

---

- Search hasn't gone away
- Many of those using AI for search haven't done search before
- We've been using ML in search for years
- Traditional techniques are still the best for many use cases
- It's like 2012 all over again!



September 6, 2023

[Charlie Hull](#)

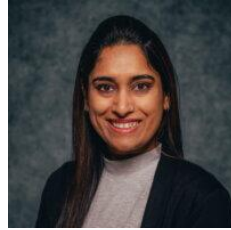
Category: [Big Data](#)

# Pragmatic solutions

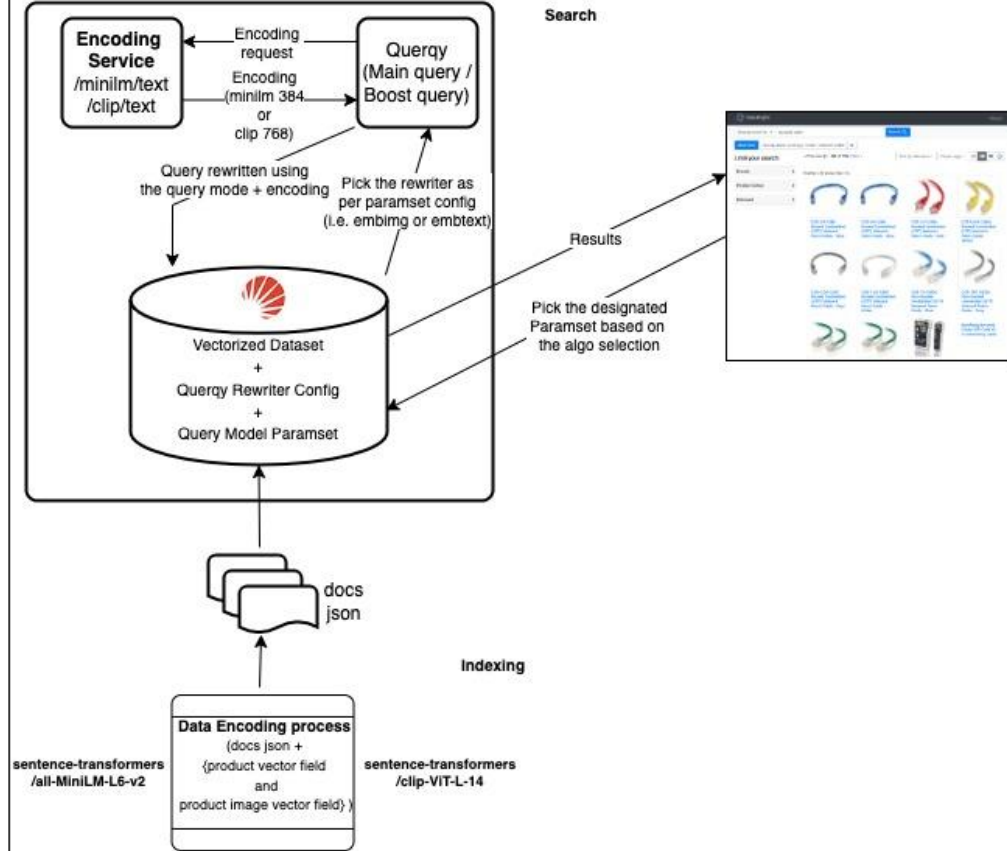


Some examples of what can be done with vectors & LLMs

# E-commerce search with vectors

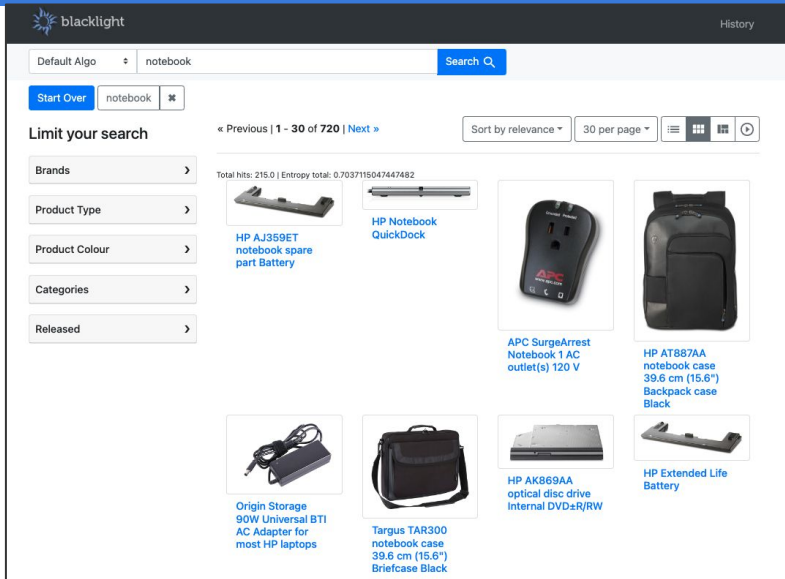


- Embedding service using FastApi
- Query rewriting using Querqy for respective models (CLIP , MiniLM and OpenAI)
- Vectorized ICECAT dataset using CLIP (image field) and MiniLM (text field - title + brand)
- Query models support
  - Match by Text Vector Search
  - Match by Image Vector Search
  - Boost by Text vectors
  - Boost by Image vectors



# Query : Notebook

<- Results with default algo



blacklight History

Default Algo notebook Search

Start Over notebook

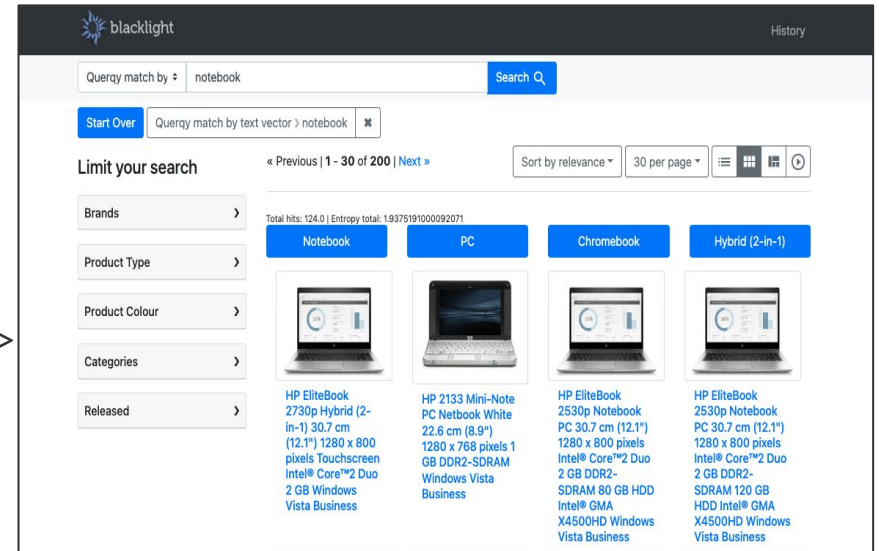
Limit your search

« Previous | 1 - 30 of 720 | Next » Sort by relevance 30 per page

Total hits: 215.0 | Entropy total: 0.7037115047447482

- HP AJ359ET notebook spare part Battery
- HP Notebook QuickDock
- APC SurgeArrest Notebook 1 AC outlet(s) 120 V
- HP AT887AA notebook case 39.6 cm (15.6") Backpack case Black
- Origin Storage 90W Universal BTI AC Adapter for most HP laptops
- Targus TAR300 notebook case 39.6 cm (15.6") Briefcase Black
- HP AK869AA optical disc drive Internal DVD±R/RW
- HP Extended Life Battery

