Enhancing Generative AI with Graphs

Agenda

- What is a graph?
- How GenAI and Graphs are used together in today's emerging solutions
- Introduction to GraphRAG
- Why GraphRAG is the hot topic in data driven retrieval
- When and where to use GraphRAG
- Demo
- Why you might not want to use GraphRAG

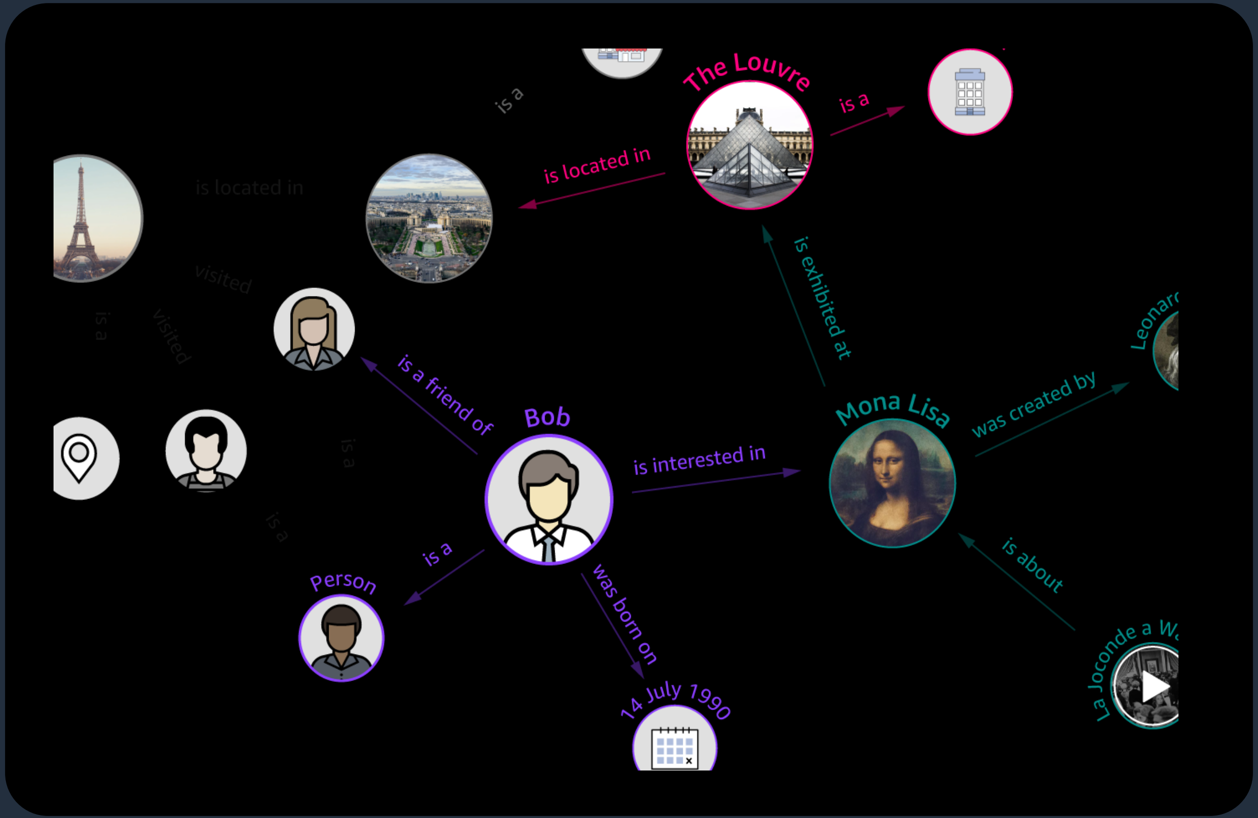
Technical challenges graphs help solve

- Combining data across silos
- Finding common connections or paths
- Working with heterogenous data with complex relationships
- Data full of many-to-many relationships

Graphs work with data like a mind map tool instead of multiple excel spreadsheets.

What is a Knowledge Graph?

Understanding the who, what, when, and where



Benefits

1. Link disparate data sources

Link disparate and heterogeneous data sources together to discover hidden connections

2. Improved search results

Increase productivity by making data easily accessible through improved search relevance

3. Augment ML/AI

Improve the efficiency and effectiveness of machine learning models by providing context and augmentation with related content

Why do I care?

In the Financial Services industry, managing and analyzing vast amounts of data is critical.

