An abstract graphic on the left side of the slide. It features a blue silhouette of a person climbing a rope. The rope is represented by several thick, curved lines in shades of blue and green. The person is positioned on the left, with their arms and legs extended as if climbing. The background is white.

# **Personalized Chat Assistant – Leveraging Azure OpenAI Model & Elasticsearch**

**Ashish Tiwari**

**DevRel Engineer, Elastic**

**X @\_ashish\_tiwari**



**Actually, search is**  
**MORE IMPORTANT THAN**  
**EVER!**



# LLMs Need to be **Grounded**

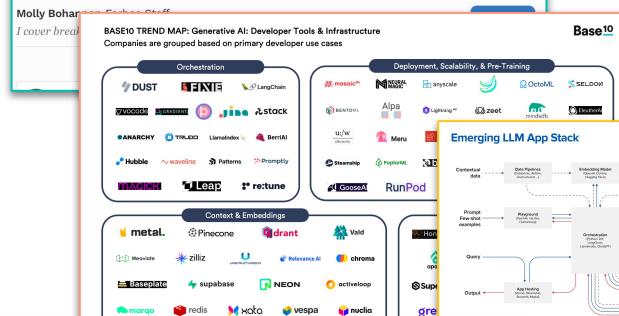


## Generative AI apps come with specific challenges:

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

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

Molly Bohan



OpenAI

Research ▾ API ▾ ChatGPT ▾ Safety Company ▾

GPT-4 generally lacks knowledge of events that have occurred after the vast majority of its data cuts off (September 2021), and does not learn from its experience. It can sometimes make simple reasoning errors which do not seem to comport with competence across so many domains, or be overly gullible in accepting obvious falsehoods as introducing security vulnerabilities into code.

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

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

Comment

ars TECHNICA

BIZ & IT TECH SCIENCE POLICY CARS GA


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



 **Thilak Rao**   
@Thilak

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



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

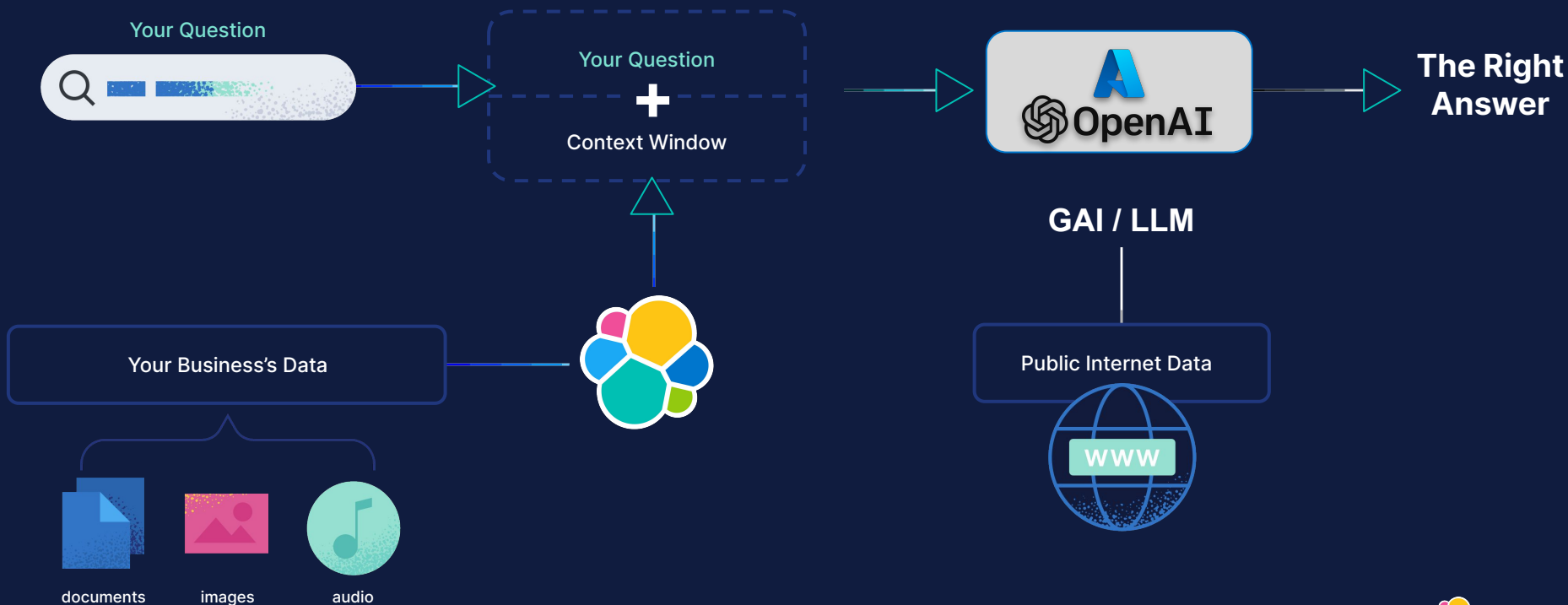
```
from collections import  
OrderedDict  
  
class LRUCache:  
    def __init__(self,  
capacity: int):  
        self.capacity =  
capacity  
        self.cache =
```

 Ask me anything... 