Results with match by text vector ->



blacklight History

Query match by notebook Search

Start Over Query match by text vector notebook

Limit your search

« Previous | 1 - 30 of 200 | Next » Sort by relevance 30 per page

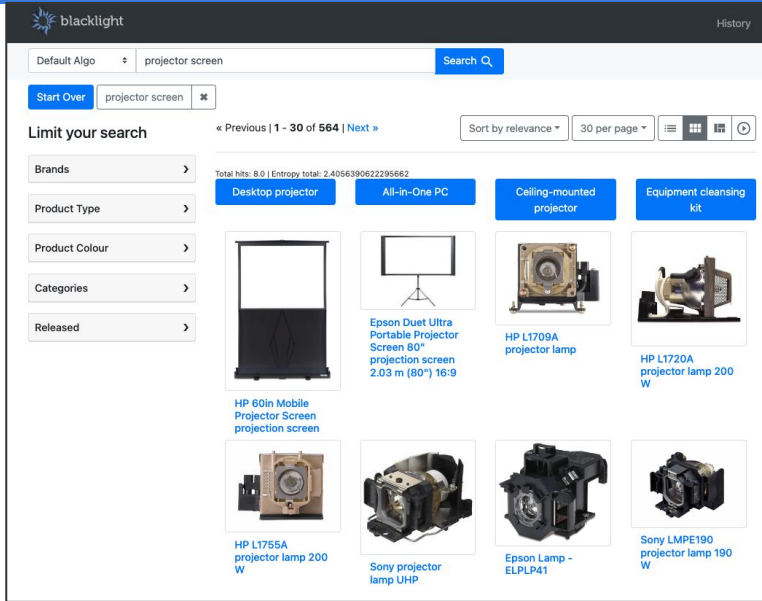
Total hits: 124.0 | Entropy total: 1.9375191000092071

Notebook PC Chromebook Hybrid (2-in-1)

- HP EliteBook 2730p Hybrid (2-in-1) 30.7 cm (12.1") 1280 x 800 pixels Touchscreen Intel® Core™2 Duo 2 GB Windows Vista Business
- HP 2133 Mini-Note PC Netbook White 22.6 cm (8.9") 1280 x 768 pixels 1 GB DDR2-SDRAM Windows Vista Business
- HP EliteBook 2530p Notebook PC 30.7 cm (12.1") 1280 x 800 pixels Intel® Core™2 Duo 2 GB DDR2-SDRAM 80 GB HDD Intel® GMA X4500HD Windows Vista Business
- HP EliteBook 2530p Notebook PC 30.7 cm (12.1") 1280 x 800 pixels Intel® Core™2 Duo 2 GB DDR2-SDRAM 120 GB HDD Intel® GMA X4500HD Windows Vista Business

# Query : Projector Screen

← Results with default algo



blacklight History

Default Algo projector screen Search Q

Start Over projector screen

Limit your search

« Previous | 1 - 30 of 564 | Next » Sort by relevance 30 per page

Total hits: 8.0 | Entropy total: 2.4056390622295662

- Desktop projector
- All-in-One PC
- Ceiling-mounted projector
- Equipment cleansing kit

HP 60in Mobile Projector Screen projection screen

Epson Duet Ultra Portable Projector Screen 80" projection screen 2.03 m (80") 16:9

HP L1709A projector lamp

HP L1720A projector lamp 200 W

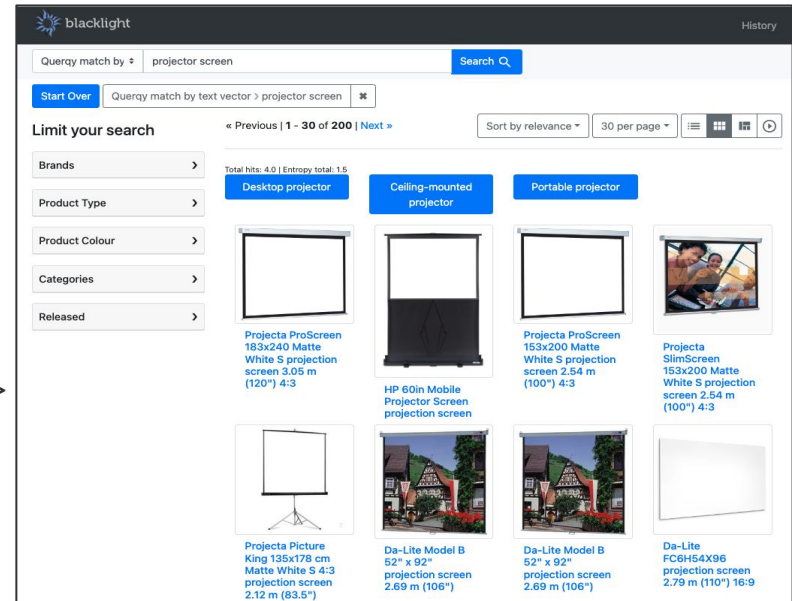
HP L1755A projector lamp 200 W

Sony projector lamp UHP

Epson Lamp - ELPLP41

Sony LMPE190 projector lamp 190 W

Results with match by text vector ->



blacklight History

Query match by projector screen Search Q

Start Over Query match by text vector > projector screen

Limit your search

« Previous | 1 - 30 of 200 | Next » Sort by relevance 30 per page

Total hits: 4.0 | Entropy total: 1.5

- Desktop projector
- Ceiling-mounted projector
- Portable projector

Projecta ProScreen 153x200 Matte White S projection screen 3.05 m (120") 4:3

HP 60in Mobile Projector Screen projection screen

Projecta ProScreen 153x200 Matte White S projection screen 2.54 m (100") 4:3

Projecta SlimScreen 153x200 Matte White S projection screen 2.54 m (100") 4:3

Projecta Picture King 135x178 cm Matte White S 4:3 projection screen 2.12 m (83.5")