- Customer data and email to provide personalized services
- Filings, research reports, market data, and news stories to make effective investment decisions
- Above plus public data to recommend unique investment opportunities or suggest bespoke financial products
- Know your customer and fraud detection

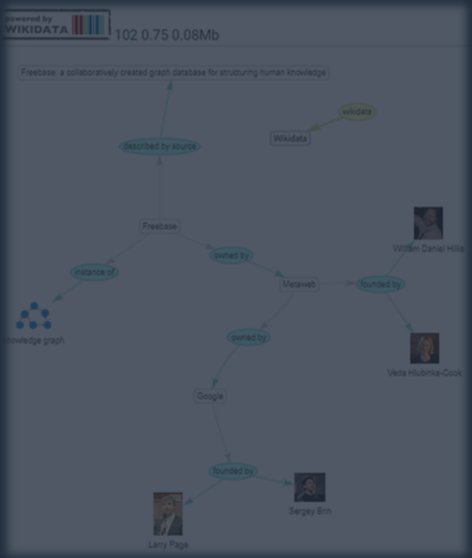
While we walk through these examples today, imagine a research assistant tool acting as a force multiplier for your team:

- Minimizing the search and discovery work
- Validate and challenge decisions with pros and cons
- Agents that can keep an eye on data around the clock and alert to emerging opportunities

Graphs enhance GenAI application

Graph Generation

Generate a graph from a given corpus of structured or unstructured data



Graph Enhanced RAG (GraphRAG)

Enhance a RAG application with relevant information to provide more comprehensive and explainable answers

Images from [Wikimedia Commons] and [Nick Youngson]

RAG is a powerful architecture pattern but has complex data challenges

CONNECTEDNESS
data spread across multiple disparate documents is hard to retrieve

SPECIFICITY
embeddings are sparse representations of data which may lack crucial details

EXPLAINABILITY
explaining the relevance of data retrieved is demanding

aws

The most relevant information to a question may be the most connected ideas, not the most similar text.

What benefit does a graph provide a RAG applications?

Vectors find relevant information using similarity in language.

- e.g. Sentences in a document that discuss similar locations/names/topics will be highly similar using a vector search.

Graphs find relevant information using connected ideas

- e.g. Entities/concepts and interactions will be highly connected using a graph search.

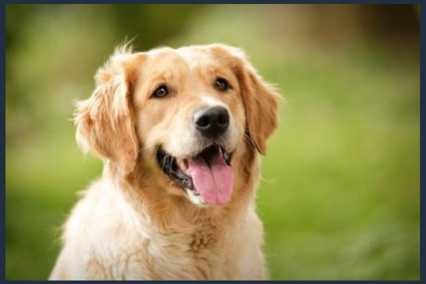
How is it doing this?

