Unlocking Relevance with Large Language Models in Search

Matt Riley

General Manager, Search

Search is **rapidly** evolving

HR helpdesk

What are key aspects of the company's 401k policy for an employee in my location? How do I enroll?

Customer Success

Are customers in Dallas buying products with the deepest discounts? Which promotions are doing well?

Ecommerce

What material list and tools do I need to build an irrigation system for 1 acre back yard in Detroit, MI?

Regulatory Compliance

Which recent transactions are not compliant with upcoming 2025 regulatory requirement X?

Corporate Finance

Summarize changes to revenue recognition and expense categorization for FY 2024

Predictive Maintenance

Based on sensor data and recent customer reviews, when do I schedule repair unit X?

elastic

Great search is **critical** for great Gen AI experiences

elastic

Why retrieval technology **matters**

Great user experience

Build engaging search, enriched by your private data and context

Control costs, hallucinations

Set context window size, use high relevance ranking to reduce irrelevant context

Enterprise toolkit for developers

Not just vector search, specialized filtering & faceting (geo, time), hybrid search & more

elastic

Retrieval augmented generation with Elasticsearch

"Does the company own my side hustle?"

1 Question as search query

Elasticsearch

Domain Specific, Private Content

2.A

original question

context window

2.B

Most relevant data to the query

3

4

The company's ownership of your side hustle depends on the circumstances under which it was created. According to the Intellectual Property Policy, if the intellectual property of your side hustle was created, conceived, or developed outside of company time and without the use of company resources, equipment, or facilities, it generally remains your property. However, if the intellectual property is directly related to your job responsibilities or if you have used company resources, equipment, or facilities in its creation, it may be considered the property of the company.

THE ANSWER WAS FORMED FROM INFORMATION FOUND IN THIS DOCUMENT

Intellectual Property Policy

elastic

History of **AI innovation** at Elastic

Support for supervised and unsupervised learning
2017

Native support for Vector Search
2019

Support for NLP and BYO Transformer models
2022

ELSER - a proprietary zero-shot retrieval model
2023

ESRE Elasticsearch Relevance Engine
2023

Support for forecasting in Observability
2018

Support for anomaly detection and AIOps
2019

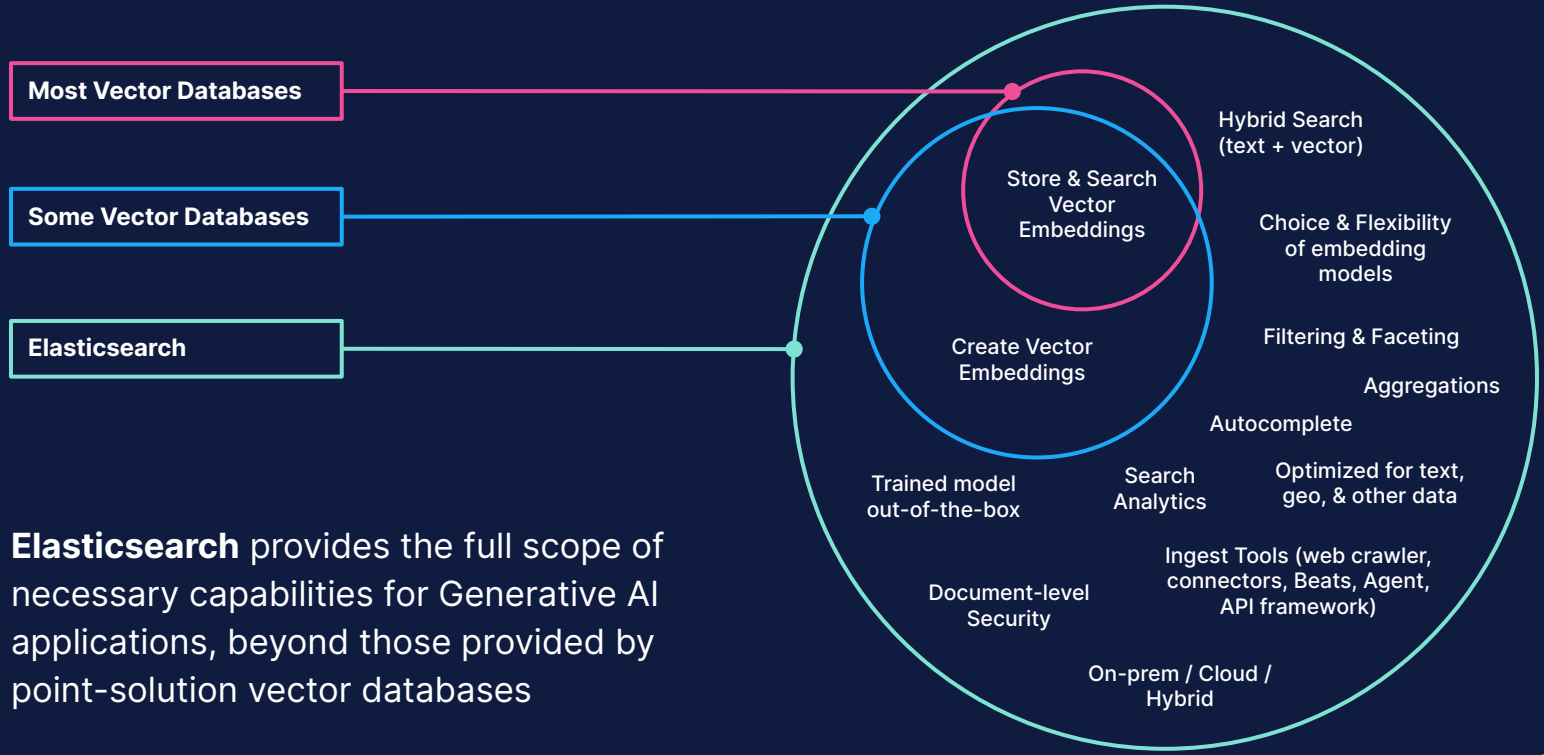
Support for ML-powered detection rules in Security
2020