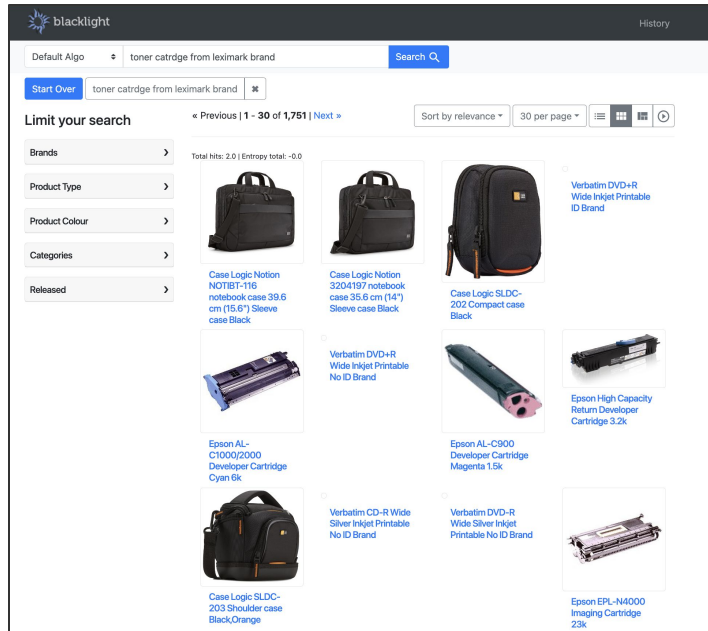
Da-Lite Model B 52" x 92" projection screen 2.69 m (106")

Da-Lite Model B 52" x 92" projection screen 2.69 m (106")

Da-Lite FC6H54X96 projection screen 2.79 m (110") 16:9

# Query : toner catrdge from leximark brand (with 2 misspells)

<- Results with default algo



blacklight History

Default Algo toner cartridge from leximark brand Search Q

Start Over toner cartridge from leximark brand

Limit your search < Previous | 1 - 30 of 1,751 | Next > Sort by relevance 30 per page

Brands >

Product Type >

Product Colour >

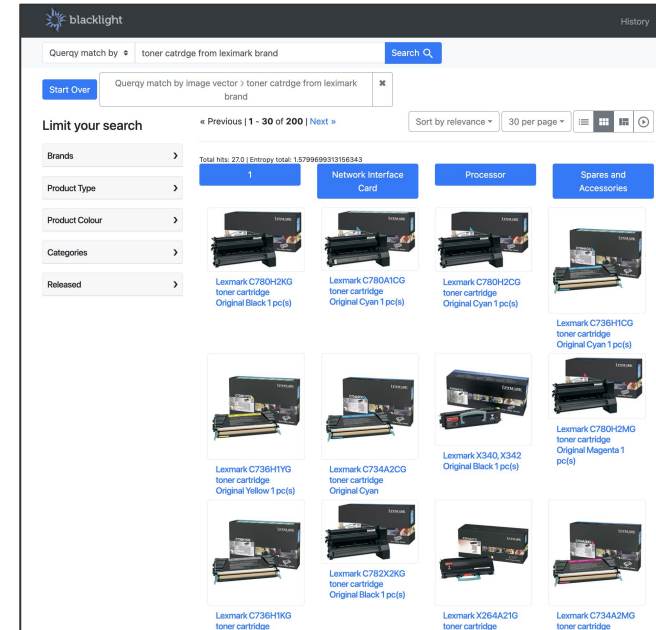
Categories >

Released >

Total hits: 20 | Entropy total: -0.0

- Verbatim DVD+R Wide Inkjet Printable ID Brand
- Case Logic Notion NOTBT-116 notebook case 39.6 cm (15.6") Sleeve case Black
- Case Logic Notion 3204197 notebook case 35.6 cm (14") Sleeve case Black
- Case Logic SLDC-202 Compact case Black
- Verbatim DVD+R Wide Inkjet Printable No ID Brand
- Epson High Capacity Return Developer Cartridge 3.2k
- Epson AL-C900 Developer Cartridge Magenta 1.5k
- Epson AL-C1000/2000 Developer Cartridge Cyan 0k
- Verbatim CD-R Wide Silver Inkjet Printable No ID Brand
- Verbatim DVD-R Wide Silver Inkjet Printable No ID Brand
- Epson EPL-N4000 Imaging Cartridge 23k
- Case Logic SLDC-203 Shoulder case Black/Orange

Results with match by image vector ->



blacklight History

Query match by toner cartridge from leximark brand Search Q

Start Over Query match by image vector > toner cartridge from leximark brand

Limit your search < Previous | 1 - 30 of 200 | Next > Sort by relevance 30 per page

Brands >

Product Type >

Product Colour >

Categories >

Released >

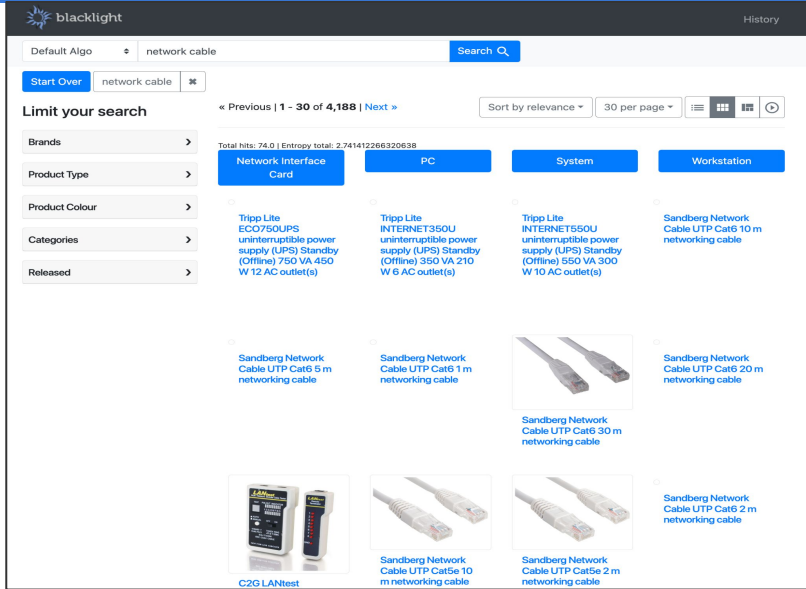
Total hits: 270 | Entropy total: 1.57996993136343

