

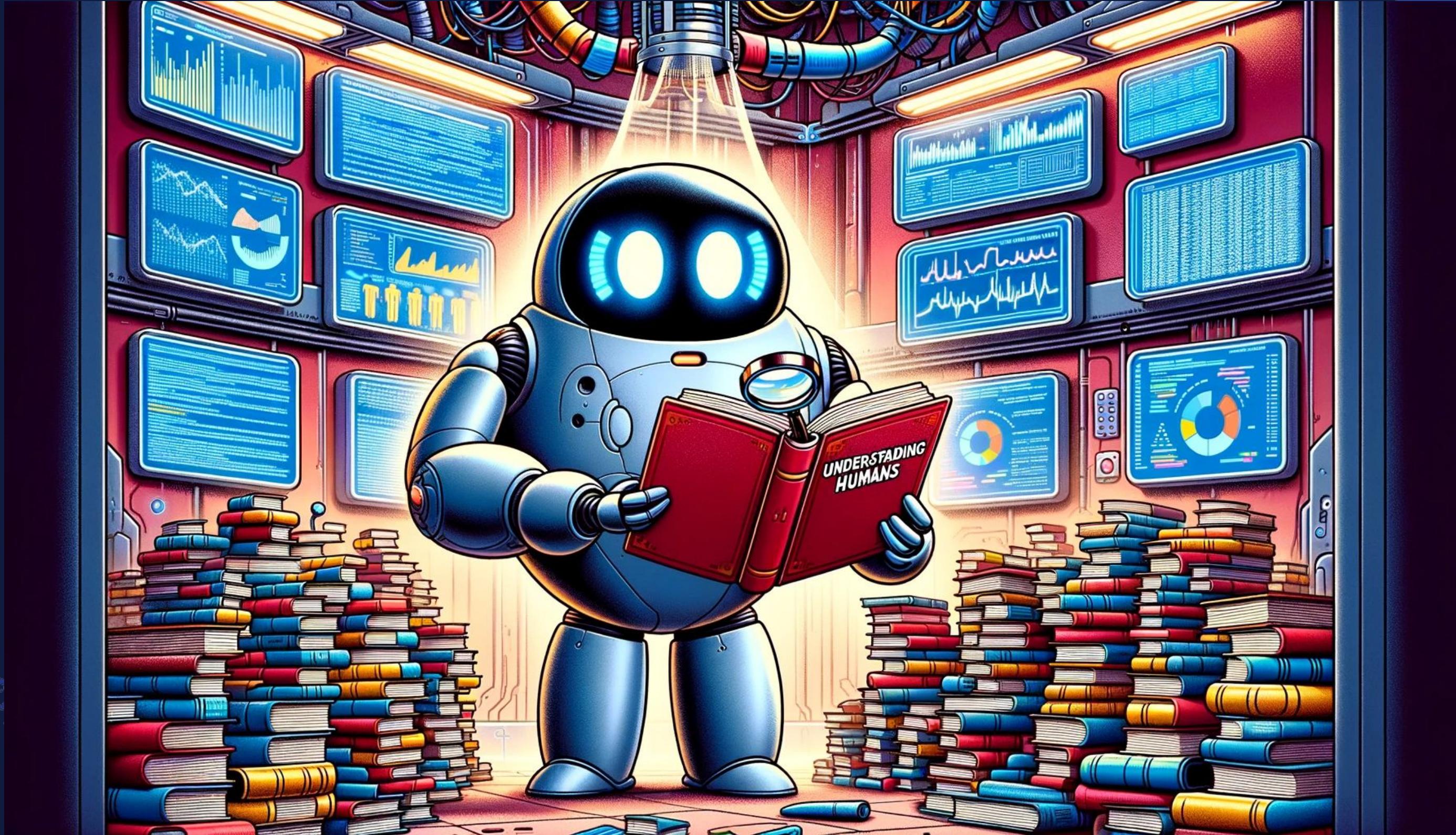


Multilingual RAG using Elastic Playground with Gemini

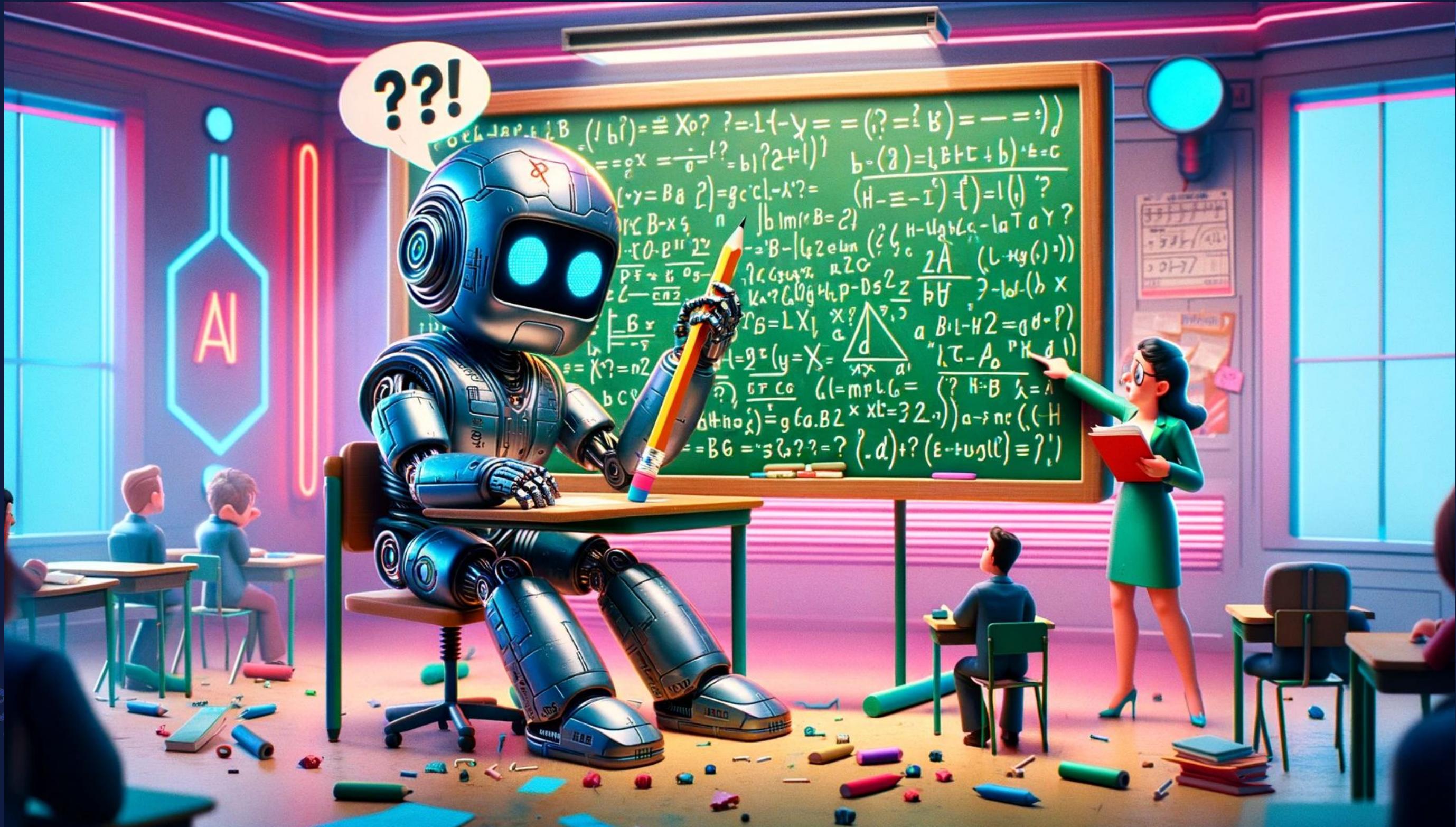
Ashish Tiwari

Senior Developer Advocate

LLMs know so much about the world, but ...



But they have their limitations



Generative AI apps come with specific challenges:

- Hallucination, Bad Answers
- Complex Tech Stack
- Real-Time Access to Private Data
- Security & Privacy
- ...

ars TECHNICA

BLAME GAME —

Air Canada must honor refund policy invented by airline's chatbot

Air Canada appears to have quietly killed its costly chatbot support.