ML-powered alerting rules
2021

Elastic Security Assistant powered by ESRE
2023

elastic

Vector search is your **starting** point...



Elasticsearch provides the full scope of necessary capabilities for Generative AI applications, beyond those provided by point-solution vector databases

elastic

Elasticsearch Relevance Engine™ for **Gen AI apps**

Text, vector, hybrid search	Textual search & vector database	RRF: hybrid scoring model (vector & textual search)	Filtering & faceting for kNN and hybrid queries
Choice of machine learning models	Host your transformer models, or use 3rd party LLMs (OpenAI)	Elastic's proprietary zero-shot ML model	Integration with 3rd party tooling like LangChain
Enterprise ready developer experience	Document and field Level security	Several major compliance frameworks covered	Build on-prem, or across cloud platforms on 50+ regions

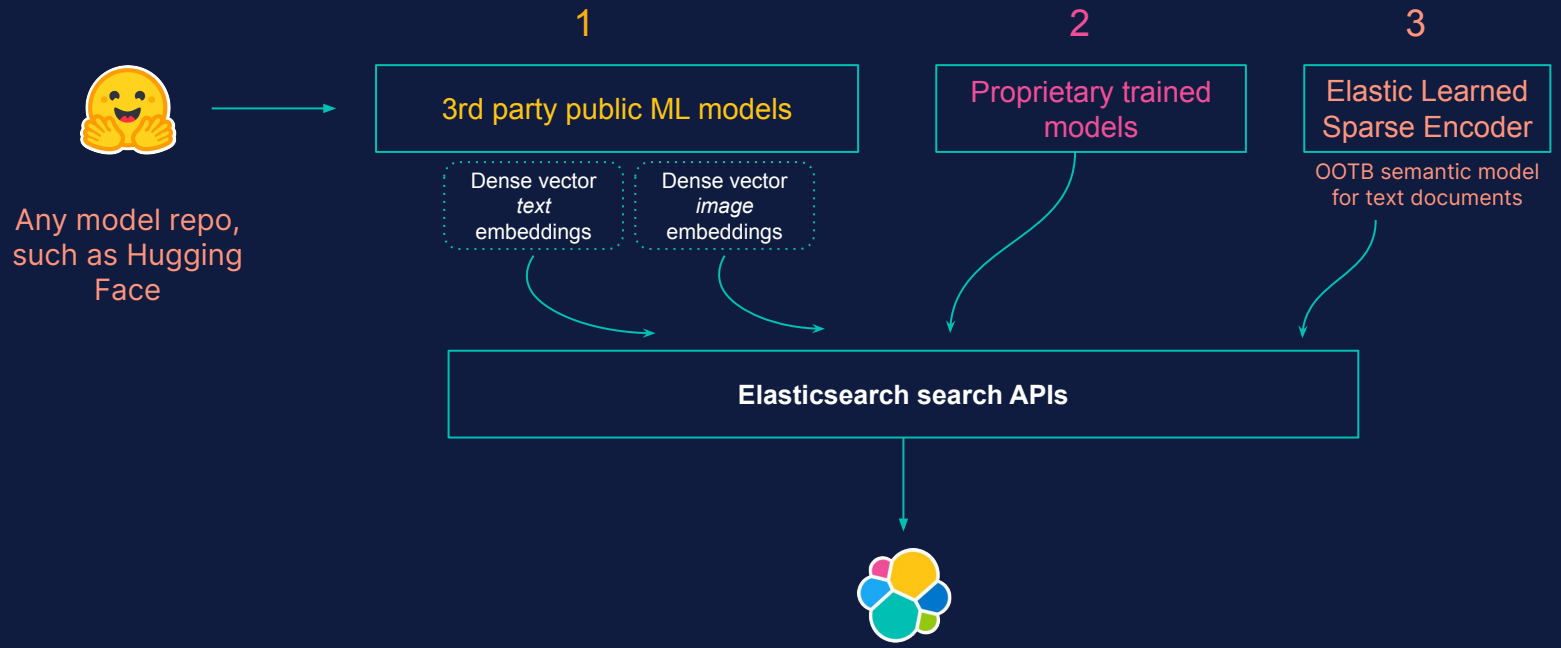
elastic

Elasticsearch Relevance Engine™ for **Gen AI apps**

Text, vector, hybrid search	Textual search & vector database	RRF: hybrid scoring model (vector & textual search)	Filtering & faceting for kNN and hybrid queries
Choice of machine learning models	Host your transformer models, or use 3rd party LLMs (OpenAI)	Elastic's proprietary zero-shot ML model	Integration with 3rd party tooling like LangChain
Enterprise ready developer experience	Document and field Level security	Several major compliance frameworks covered	Build on-prem, or across cloud platforms on 50+ regions

elastic

Choose ML models for **your** use-case



elastic

Elasticsearch Relevance Engine™ for **Gen AI apps**

Text, vector, hybrid search	Textual search & vector database	RRF: hybrid scoring model (vector & textual search)	Filtering & faceting for kNN and hybrid queries
Choice of machine learning models	Host your transformer models, or use 3rd party LLMs (OpenAI)	Elastic's proprietary zero-shot ML model	Integration with 3rd party tooling like LangChain
Enterprise ready developer experience	Document and field Level security	Several major compliance frameworks covered	Build on-prem, or across cloud platforms on 50+ regions

elastic

Elastic Learned Sparse Encoder: **semantic search** out-of-the-box

Data sets (BEIR benchmark)

Search results ranking models

	Average	TREC-COVID	NFCorpus	NQ	HotpotQA	FiQA	ArguAna	Touche-2020	DBPedia	SCIDOCS	FEVER	Climate-FEVER	SciFact