Similarity in vector space compares mathematical closeness

Zucchini is similar to *summer squash* and *courgette*

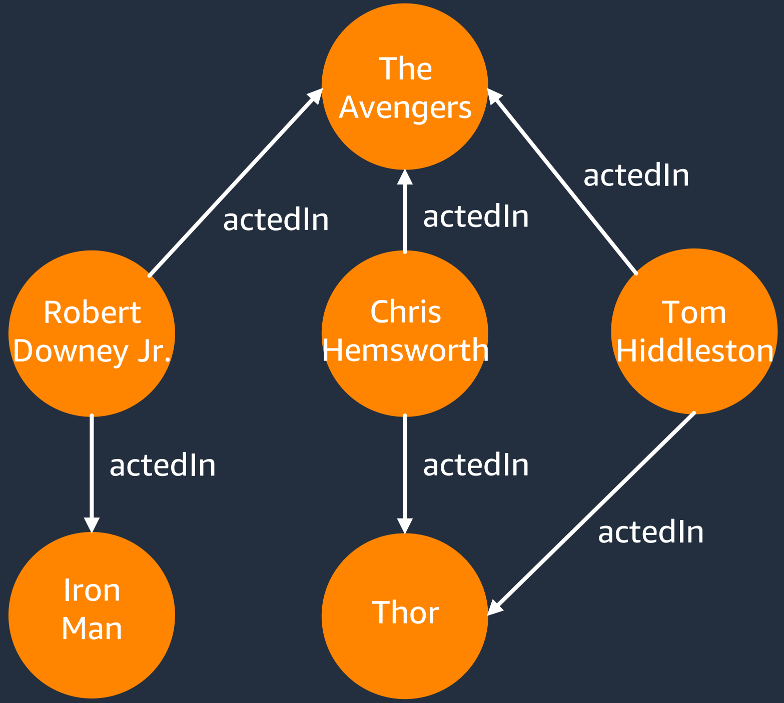
Vectors can represent …

… A text embedding model
… An image embedding model

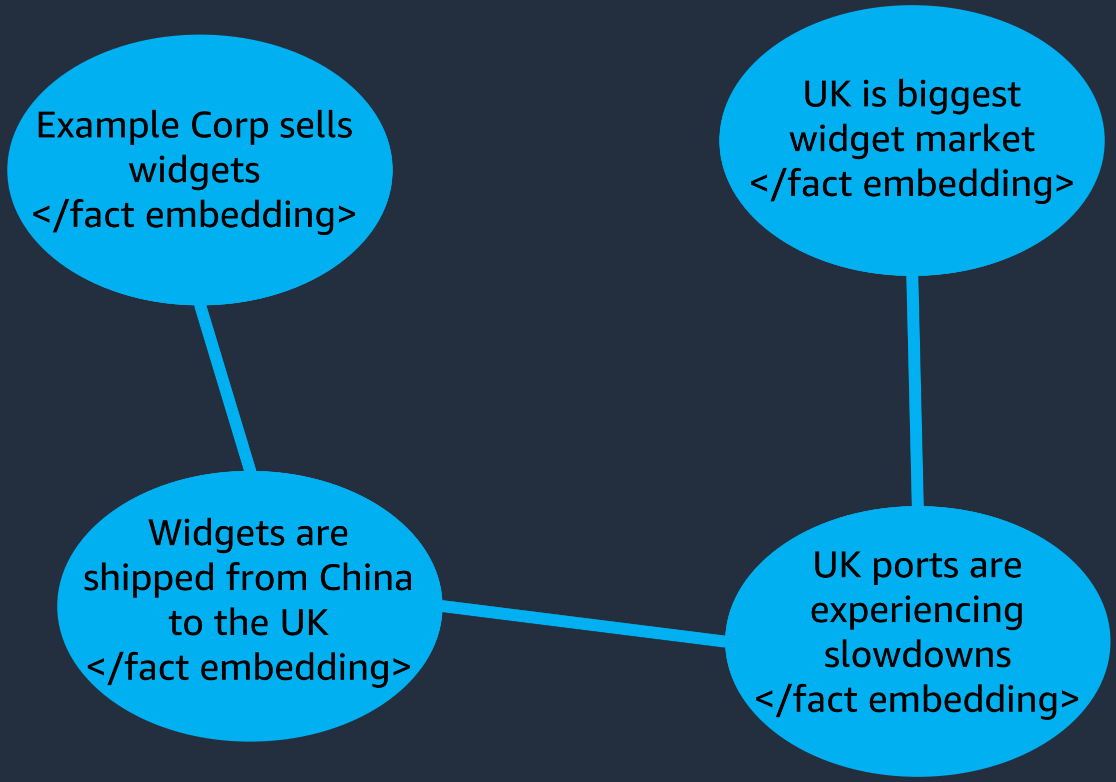


Relatedness in graph space compares shared connections

Chris Hemsworth and *Tom Hiddleston* are related because they've starred in multiple movies together