# **RAG – Retrieval Augmented Generation**

# Retrieval Augmented Generation





# **Why Elasticsearch?**



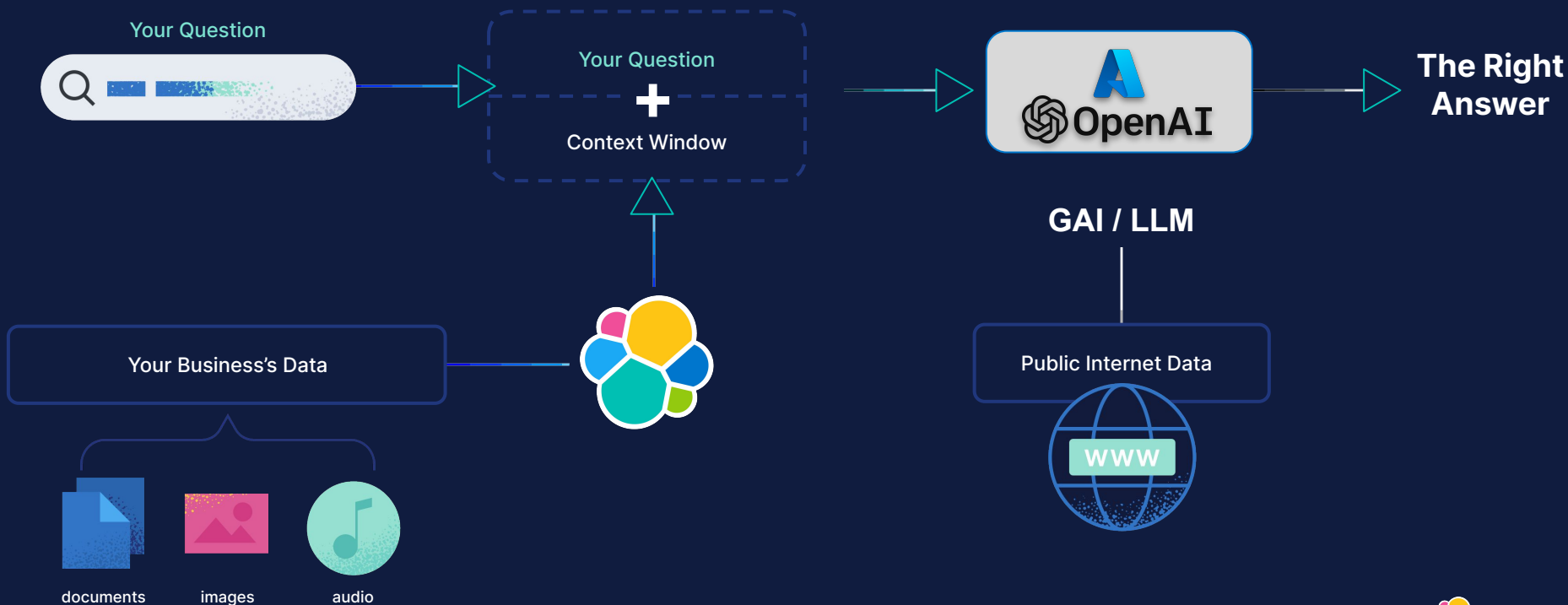
# The Elastic Search Platform



# We've Come a Long Way

results pinning permissions-based search results clustered indices language analyzers synonyms language identification highlighters bkd trees inference match enrich processors cross-cluster search aggregations data encryption field level security filters full-text search document store type ahead knn search percolators geospatial data types vector search dynamic mapping runtime fields asynchronous search document level security analyzers suggesters inverted index tokenizers relevance scoring query DSL auto Corrections lookup runtime field geo-match enrich query profiler optimizer

# Retrieval Augmented Generation





# Flowise AI

*Open source UI visual tool to build  
your customized LLM orchestration  
flow & AI agents*