BM25	0.416	0.688	0.327	0.326	0.602	0.254	0.472	0.347	0.287	0.165	0.649	0.186	0.69
RRF (BM25/Dense)	0.449	0.697	0.317	0.445	0.611	0.318	0.474	**0.354**	0.353	0.159	0.746	0.238	0.671
Linear (BM25/Dense)	0.471	0.787	0.335	0.485	0.62	0.341	0.444	0.346	0.378	0.164	**0.778**	**0.272**	0.698
SPLADE	N/A	0.726	0.347	**0.537**	0.687	**0.347**	**0.526**	0.246	**0.436**	0.158			0.703
ELSER	0.471	0.747	**0.351**	0.524	0.67	0.339	0.5	0.263	0.415	0.156	0.777	0.218	0.695
RRF (BM25/ELSER)	**0.478**	**0.797**	**0.352**	0.468	0.674	0.311	0.497	0.347	0.411	**0.166**	0.762	0.24	**0.712**

Search quality *worse* than benchmark (BM25)

Search quality *better* than benchmark (BM25)

Want to learn more?

Search “elasticsearch sparse encoder blog”

elastic

What are our customers **building**?

elastic

Gen AI apps powered by search

Search for Customer service

"Feedback from our engineers is extremely positive. They now use Topic Search to solve 90% of service requests. They can deliver a better customer experience by easily finding on-target information and fixing issues much faster than before."

Sujith Joseph, Principal Enterprise Search & Cloud Architect, Cisco Systems

Legal e-Discovery search

"I'm thrilled about the benefits we can bring to customers through our investments to harness Elasticsearch within RelativityOne. We're excited about the potential to deliver powerful, AI-augmented search results to our customers."

Chris Brown, Chief Product Officer, Relativity

elastic

Where do we go from here?

(**Spoiler alert**: Full set of gen AI dev tools, and more than being the best vector database out there)

Coming soon, to a Gen AI search app near you...

Inference API

Pick and mix service providers that give you the dense vector embeddings you want

3rd party integrations

More integrations to popular 3rd party projects, resulting in easier AIOps for you

Result reranking capabilities

Making our text + vector hybrid search story even stronger

Int8 quantization and memory reduction

Reduction in disk usage, near-zero accuracy loss, and faster search

Passage Vectors

Automatic diversification over parent doc, complex passage and parent filtering with Lucene primitives

Improving Retrieval Quality for Search

9/27, Wednesday, 2 pm,
Downstairs, Fisher

Priscilla Parodi,
Principal Developer Advocate, Elastic

Visit our booth for a demo!

elastic