- 1
- Network interface Card
- Processor
- Spares and Accessories

- Lexmark C780H2KG toner cartridge Original Black 1 pc(s)
- Lexmark C780A1CG toner cartridge Original Cyan 1 pc(s)
- Lexmark C780H2CG toner cartridge Original Cyan 1 pc(s)
- Lexmark C736H1CG toner cartridge Original Cyan 1 pc(s)
- Lexmark C780H2MG toner cartridge Original Magenta 1 pc(s)
- Lexmark C736H1YG toner cartridge Original Yellow 1 pc(s)
- Lexmark C734A2CG toner cartridge Original Cyan
- Lexmark X340, X342 Original Black 1 pc(s)
- Lexmark C782X2KG toner cartridge Original Black 1 pc(s)
- Lexmark C736H1KG toner cartridge
- Lexmark X264A21G toner cartridge
- Lexmark C734A2MG toner cartridge

# Query : Network Cable

<- Results with default algo



blacklight

Default Algo network cable Search

Start Over network cable

Limit your search « Previous | 1 - 30 of 4,188 | Next » Sort by relevance 30 per page

Brands

Product Type

Product Colour

Categories

Released

Total hits: 74.0 | Entropy total: 2.741412266320638

Network Interface Card PC System Workstation

Tripp Lite ECO750UPS uninterruptible power supply (UPS) Standby (Offline) 750 VA 450 W 12 AC outlet(s)

Tripp Lite INTERNET350U uninterruptible power supply (UPS) Standby (Offline) 350 VA 210 W 6 AC outlet(s)

Tripp Lite INTERNET550U uninterruptible power supply (UPS) Standby (Offline) 550 VA 300 W 10 AC outlet(s)

Sandberg Network Cable UTP Cat6 10 m networking cable

Sandberg Network Cable UTP Cat6 5 m networking cable

Sandberg Network Cable UTP Cat6 1 m networking cable

Sandberg Network Cable UTP Cat6 20 m networking cable

Sandberg Network Cable UTP Cat6 30 m networking cable

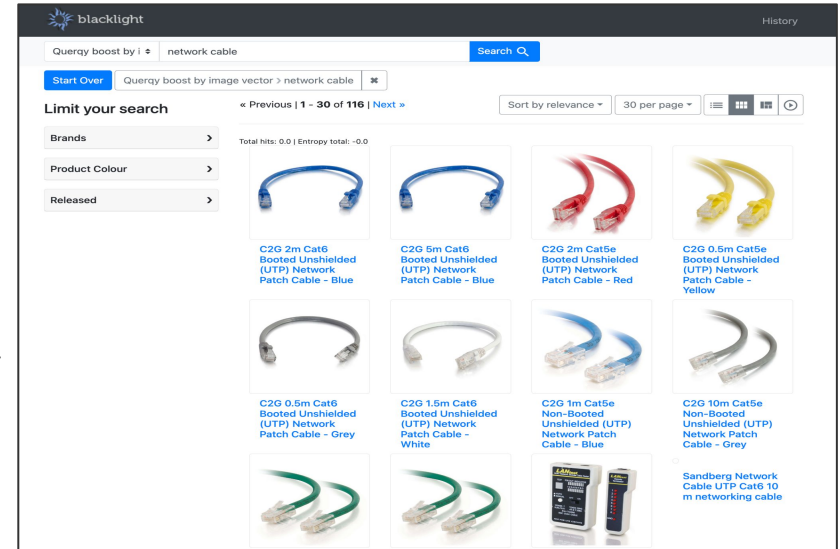
Sandberg Network Cable UTP Cat6 2 m networking cable

Sandberg Network Cable UTP Cat5e 10 m networking cable

Sandberg Network Cable UTP Cat5e 2 m networking cable

C2G LANtest

Results with boost by image vector ->



blacklight

Query boost by i network cable Search

Start Over Query boost by image vector network cable

Limit your search « Previous | 1 - 30 of 116 | Next » Sort by relevance 30 per page

Brands

Product Colour

Released

Total hits: 0.0 | Entropy total: -0.0

C2G 2m Cat6 Booted Unshielded (UTP) Network Patch Cable - Blue

C2G 5m Cat6 Booted Unshielded (UTP) Network Patch Cable - Blue

C2G 2m Cat5e Booted Unshielded (UTP) Network Patch Cable - Red

C2G 0.5m Cat5e Booted Unshielded (UTP) Network Patch Cable - Yellow

C2G 0.5m Cat6 Booted Unshielded (UTP) Network Patch Cable - Grey

C2G 1.5m Cat6 Booted Unshielded (UTP) Network Patch Cable - White

C2G 1m Cat5e Non-Booted Unshielded (UTP) Network Patch Cable - Blue

C2G 10m Cat5e Non-Booted Unshielded (UTP) Network Patch Cable - Grey

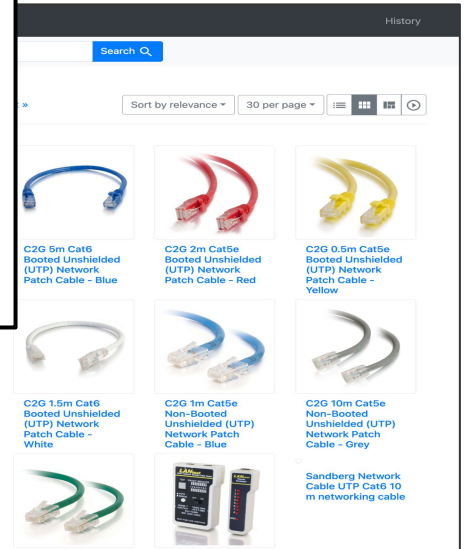
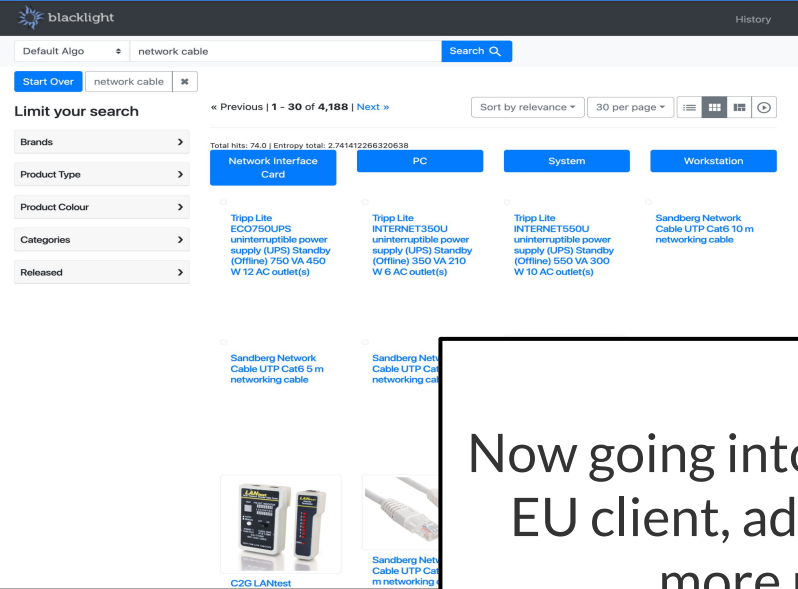
Sandberg Network Cable UTP Cat6 10 m networking cable

# Query : Network Cable

<- Results with default algo

Now going into production for a major EU client, addressing zero results - more news next year!