Example: Example Corp. Quarterly Report Data

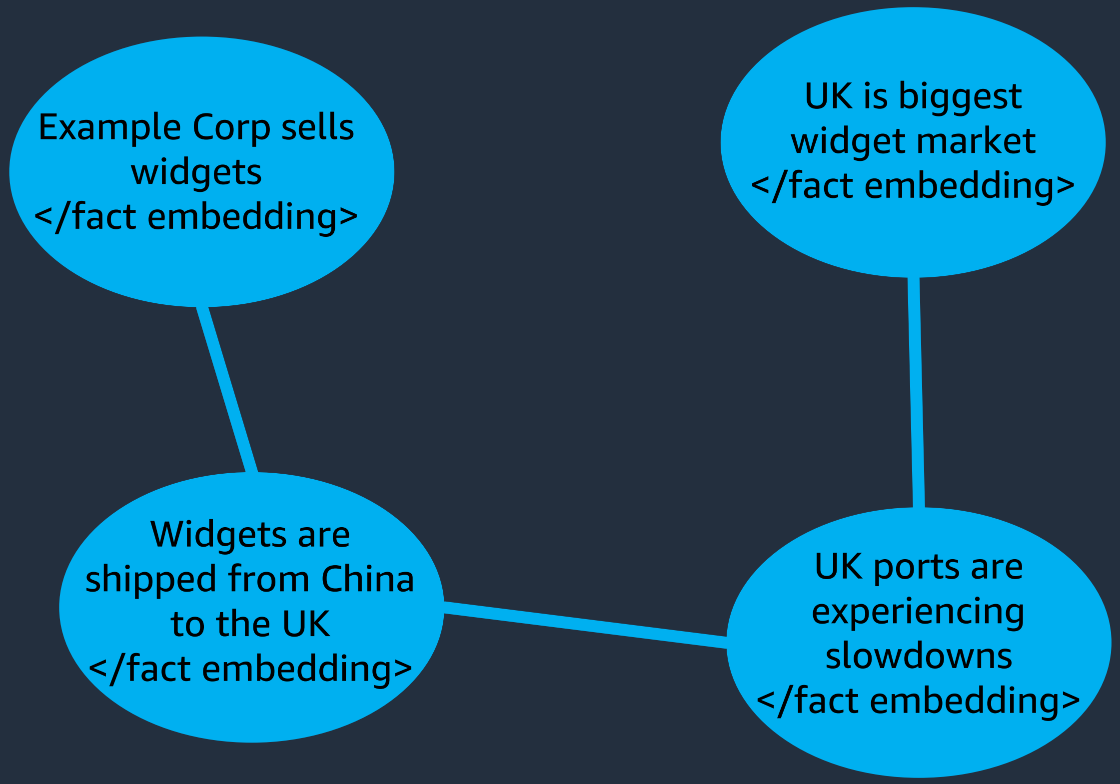


Vector Search: What is the outlook for widget sales in the UK?

STEP 1: AN EMBEDDING IS CREATED OF THE QUESTION BEING ASKED

Vector Space

</question embedding>

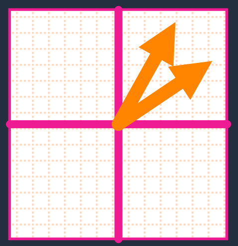


Vector Search: What is the outlook for widget sales in the UK?