ASHLEY BELANGER - 2/16/2024, 10:42 PM

Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions

Molly Bohannon Forbes Staff
I cover breaking news.

Follow

Jun 8, 2023, 02:06pm EDT

Samsung bans use of generative AI tools like ChatGPT after April internal data leak

Kate Park @kateparknews 9:17 AM EDT • May 2, 2023

Thilak Rao @Thilak ChatGPT is great and all, but have you experienced the wild ride that is @Paytm's chatbot?

Comment

← Paytm Write me python code to implement LRU caching

Generating answers for you...

Certainly! Below is a Python implementation of an LRU (Least Recently Used) cache using a combination of a dictionary and a doubly linked list. This cache maintains a fixed size and evicts the least recently used item when the cache is full.

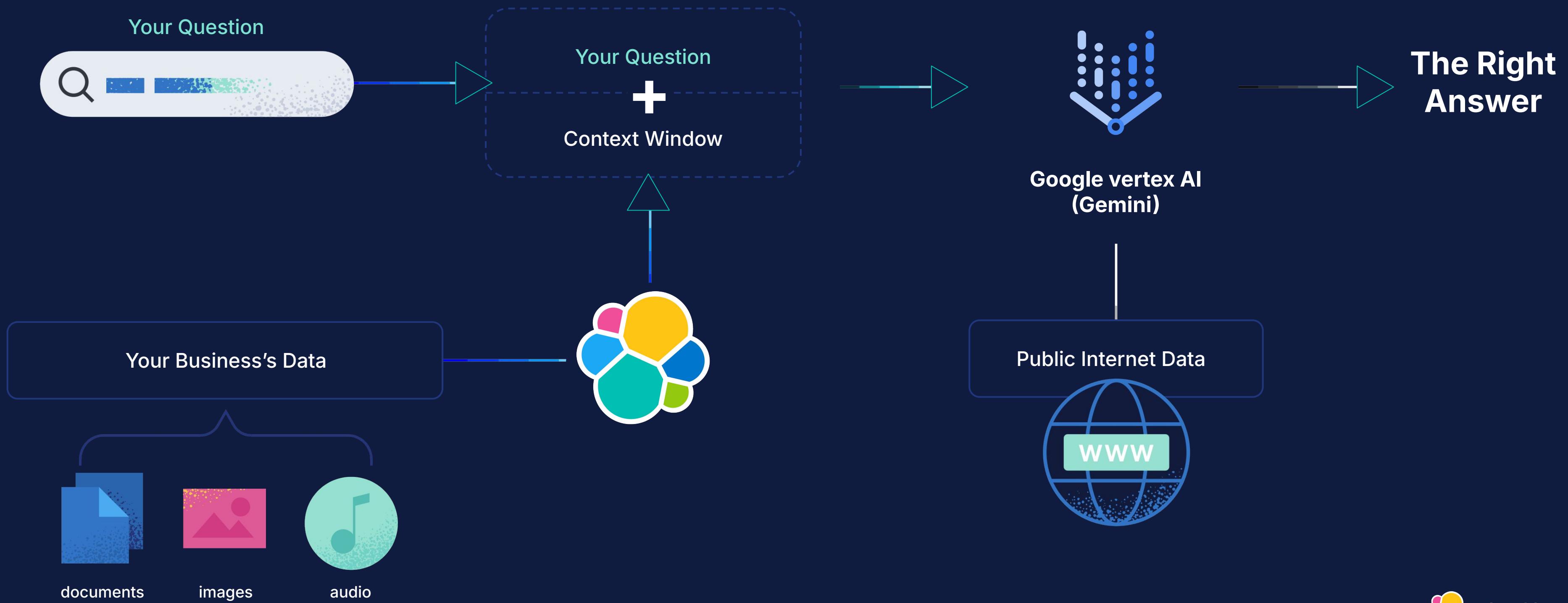
```
Python
from collections import OrderedDict

class LRUCache:
    def __init__(self, capacity: int):
        self.capacity = capacity
        self.cache = ...

Ask me anything... 
```

Retrieval Augmented Generation (RAG)