Results with boost by image vector ->

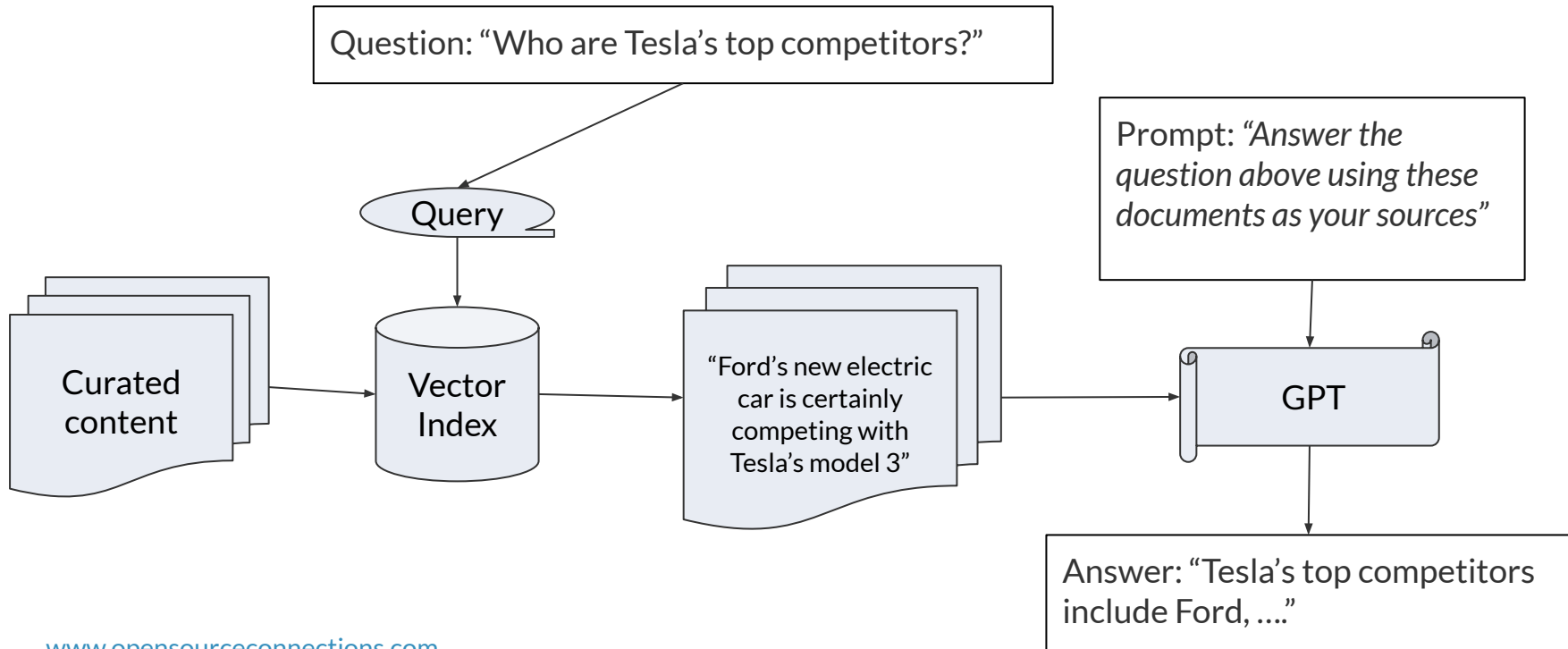


# Retrieval Augmented Generation (RAG)

---

1. User asks a question
2. Retrieve results from a vector database
3. Ask a LLM to answer the question *using the retrieved results*
4. Addresses hallucination & grounds the answer in truth you control - hopefully!

# Retrieval Augmented Generation (RAG)



Query: Who are Tesla's competitors in 2022?

Retrieve



Generate >>

Validate >>

20 Documents

**Tesla's U.S. electric vehicle market share will drop to 18% by 2026, BofA estimates** #1 (was: #1)  
(0.6362432)

+ Tesla's share of the U.S. electric vehicle market will drop to 18% by 2026, Bank of America analysts estimated in a note Tuesday. That would be a significant drop from the 32% share of the market BofA reported Tesla had in 2022 and even further down from its peak of 36% market share in 2019. That shift is expected as incumbent automakers such as Ford and General Motors are expected to grow their share of the U.S. EV market, according to BofA. Ford and GM, which hold 8% and 5% market share.

In 2022, Tesla's competitors in the electric vehicle market included incumbent automakers such as Ford and General Motors. Both Ford and General Motors held 8% and 5% market share respectively in 2022 [1]. Ford, in particular, has been directly competing with Tesla, having beaten Tesla to the pickup segment by beginning production of its F-150 Lightning, the electric version of its consistently popular trucks, in April 2022 [2]. Furthermore, General Motors has been competing with Tesla with its Bolt sedan [3]. In the Chinese market, Tesla is also competing with multiple local companies [4].

**Ford EVs will use Tesla charging tech in surprise partnership between rival automakers** #2 (was: #2)  
(0.6361465)

+ Tesla previously discussed opening its private network to other EVs. While House officials announced in February that Tesla committed itself to 1,000 of its charging stations by the end of 2024 to help Tesla EV drivers. Previously the company's chargers in the U.S. were not compatible and made it difficult to use with Tesla's EVs. In Tesla's first quarter shareholder deck, the company disclosed that it has roughly 3,000 Supercharger connectors available at 4,000 Supercharger Stations. The company.

**The \$25,000 electric vehicle is coming, with big implications for the auto market and car buyers** (0.6277465) #3 (was: #3)

+ While Tesla hopes the entry-level car will cement its role as a catalyst that can cause a reorgnization of the market, automakers have been reducing their footprints in the less profitable low end of the market, preferring to concentrate on larger vehicles with wider profit margins. A spokesman for Hyundai's U.S. operation said in an e-mail that the company has no plans to introduce a lower-end EV. He said Ford's have been announced other EVs will use the Equinox as its starting.

**Elon Musk's visit underscores China's importance to global EV market, analysts say** (0.6253409) #4 (was: #4)

+ It won't be easy for Tesla, however. He noted that China the most competitive market for EVs, with Tesla competing with multiple top automakers for supremacy. "Tesla is, unlike other players in the world, not the only top dog in this market," he added. When asked if Tesla's strategy of cutting prices is appropriate, Rees said Tesla is "fighting with an underdog" — Model 3 was launched three years ago and Model Y two years.