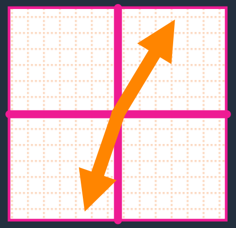
STEP 2: SIMILARITY SEARCH IS RUN TO FIND THE MOST SIMILAR LANGUAGE

Vector Space

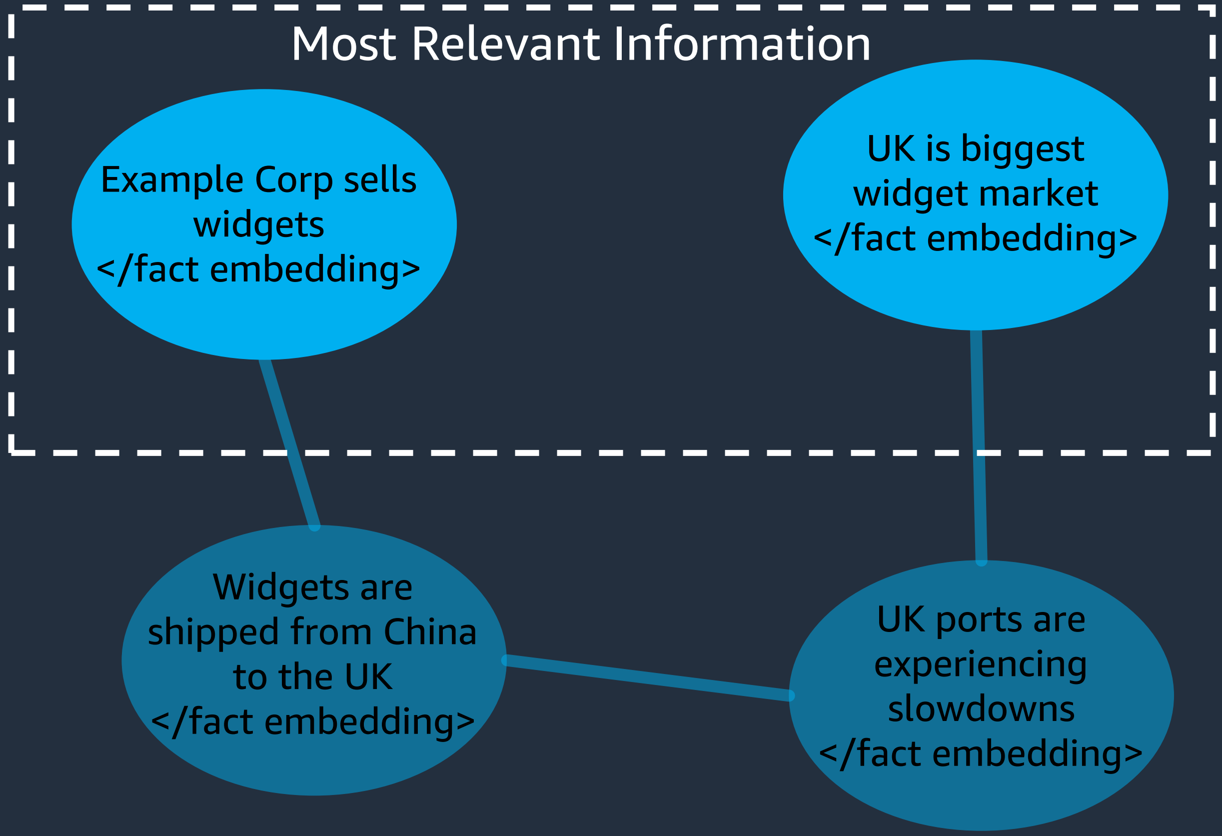
</question embedding>



More Similar



Less Similar

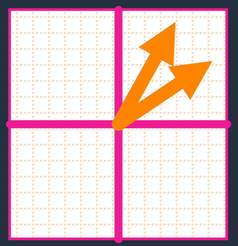


Vector Search: What is the outlook for widget sales in the UK?