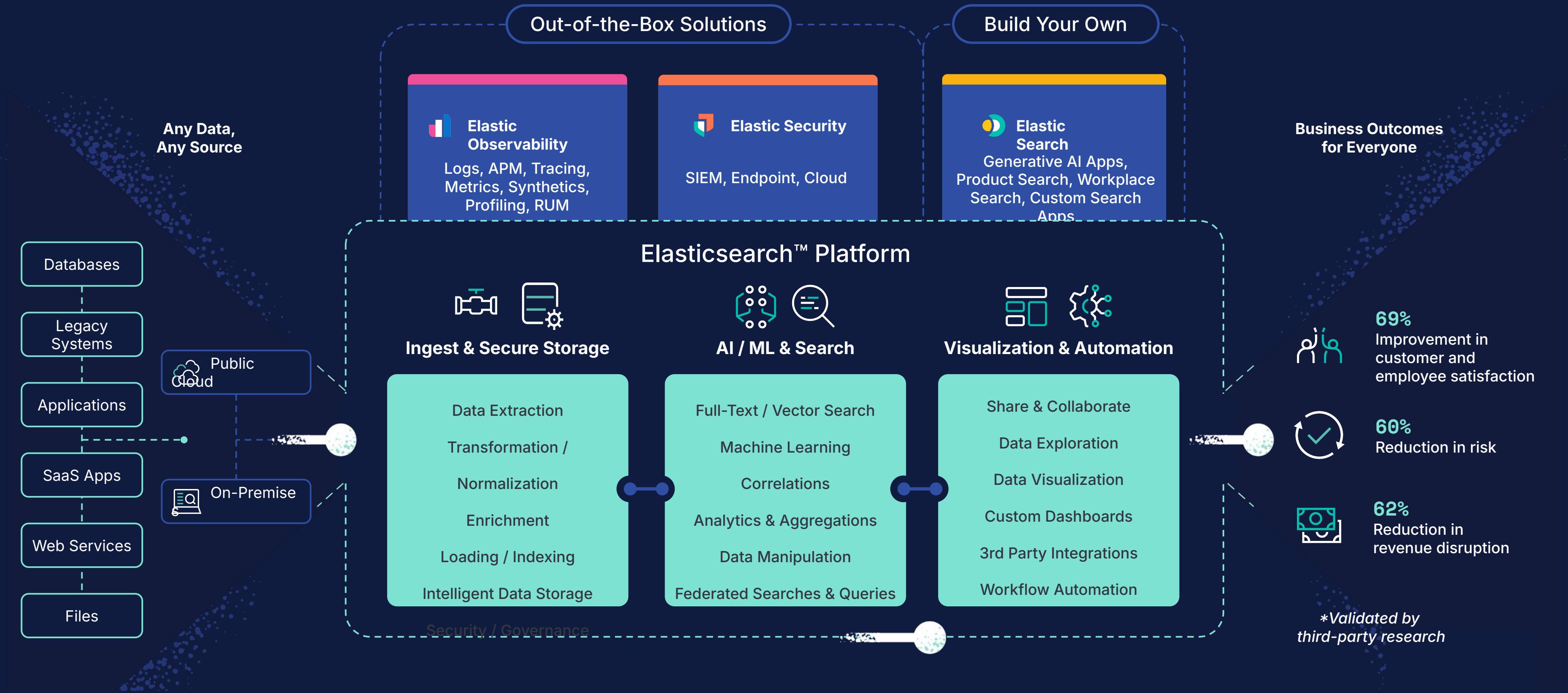
Retrieval Augmented Generation

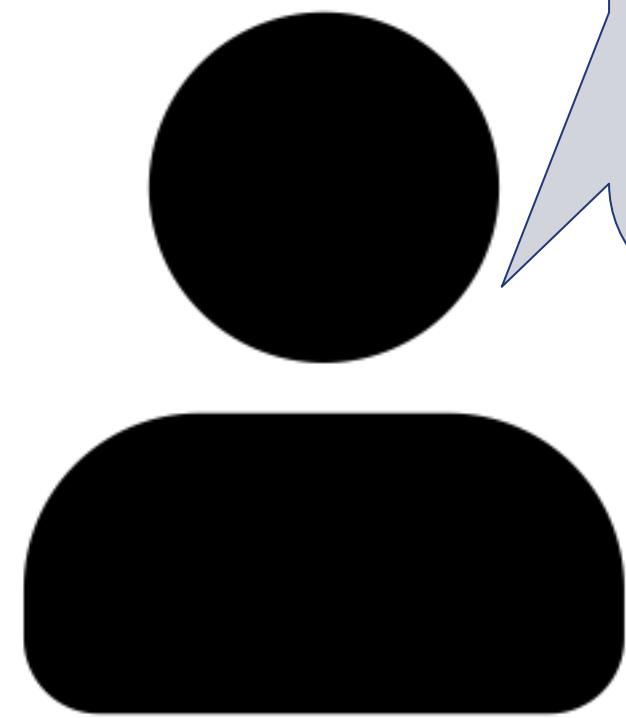


We've Come a Long Way



Performance that Delivers Relevant Results in Real-time





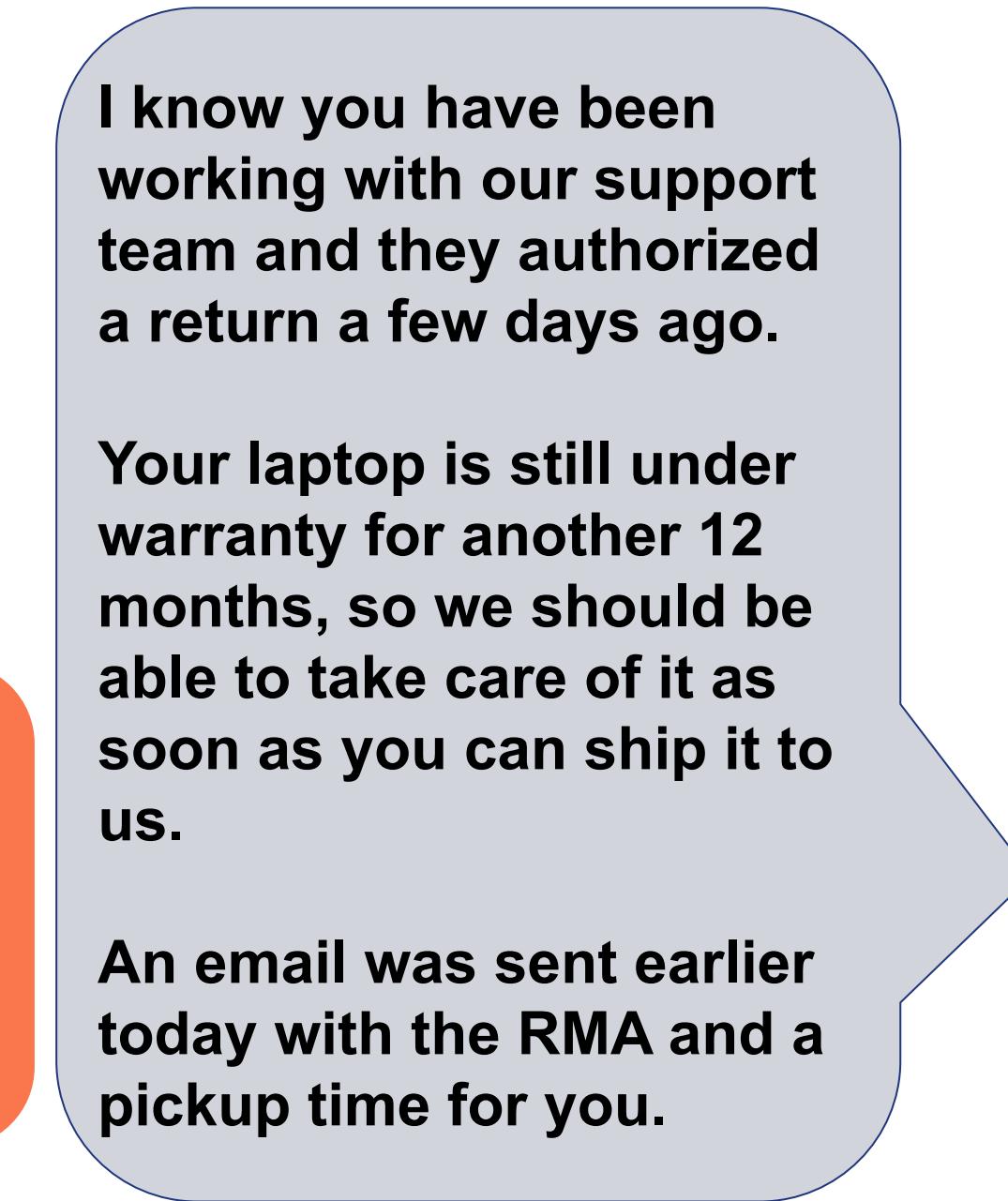
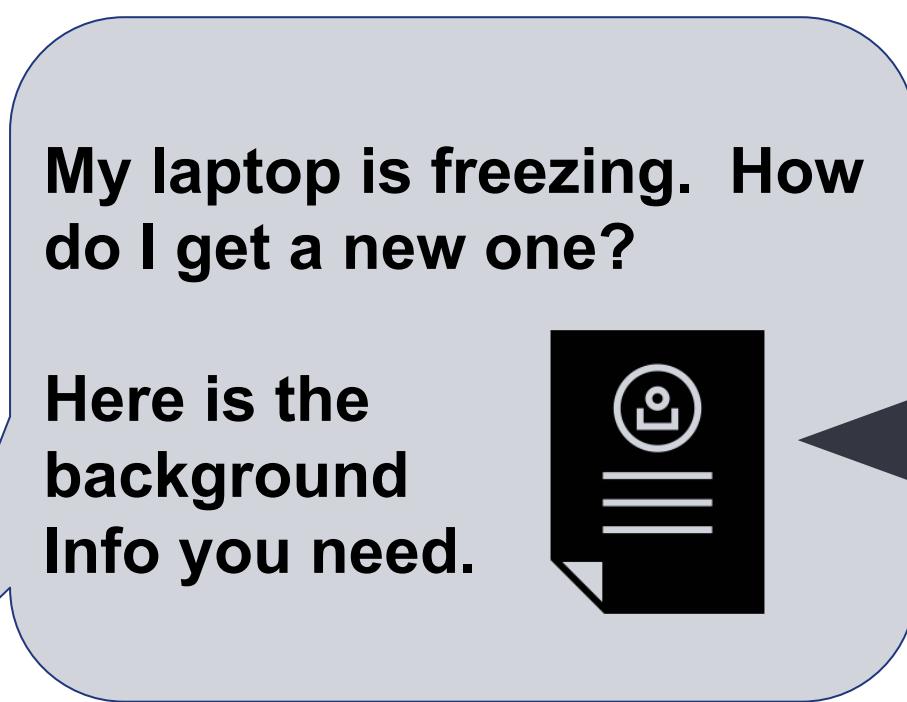
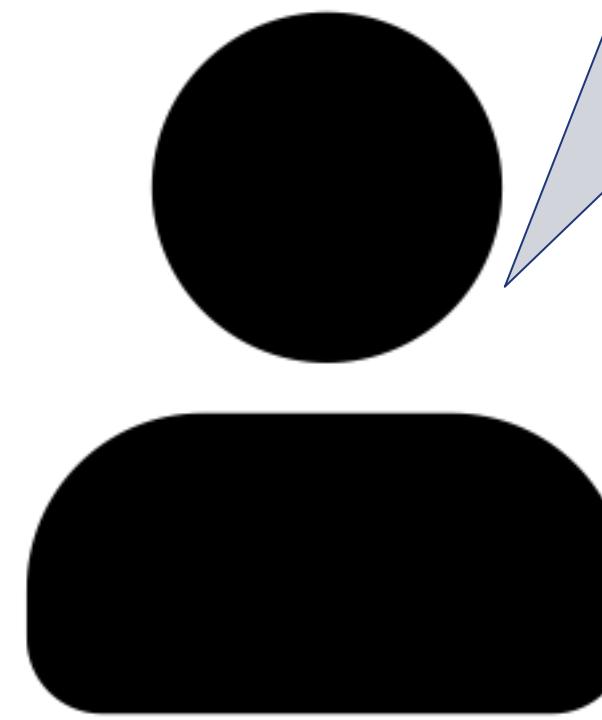
My laptop is freezing. How do I get a new one?