**The \$25,000 electric vehicle is coming, with big implications for the auto market and car buyers** (0.6204887) #5 (was: #5)

+ The market leader Tesla, the key to pulling costs down from the \$47,000 for price of the Model 3 standard range begins with new or redesigned batteries, costly greater scale and advances in battery technology. Rees and his said, they said battery costs have another 20 percent to fall after years of declines. In the No. 2 U.S. EV maker, Ford expects single scale economies to improve EVs, spending and margins by 20 percentage points by 2026, according to a presentation to analysts on...

**What Tesla charging partnerships with Ford and GM mean for the EV industry** (0.61279804) #6 (was: #6)

- Chunk content according to some algorithm (here the content was already provided in paragraphs/sentences, and we just used the paragraphs).
- Create embeddings for the chunks (using MiniLM with the sentence-transformers library).
- Index the chunks in OpenSearch, here using the Lucene HNSW backend (OpenSearch also supports nmslib and FAISS backends).
- On search, embed the query using MiniLM and retrieve the top 20 chunks by vector similarity.
- "Generate" button generates a prompt with the query and the top 5 chunks and sends it to OpenAI (or a local GPT model, e.g. Mistral-7b).
- Have some buttons that allow you to "rate" chunks and shuffle them around and move the 5 chunks threshold (marked with a red line) up or down to see how that changes GPT output.
- "Validate" button which sends another prompt to GPT and performs some sentence-by-sentence evaluation of each chunk to see which chunk contains the most sentences that OpenAI considers helpful in answering the question.

Query: Who are Tesla's competitors in 2022?

Retrieve



Generate &gt;&gt;

Validate &gt;&gt;

20 Documents

 Tesla's U.S. electric vehicle market share will drop to 18% by 2026, *BofA* estimates #1 (was: #1)  
 (0.6362432)

 + Tesla's share of the U.S. electric vehicle market will drop to 18% by 2026, Bank of America analysts estimated in a note Tuesday. That would be a significant drop from the 23% share of the market *BofA* reported Tesla had in 2022 and even further down from its peak of 26% market share in 2019. That shift is expected as incumbent automakers such as Ford and General Motors are expected to grow their share of the U.S. EV market, according to *BofA*. Ford and GM, which hold 8% and 5% market share.

In 2022, Tesla's competitors in the electric vehicle market included incumbent automakers such as Ford and General Motors. Both Ford and General Motors held 8% and 5% market share respectively in 2022 [1]. Ford, in particular, has been directly competing with Tesla, having beaten Tesla to the pickup segment by beginning production of its F-150 Lightning, the electric version of its consistently popular trucks, in April 2022 [2]. Furthermore, General Motors has been competing with Tesla with its Bolt sedan [3]. In the Chinese market, Tesla is also competing with multiple local companies [4].

 Ford EVs will use Tesla charging tech in surprise partnership between rival automakers #2 (was: #2)  
 (0.6361465)

+ Tesla previously discussed opening its private network to other EVs. White House officials announced in February that Tesla committed itself to 1,000 of its charging stations by the end of 2022 to non-Tesla EV drivers. Previously the company's chargers in the U.S. were limited to and made to be compatible with Tesla's EVs. In Tesla's first quarter shareholder deck, the company disclosed that it has roughly 3,000 Supercharger connectors worldwide at 4,347 Supercharger Stations. The company.

The \$25,000 electric vehicle is coming, with big implications for the auto market and car buyers (0.6277465) #3 (was: #3)

+ While Tesla hopes the entry-level car will cement its role as a catalyst that can serve as segments of the market, automakers have been reducing their footprints in the less profitable low end of the market, preferring to concentrate on larger vehicles with wider profit margins. A spokesman for Hyundai's U.S. operation said in an e-mail that the company has no plans to introduce a lower-end EV. He said Ford's have been announced other EVs will use the Equinox as its starting.

Elon Musk's visit underscores China's importance to global EV market, analysts say (0.6253409) #4 (was: #4)

+ It won't be easy for Tesla, however. He noted that China the most competitive market for EVs, with Tesla competing with multiple big automakers for supremacy. "Tesla is, unlike other players in the world, not the only top dog in this market," he added. When asked if Tesla's strategy of cutting prices is appropriate, Rees said Tesla is "fighting with an underdog" — Model 3 was launched three years ago at \$37,000.

The \$25,000 electric vehicle is coming, with big implications for the auto market and car buyers (0.6204887) #5 (was: #5)

+ The market leader Tesla, the key to pulling costs down from the \$47,000 list price of the Model 3 standard range begins with new or redesigned features, costly greater scale and advances in battery technology. Rees said that cost. Rees said battery costs have fallen 50 percent in last three years of decline. At the No. 2 U.S. EV maker, Ford expects single scale economies to improve EVs, operating profit margins by 20 percentage points by 2025, according to a presentation to analysts on.

What Tesla charging partnerships with Ford and GM mean for the EV industry (0.61279804) #6 (was: #6)

Prompt:

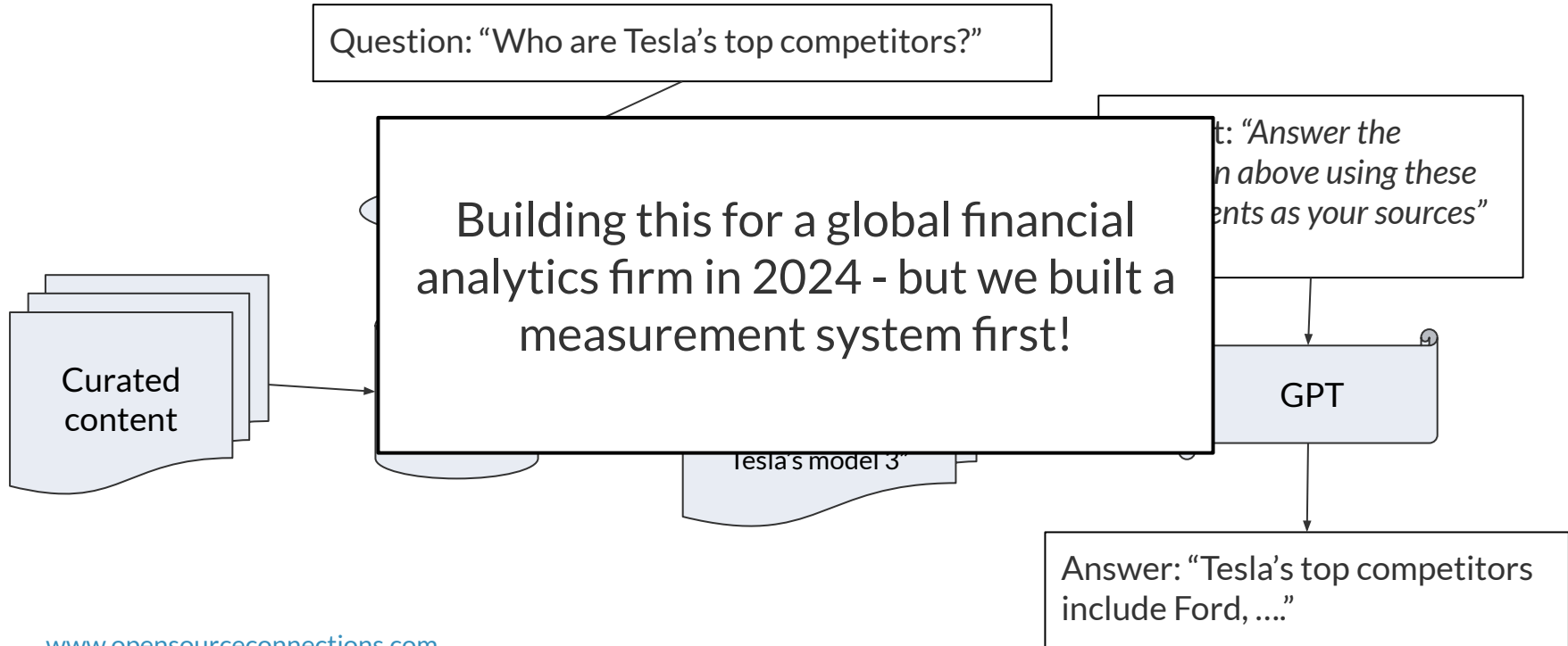
"You are an expert. You answer questions truthfully based on provided documents. You are talkative and provide lots of specific details from the context.