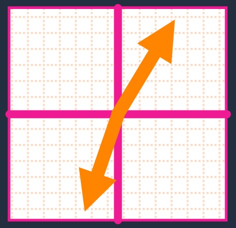
STEP 3: MOST RELEVANT INFORMATION SENT TO LLM FOR RESPONSE

Vector Space

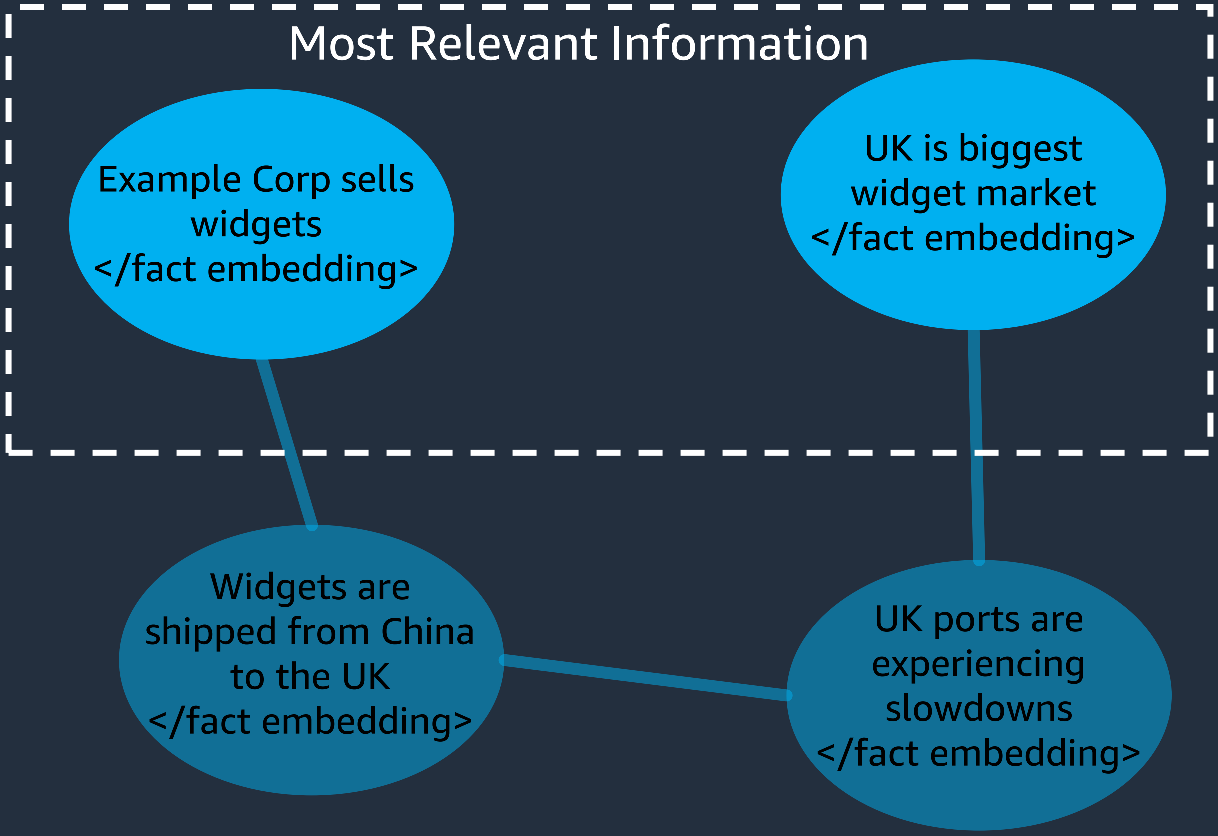
</question embedding>

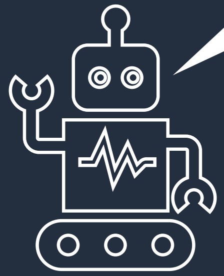


More Similar



Less Similar





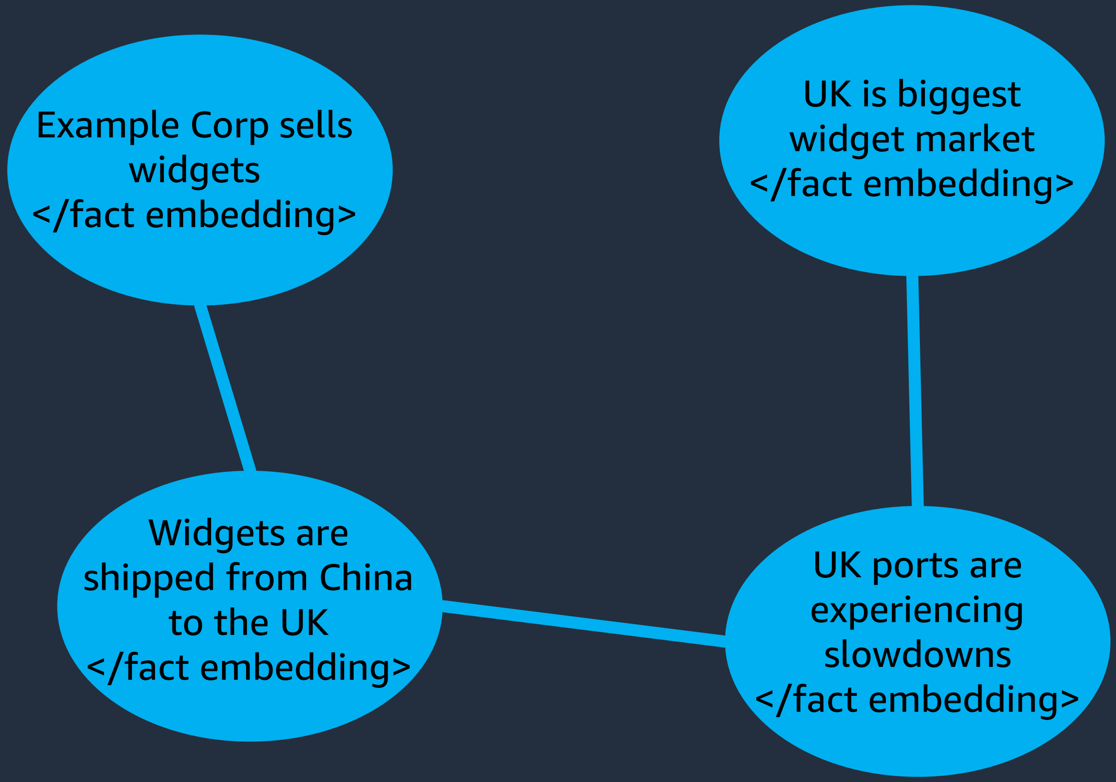
LLM Response

Graph Search: What is the outlook for widget sales in the UK?