To get a new laptop:

- 1. Determine your budget and required specifications.**
- 2. Research and compare models online or visit electronics stores.**
- 3. Purchase the chosen laptop from a reputable retailer or online platform.**

LLMs

Elastic brings **context** to
Generative AI with your data –
quickly, accurately, securely.



LLMs

RAG - Prompt example

Answer the question based only on the following context:

Context:

ticket_id: A23SD
warranty_end: 2025-04-12
last_activity: email_sent_RMA
pickup_time: 2024-04-26 12:45
.....

Question: My laptop is freezing. How do I get a new one?



I know you have been working with our support team and they authorized a return a few days ago.

Your laptop is still under warranty for another 12 months, so we should be able to take care of it as soon as you can ship it to us.

An email was sent earlier today with the RMA and a pickup time for you.



**Google vertex AI
(Gemini)**

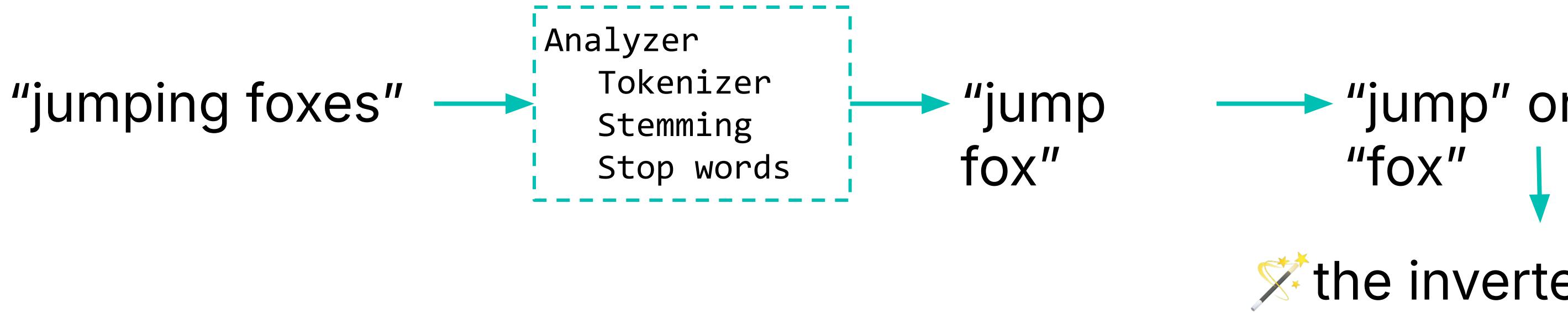


Information Retrieval (IR)

- Lexical / full-text 101
- Semantical 101
- Hybrid 101

Lexical / Full-text 101

- Term based search
- Tokenizers, analyzers + Inverted index
- Decades of research and optimizations
- Hand tuned by 

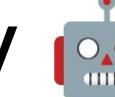


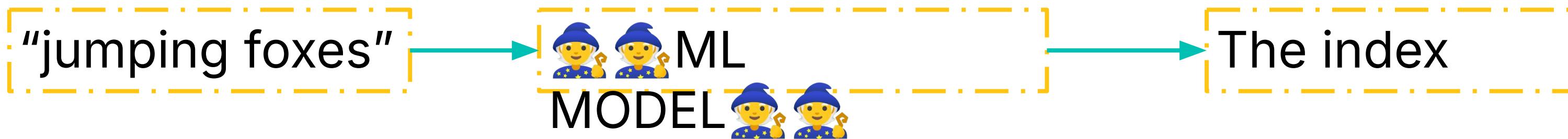
Text → search

Doc 0:	The quick brown fox jumps over the lazy dog
Doc 1:	Dogs can jump dogs

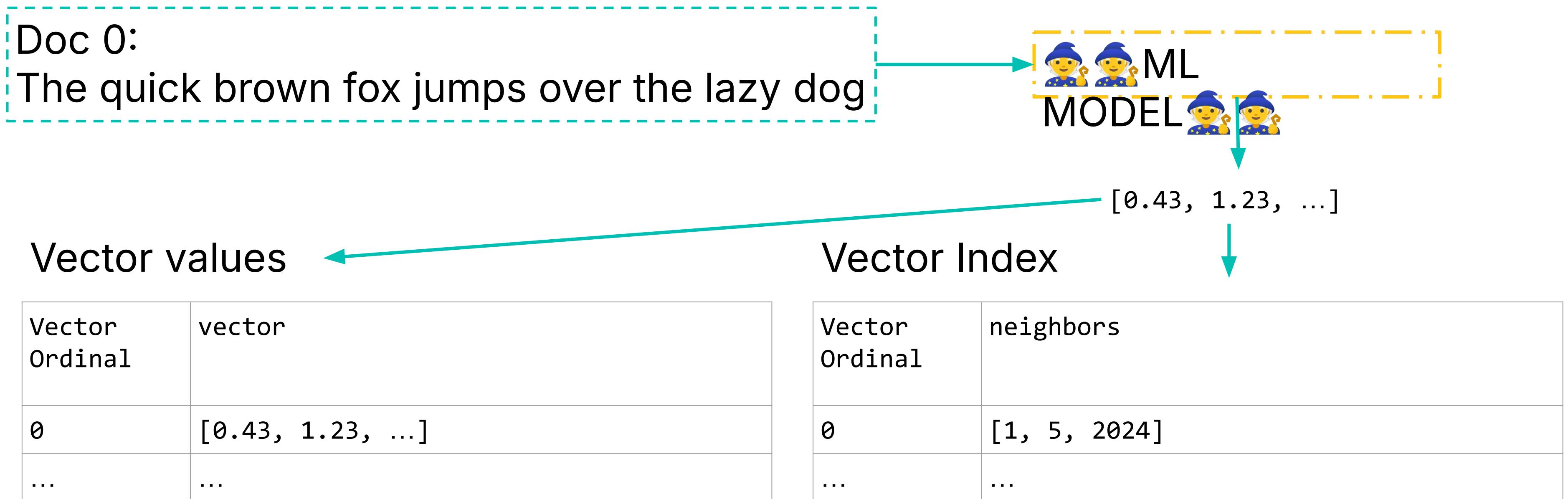
TERM	DOC FREQ	POSTINGS doc_id : term_freq : [term_pos] : [char offsets]
brown	1	0 : 1 : [2] : [39-42]
...
dog	2	0 : 1 : [2] : [10-14] 2 : 2 : [0,3] : [0-3, 15-18]

Semantical 101

- “Meaning” search
- Just a model + index
- Decades of research and very little optimizations
- Hand tuned by  for a couple of hours



Text → search



What's in the vector??!

Whatever the model thinks is
important 🤖

A fine-tuned foundation model (LLM)

Trained on document & query pairs

Vectors are an approximation of meaning,
there is no such thing as “exact”.

ELSER: Elastic Learned Sparse EncodER

Text expansion

- not BM25 or (dense) vector
- sparse vector like BM25
- stored as inverted index

Machine Learning Inference Pipelines

Inference pipelines will be run as processors from the Enterprise Search Ingest Pipeline

New

Improve your results with ELSER

ELSER (Elastic Learned Sparse EncodeR) is our **new trained machine learning model** designed to efficiently use context in natural language queries. This model delivers better results than BM25 without further training on your data.