"For each document check whether it is related to the question. Only use documents that are related to the question to answer it. Ignore documents that are not related to the question. If the answer exists in several documents, summarize them. Only answer based on the documents provided. **Don't make things up.**"

" Always use references in the form [NUMBER OF DOCUMENT] when using information from a document. e.g. [3], for Document[3]. The reference must only refer to the number that comes in square brackets before the passage. Otherwise, do not use brackets in your answer and reference ONLY the number of the passage without mentioning the word passage."

" If the documents can't answer the question or you are unsure say: 'The answer can't be found in the text'."

# Retrieval Augmented Generation (RAG)



# The importance of measurement

---

Lord Kelvin:

*[A Lecture delivered at the Institution of Civil Engineers on May 3, 1883; being one of a series of Six Lectures on "The Practical Applications of Electricity."]*

I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it ; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind

# Measuring AI-powered search with Quepid

Quepid is OSC's free, open source platform for gathering human judgments, calculating metrics & tuning queries

[www.quepid.com](http://www.quepid.com)

We recently added AI search engine



```
Query | Tuning Knobs | Settings | History | Annotations

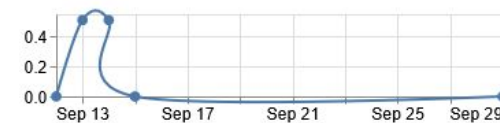
Query Sandbox:

1  {
2  "query": [
3  {
4    "query": "#$query#",
5    "start": 0,
6    "numResults": 10,
7    "corpusKey": [
8    {
9      "customerId": [REDACTED],
10     "corpusId": 23,
11     "lexicalInterpolationConfig": {
12       "lambda": 0.025
13     },
14     "dim": []
15   }
16 ]
17 }
18 ]
19 }
```

## Case Vectara Test

0.0 nDCG@10

12 Sep - 29 Sep



Search Engine

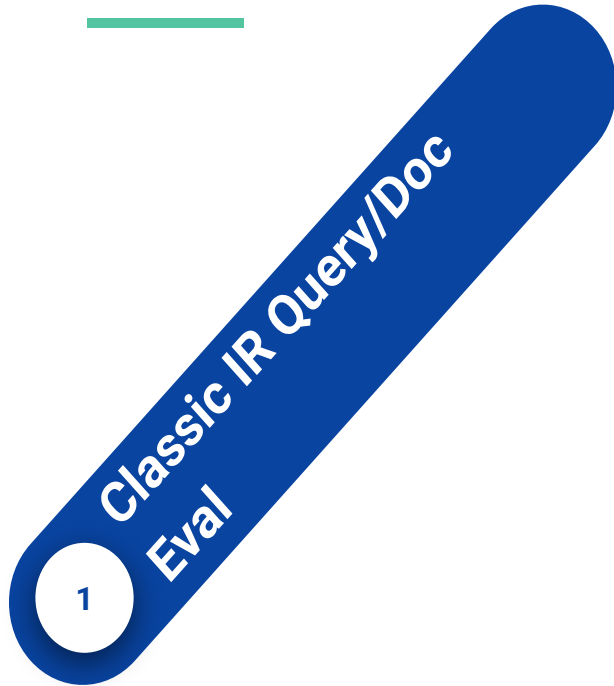
Change Search Engine:

- Solr
- Elasticsearch
- OpenSearch
- Vectara

<https://opensourceconnections.com/blog/2023/10/11/learning-to-measure-ai-powered-search-with-vectara/>



# We MUST Measure at Multiple Points.....



Does our retrieval system return vaguely decent candidates?

NCDG@  
ERR@  
P@

Implicit and Explicit evaluation....

# We MUST Measure at Multiple Points.....

---

Which LLM works best for us?

How do we store our vectors?

What chunking strategy do we need to adopt?

What is our tolerance for hallucinations that we are seeing?

2

Understand the LLM Inputs  
(chunks, data quality)

# We MUST Measure at Multiple Points.....

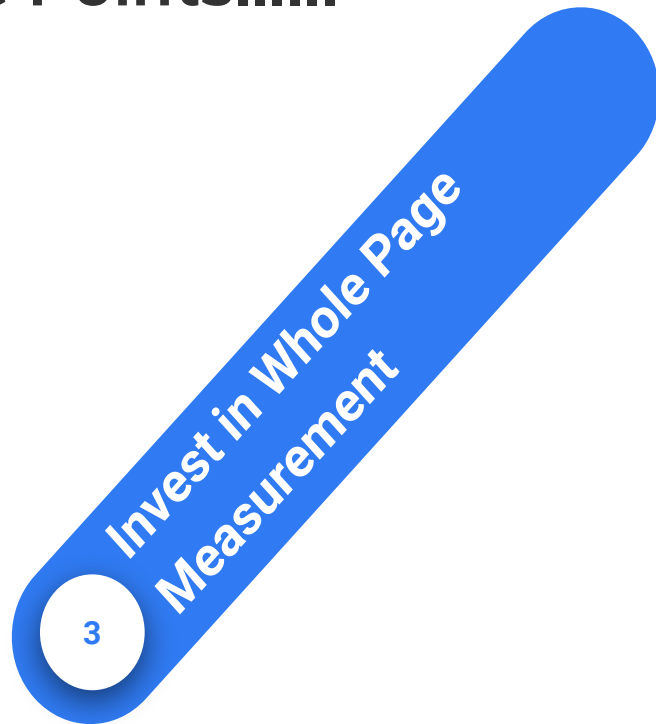
---

How useful are the summarizations that are generated?