STEP 1: AN EMBEDDING IS CREATED OF THE QUESTION BEING ASKED

Vector Space

</question embedding>

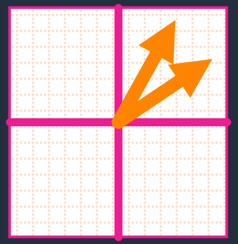


Graph Search: What is the outlook for widget sales in the UK?

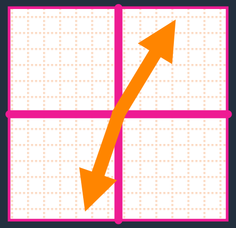
STEP 2: SIMILARITY SEARCH IS RUN TO FIND THE STARTING NODES

Vector Space

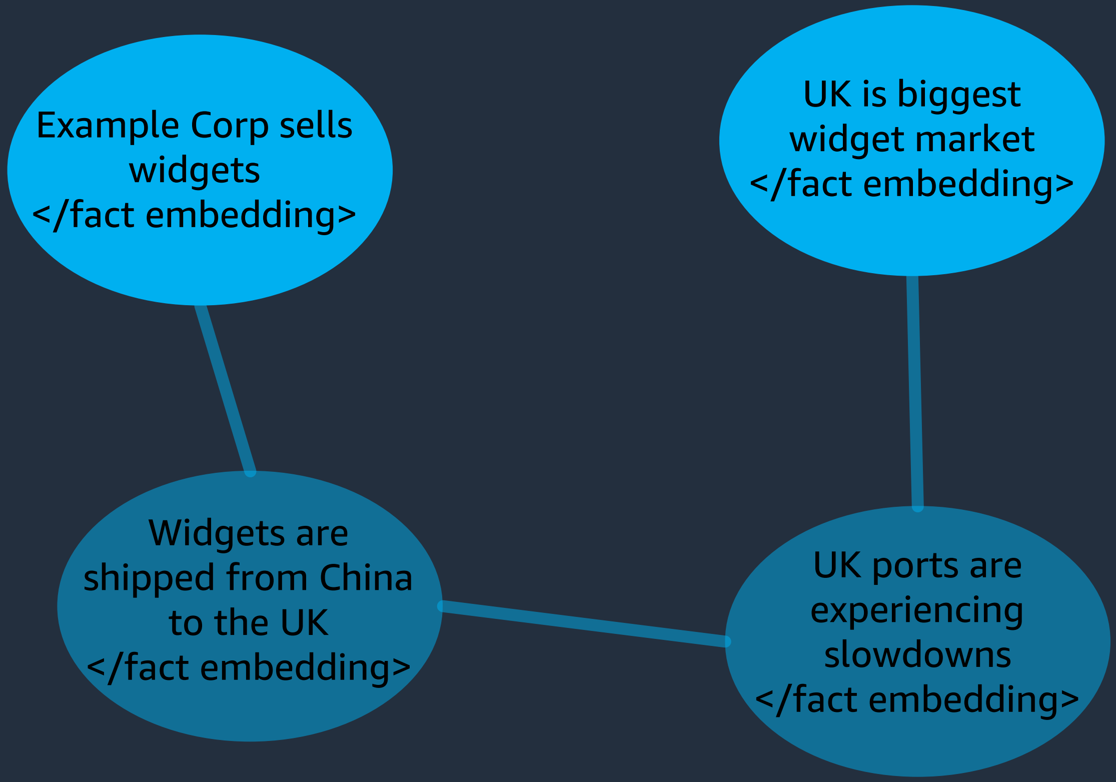
</question embedding>



More Similar



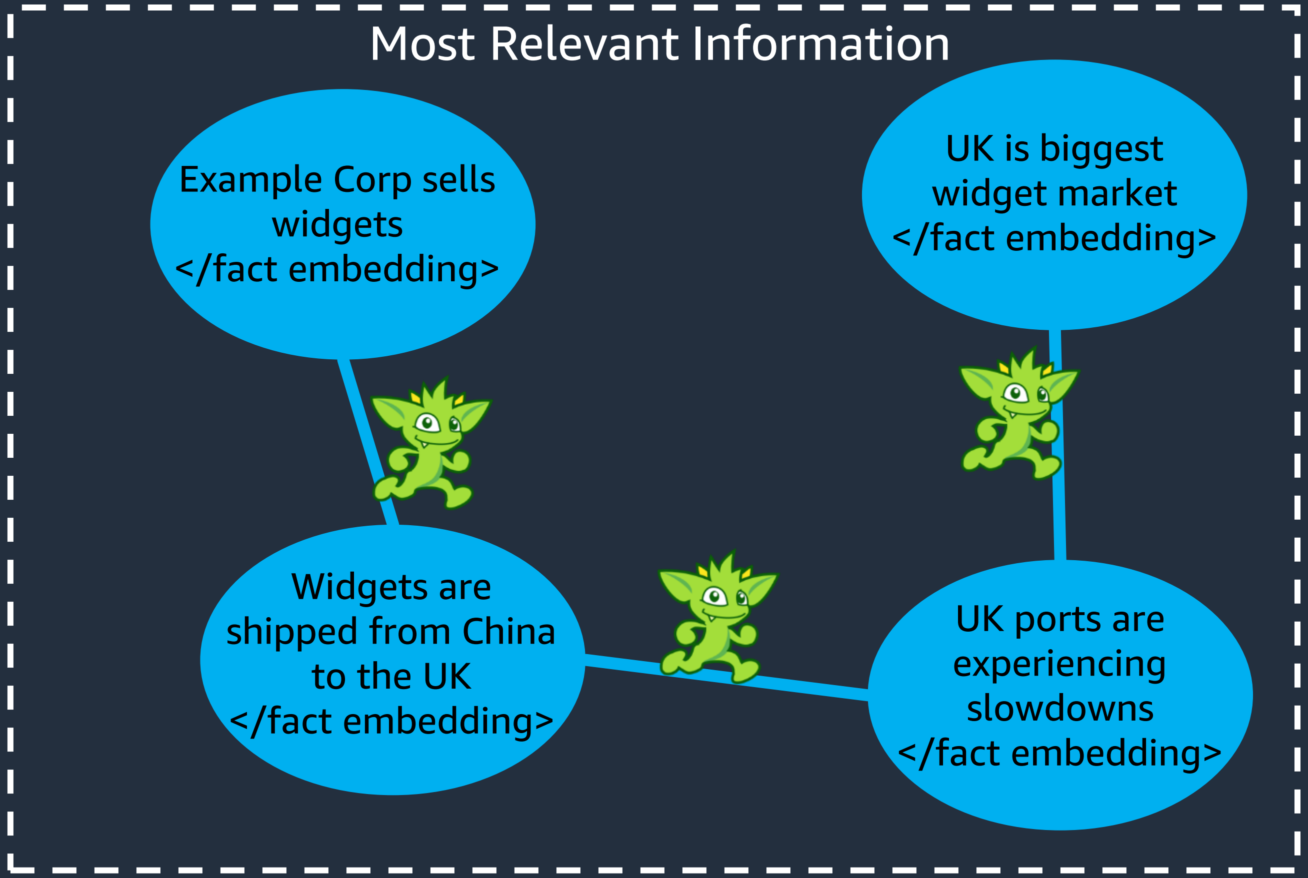
Less Similar



Graph Search: What is the outlook for widget sales in the UK?