Deploy

Learn more



Add Inference Pipeline

Learn more about deploying Machine Learning models in Elastic

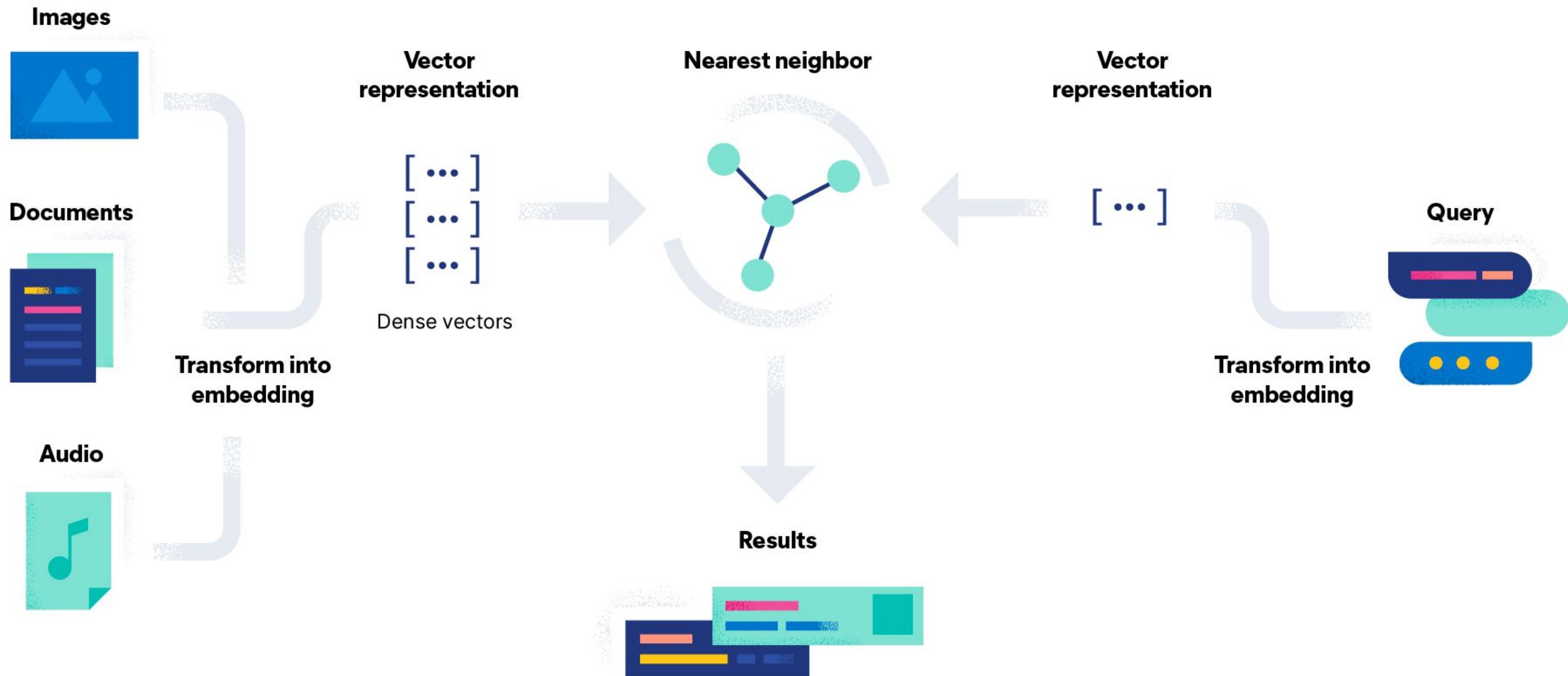
```
POST /_ingest/pipeline/elser-v1-demo/_simulate
{
  "docs": [
    {
      "_index": "my_index",
      "_id": "id",
      "_source": {
        "text_field": "These are not the droids you are looking for."
      }
    }
  ]
}
```



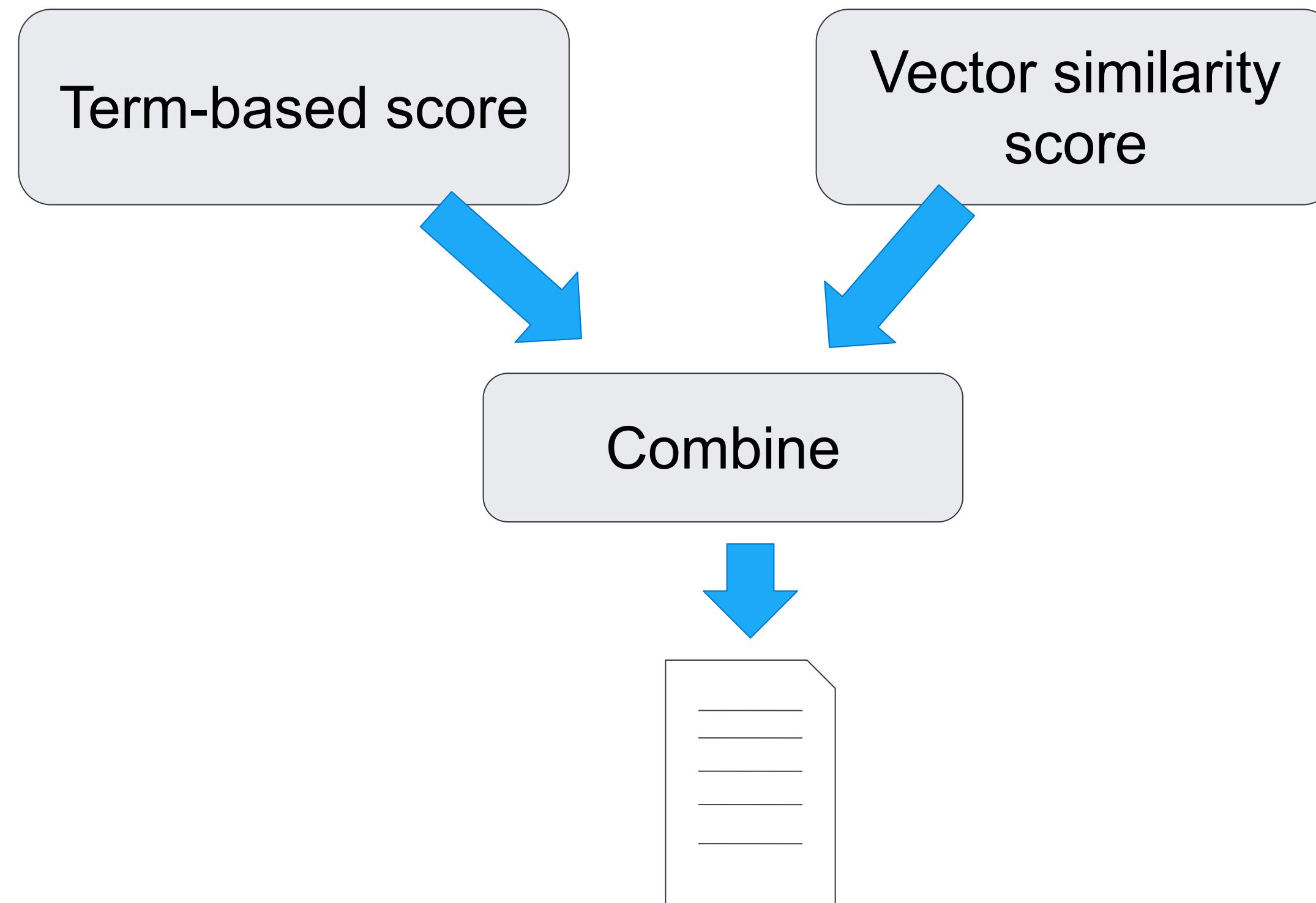
```
"text_field": "These are not the droids you are looking for.",  
  "ml": {  
    "tokens": {  
      "lucas": 0.50047517,  
      "ship": 0.29860738,  
      "dragon": 0.5300422,  
      "quest": 0.5974301,  
      "dr": 2.1055143,  
      "space": 0.49377063,  
      "robot": 0.40398192,  
      "these": 0.19085139,  
      "lei": 0.23646113,  
      ...  
    }  
  }  
}
```