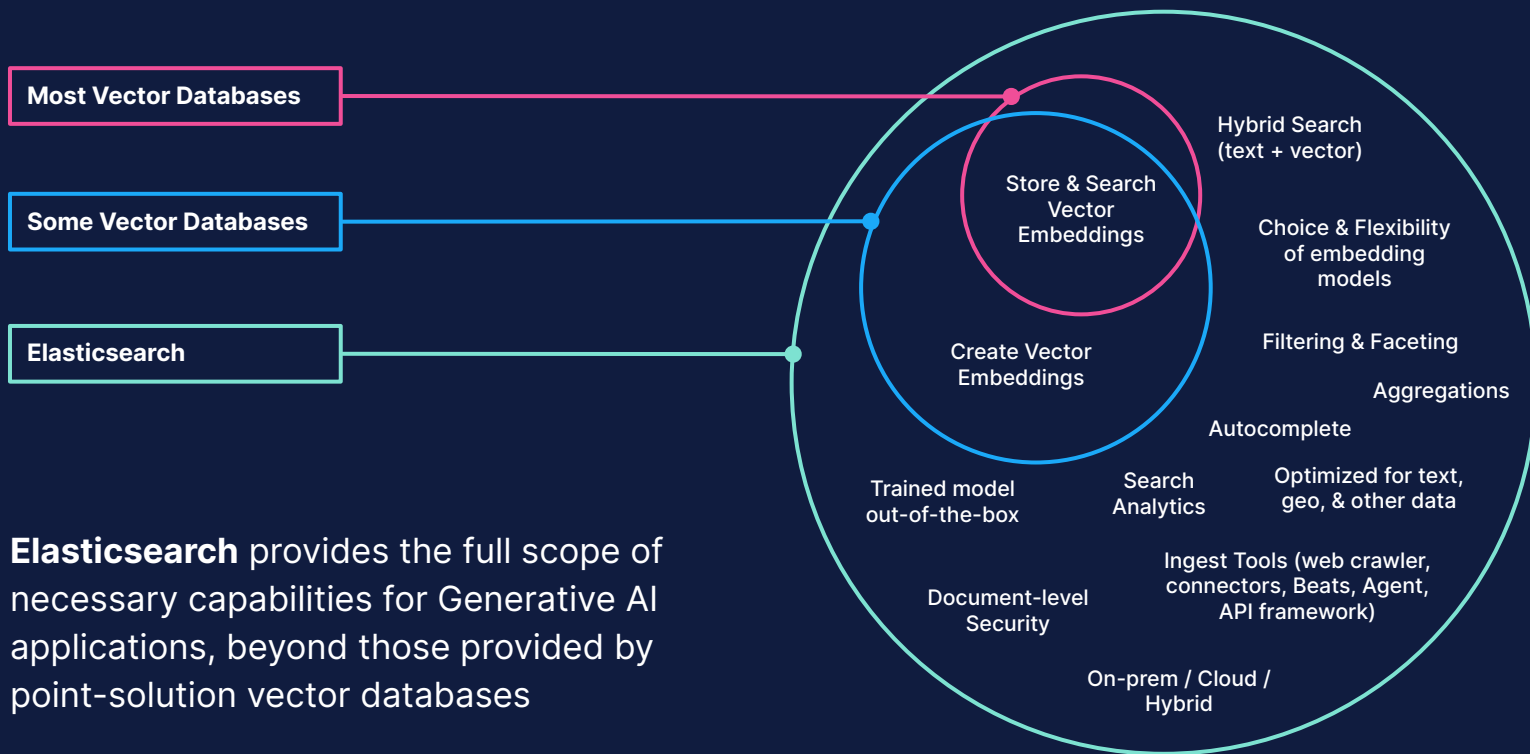


# Demo



# **Why Elasticsearch?**

# Elastic has ALL the capabilities you need





We're using Elastic as a vector database because of its inherent flexibility, scalability, and reliability. Elastic continually elevates the game by rapidly delivering new features that support Machine Learning and generative AI.

**- Peter O'Connor, Engineering Manager of Platform Engineering, Stack Overflow**



Search is critical for elevating Udemy's user experience — matching users to relevant educational content, which is why Elastic has been a long-term partner of ours. We've used Elastic as our vector database since upgrading to Elastic Cloud last year, and it has opened up new opportunities for our business. We've seen increased query speed and resource efficiency as we've scaled vector search across our innovative education solutions.

**- Software Engineering Team, Udemy**





## Resources

- **Azure + Elastic** -  
<https://www.elastic.co/virtual-events/unleash-your-creativity-with-generative-ai-and-elastic-on-microsoft-azure>
- **Generative AI** - <https://www.elastic.co/what-is/generative-ai>
- **Generative AI Blog** - <https://www.elastic.co/blog/category/generative-ai>
- **Workshop guide** - <https://ashish.one/talks/chatgpt-elasticsearch/>
- **ELSER** - <https://www.elastic.co/guide/en/machine-learning/current/ml-nlp-elser.html>
- **Vector Search** - <https://www.elastic.co/what-is/vector-search>
- **Elasticsearch + ChatGPT** - <https://www.elastic.co/blog/chatgpt-elasticsearch-openai-meets-private-data>
- **Relevance Scoring** - <https://www.elastic.co/blog/whats-new-elasticsearch-8-8-0>
- **Elasticsearch + VertexAI** -  
<https://cloud.google.com/blog/products/ai-machine-learning/interactive-search-with-google-cloud-and-elasticsearch>



# Elastic Chennai User Group





# Thank You

 [in/ashishtiwari93](https://www.linkedin.com/in/ashishtiwari93)

 [@\\_ashish\\_tiwari](https://twitter.com/_ashish_tiwari)