Did the extracted content from documents make sense?

Are we sourcing a good diversity of documents to provide the LLM?

Do we provide relevant “next steps” for the users to go deeper when appropriate?



# We MUST Measure at Multiple Points.....

---

1

Classic IR Query/Doc Eval

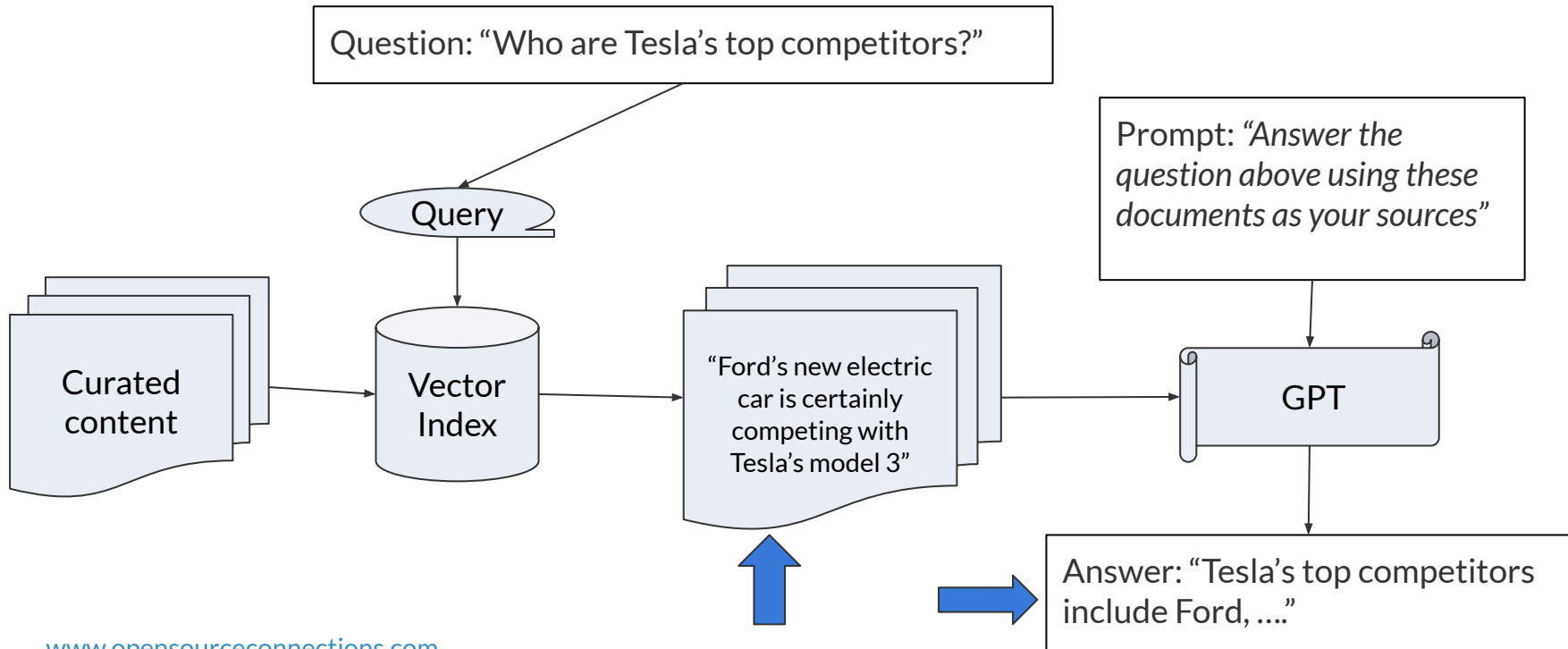
2

Understand the LLM Inputs  
(chunks, data quality)

3

Invest in Whole Page  
Measurement

# Retrieval Augmented Generation



# Retrieval Augmented Generation

## ragas score

generation

### faithfulness

how factually accurate is the generated answer

### answer relevancy

how relevant is the generated answer to the question

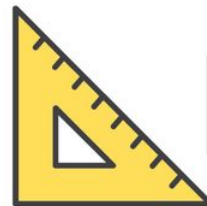
retrieval

### context precision

the signal to noise ratio of retrieved context

### context recall

can it retrieve all the relevant information required to answer the question



# ragas

*Evaluation framework for your Retrieval Augmented Generation (RAG) pipelines*

release [v0.0.20](#) Made with [Python](#) license [Apache-2.0](#) [CC](#) [Open in Colab](#) [explodinggradients](#) 473 members

<https://github.com/explodinggradients/ragas>

# Takeaways

---

# Takeaways

---

- Search is exciting again!
  - Perhaps call it AI if you want funding...
  - Lots of new people coming to search

# Takeaways

---

- Search is exciting again!
- AI techniques help solve some of search's oldest challenges
  - There is real promise here if we stay calm and pragmatic

# Takeaways

---

- Search is exciting again!
- AI techniques help solve some of search's oldest challenges
- You need to keep up
  - Things are changing very quickly
  - Don't believe all you read

# Takeaways

---

- Search is exciting again!
- AI techniques help solve some of search's oldest challenges
- You need to keep up
- Some things haven't changed
  - Search doesn't operate in a vacuum
  - What user problem are you trying to solve?

# Takeaways

---

- Search is exciting again!
- AI techniques help solve some of search's oldest challenges
- You need to keep up
- Some things haven't changed
- Just because you *have* AI doesn't mean you should use it
  - Significant risks exist
  - It can be slow & costly to implement

# Takeaways

---

- Search is exciting again!
- AI techniques help solve some of search's oldest challenges
- You need to keep up
- Some things haven't changed
- Just because you *have* AI doesn't mean you should use it
- Pragmatic, cheap and open source solutions can work

# Takeaways

---

- Search is exciting again!
- AI techniques help solve some of search's oldest challenges
- You need to keep up
- Some things haven't changed
- Just because you *have* AI doesn't mean you should use it
- Pragmatic, cheap and open source solutions can work
- Remember to measure!

# Thank you.

- 
- Email me at [chull@opensourceconnections](mailto:chull@opensourceconnections) if you need help building Search & AI
  - Join Relevance Slack at [www.opensourceconnections.com/slack](http://www.opensourceconnections.com/slack)
  - Check out the Haystack conference at [www.haystackconf.com](http://www.haystackconf.com)