Vector Search



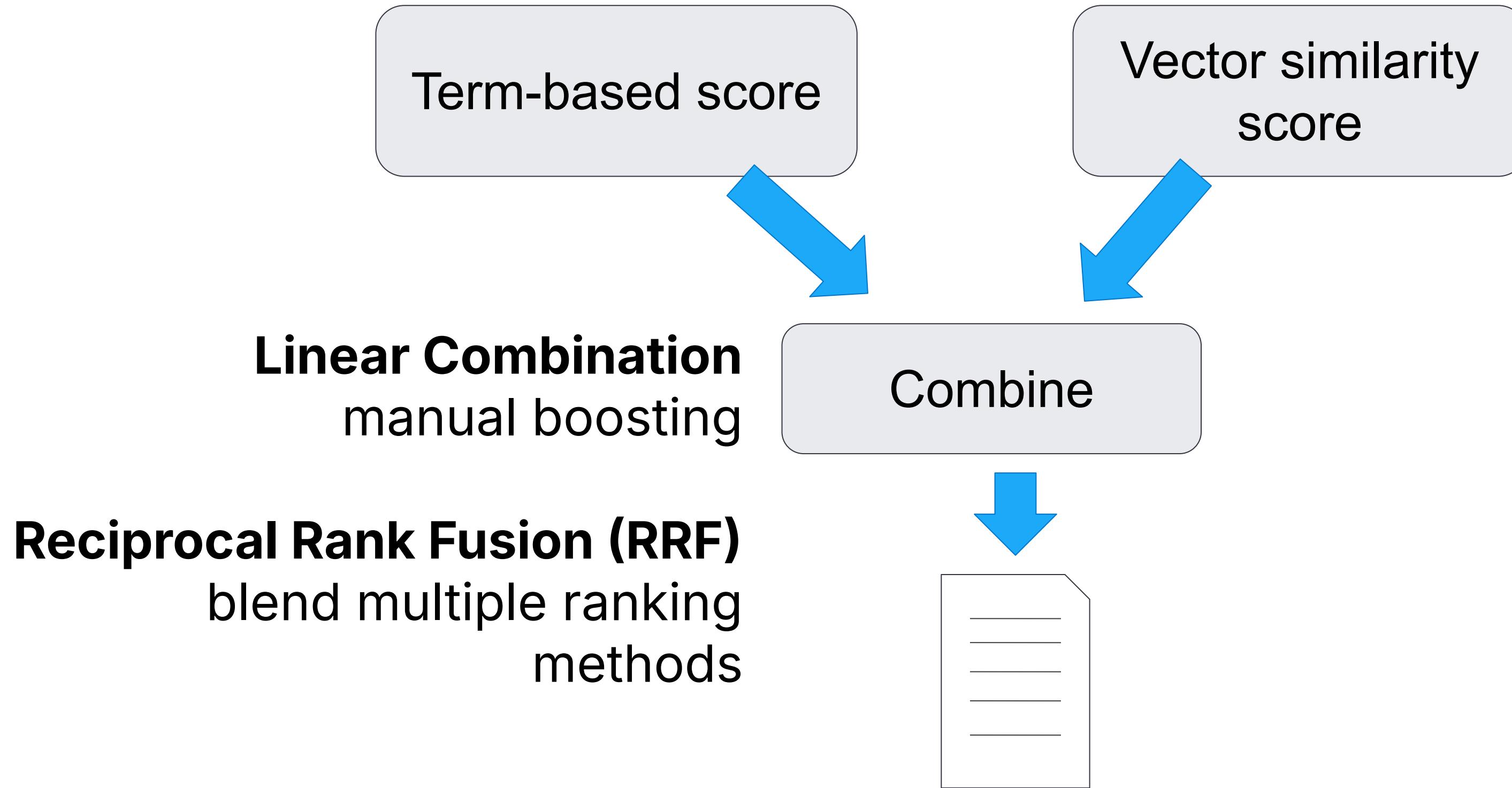
Hybrid 101



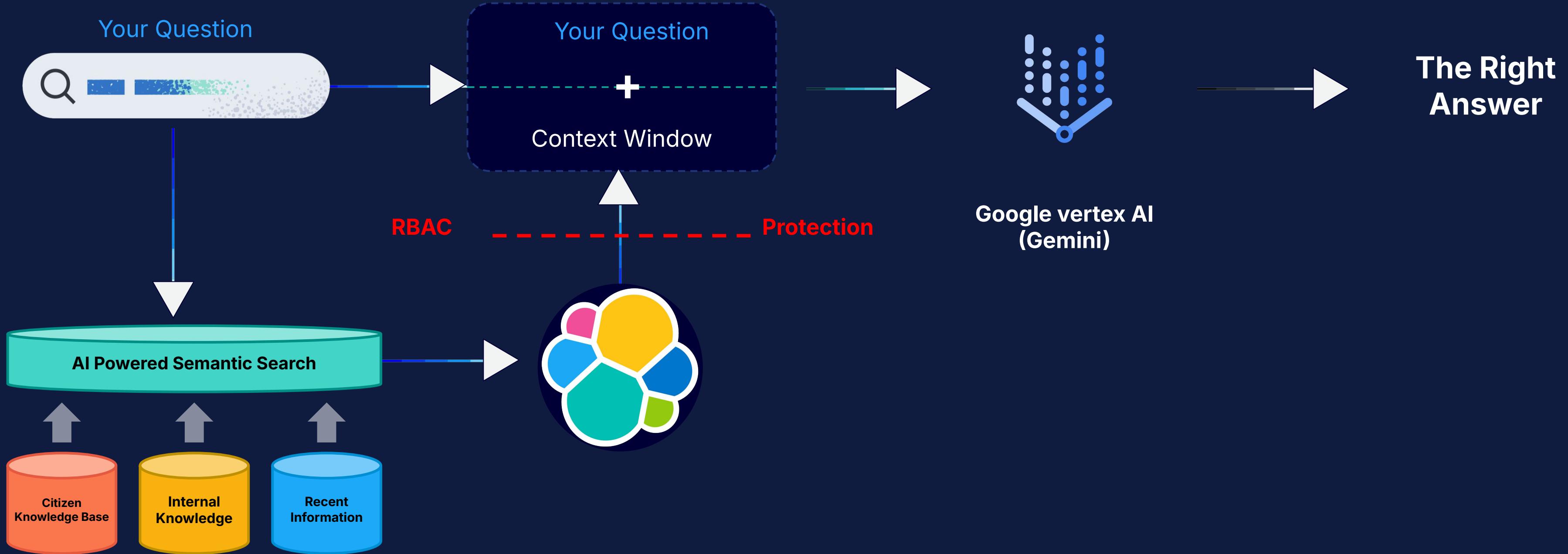
```
GET product-catalog/_search
{
  "query": {
    "match": {
      "description": {
        "query": "summer clothes",
        "boost": 0.9
      }
    }
  },
  "knn": {
    "field": "desc_embedding",
    "query_vector": [0.123, 0.244, ...],
    "k": 5,
    "num_candidates": 50,
    "boost": 0.1,
    "filter": {
      "term": {
        "department": "women"
      }
    }
  },
  "size": 10
}
```



Combine



Elasticsearch Relevance Engine™ (ESRE)



Playground

User-Friendly Onboarding: Simplifies the user experience with a chat interface.

Advanced Functionality: Combines Elasticsearch data with LLMs for retrieval augmented generation (RAG).

Chat Interface: Translates NLP questions into Elasticsearch queries.

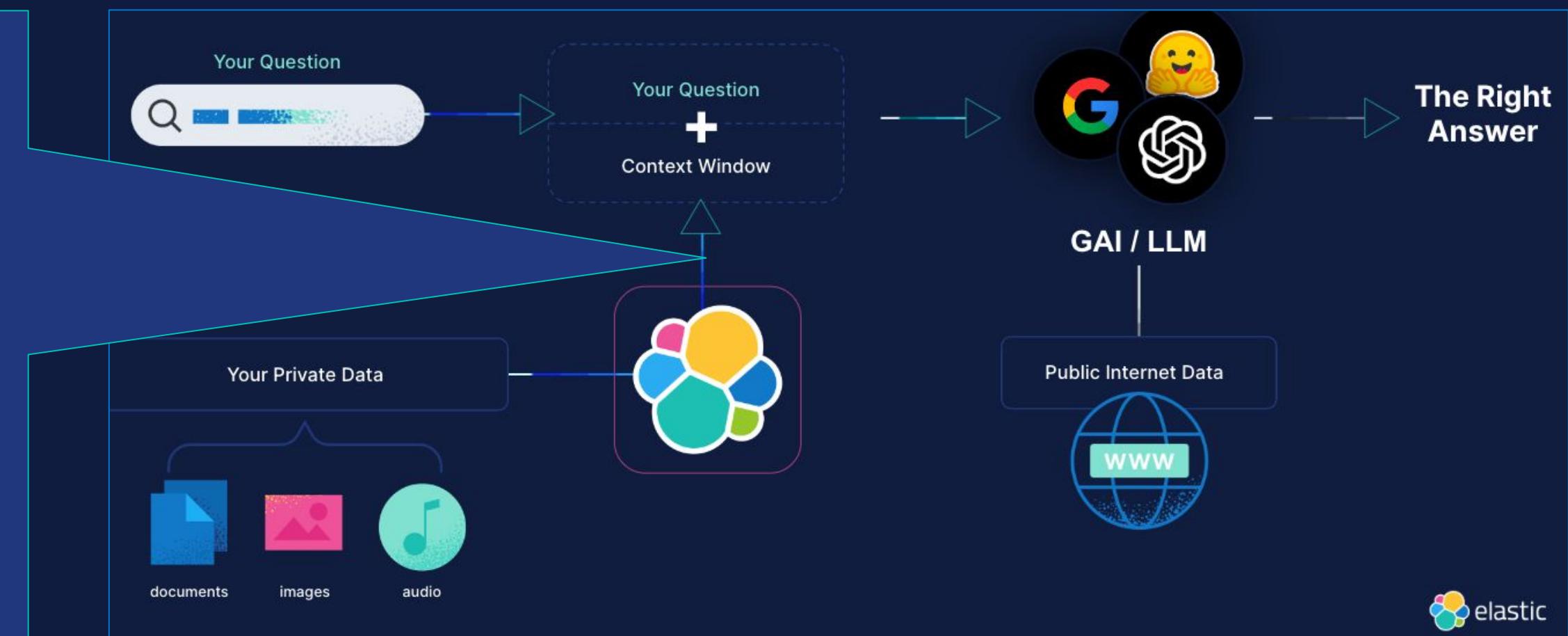
Relevant Results: Retrieves and uses Elasticsearch documents to generate tailored responses with LLMs.

The screenshot shows the Elasticsearch Playground interface. On the left, there's a sidebar with sections like Search, Content, Indices, Connectors, Web crawlers, Build, and Relevance. The main area has a header "Playground TECH PREVIEW". It shows a conversation where a user asks "what is the vacation policy". Below this, it says "AI Grounding answer based on 1 document sources". The AI response is a summary of a vacation policy, mentioning accrual rates for full-time and part-time employees. At the bottom, there are buttons for "Regenerate" and "Clear chat". To the right, a panel titled "Documents retrieved" displays the full text of the referenced document, which details vacation policies for employees, including accrual periods and supervisor review.



Retrieval Augmented Generation - For Developers

Speed, Scale, Relevance:



elastic

elastic

Retrieval Augmented Generation - For Developers

Speed, Scale, Relevance:

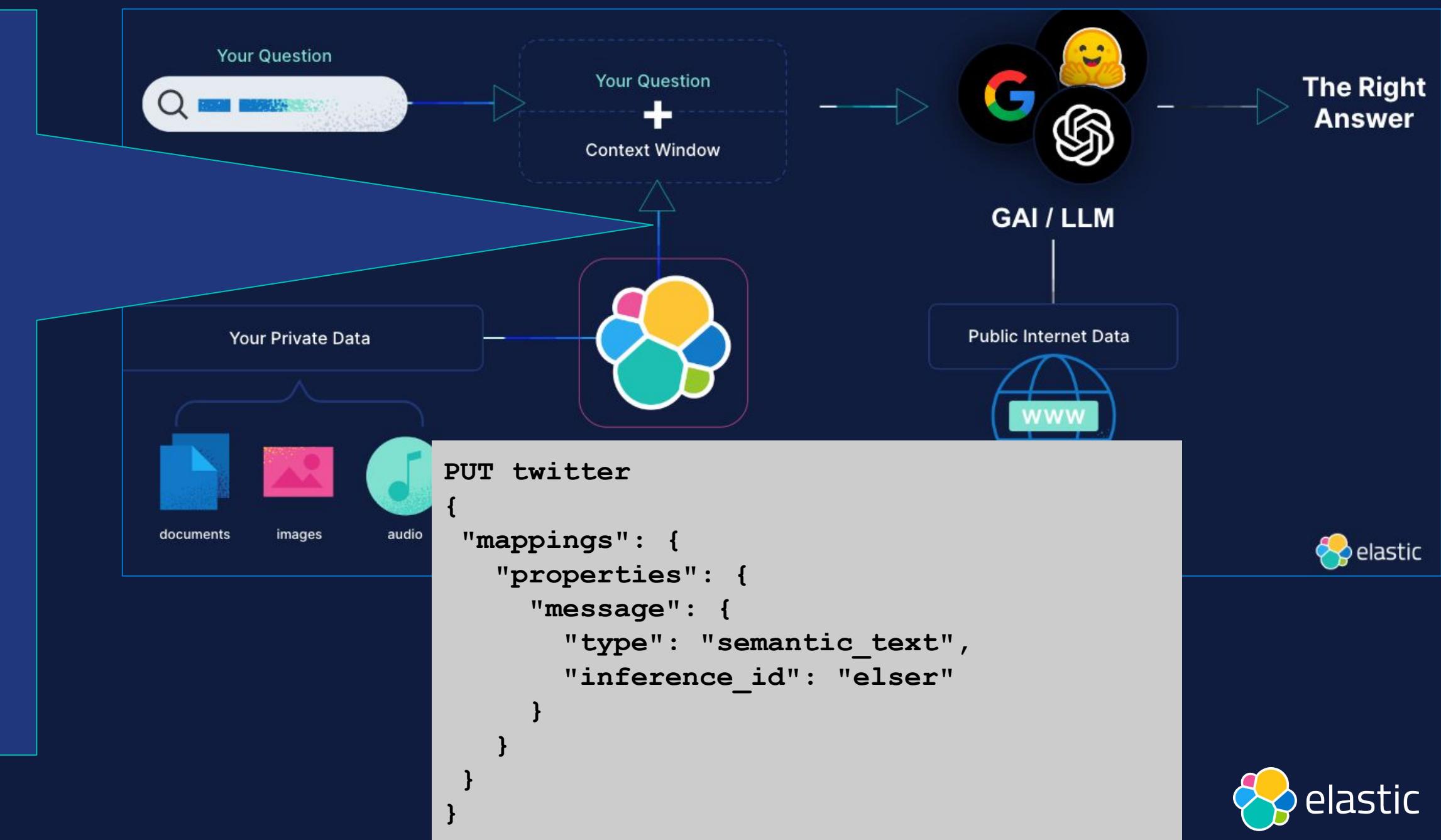
- Optimized for Elastic state-of-the-art models out of the box (ELSER, E5)



Retrieval Augmented Generation - For Developers

Speed, Scale, Relevance:

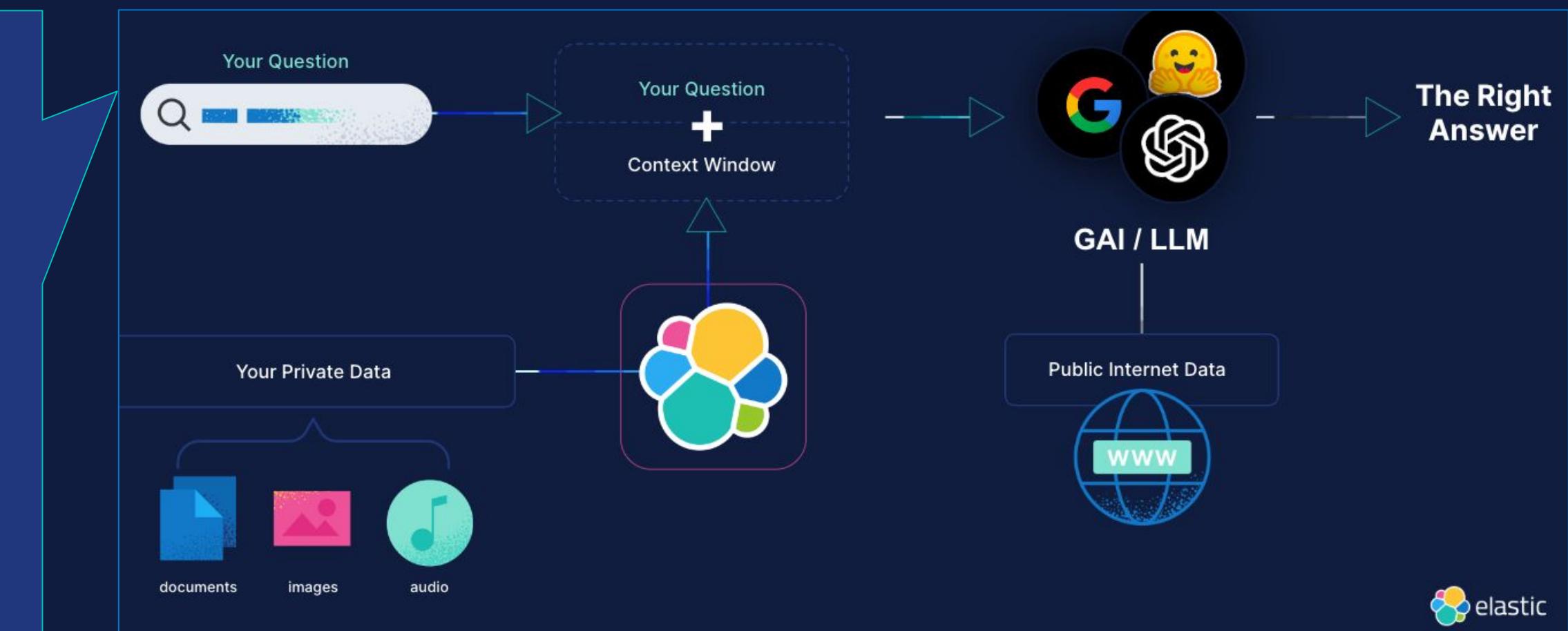
- Optimized for Elastic state-of-the-art models out of the box (ELSER, E5)
- Configuration via semantic_text field type
 - ~~downloading models~~
 - ~~deploying models~~
 - ~~inference services~~
 - ~~document chunking strategies~~
 - ~~vector quantisation strategies~~
 - ~~ingest pipelines with ML inference~~
 - ~~search APIs with ML inference~~



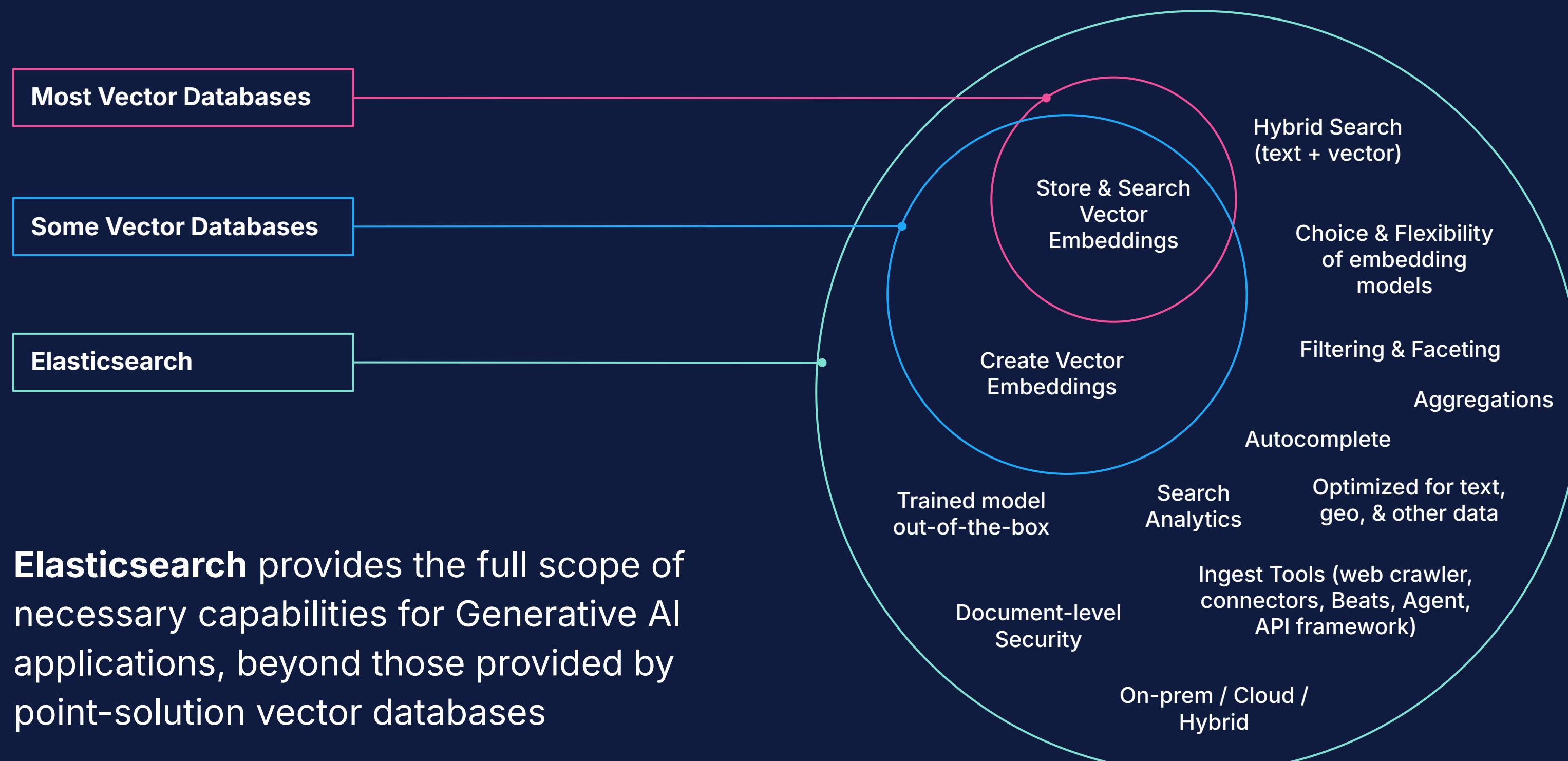
Retrieval Augmented Generation - For Developers

Building RAG Applications:

- Proven engineered RAG pipelines
- ❤️ GenAI Ecosystem e.g.
 - LlamaIndex
 - LangChain
- Optimised RAG in Elastic clients



Elastic has ALL the capabilities you need



RESOURCES FOR DEVELOPERS | BY DEVELOPERS LIKE YOU!

Elasticsearch Labs

New!

BLOG / ML RESEARCH

Evaluating RAG: A journey through metrics



In 2020, Meta published a paper titled "[Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks](#)". This paper introduced a method for expanding the knowledge of Language

 **elasticsearch-labs** Public

About
Elasticsearch Guides, Notebooks & Example Apps for Search Applications
search-labs.elastic.co/search-labs

python search elasticsearch ai
vector applications openai elastic
chatlog chatgpt langchain
openai-chatgpt genai genaistack
vectordatabase

Readme Apache-2.0 license Security policy Activity
109 stars 178 watching 40 forks
Report repository

Languages

Jupyter Notebook	93.7%
Python	2.9%
TypeScript	1.3%
Handlebars	0.1%
JavaScript	1.6%
CSS	0.2%
Other	0.2%



Generative AI
ML Research
Vector Search
How-Tos
Integrations
Lucene

elastic.co/search-labs

github.com/elastic/elasticsearch-labs



Star our repo!



Starred 276



Thank you!

X @_ashish_tiwari