STEP 3: GRAPH IS TRAVERSED TO FIND THE CONNECTED IDEAS

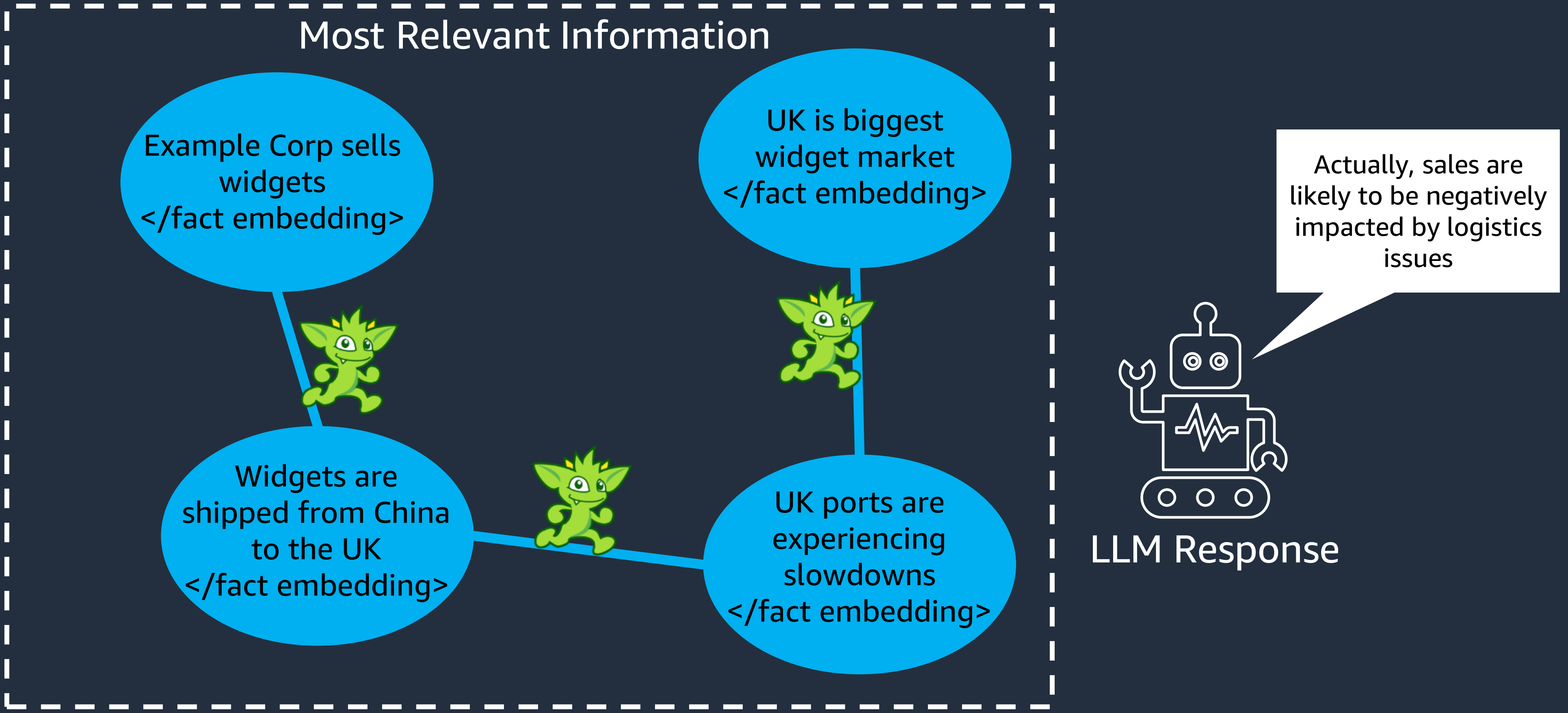
Graph Space



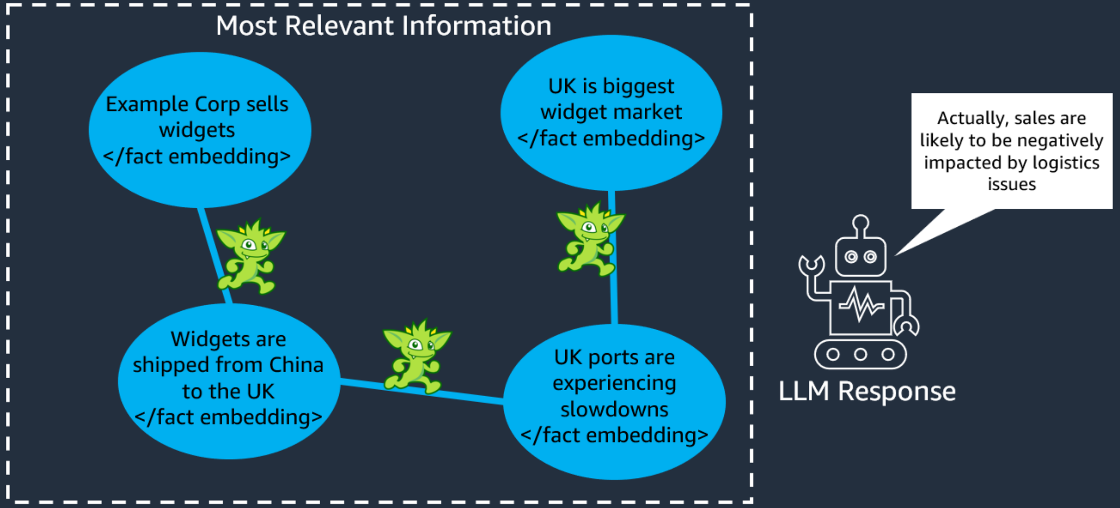
Graph Search: What is the outlook for widget sales in the UK?

STEP 4: MOST RELEVANT INFORMATION SENT TO LLM FOR RESPONSE

Graph Space

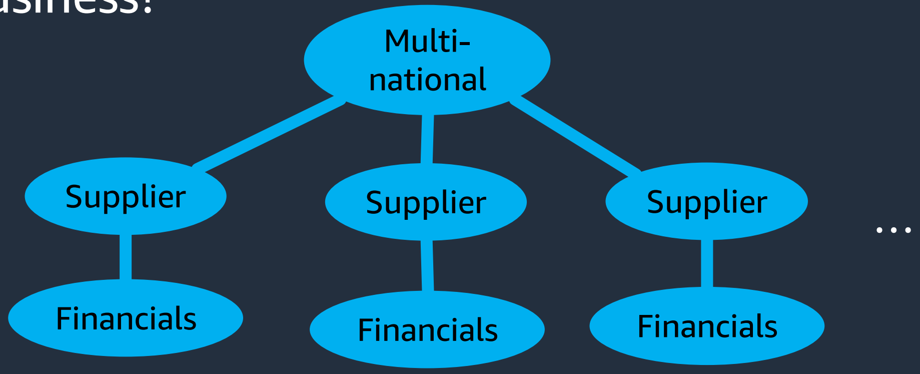


What types of questions does GraphRAG excel at?



Inference query

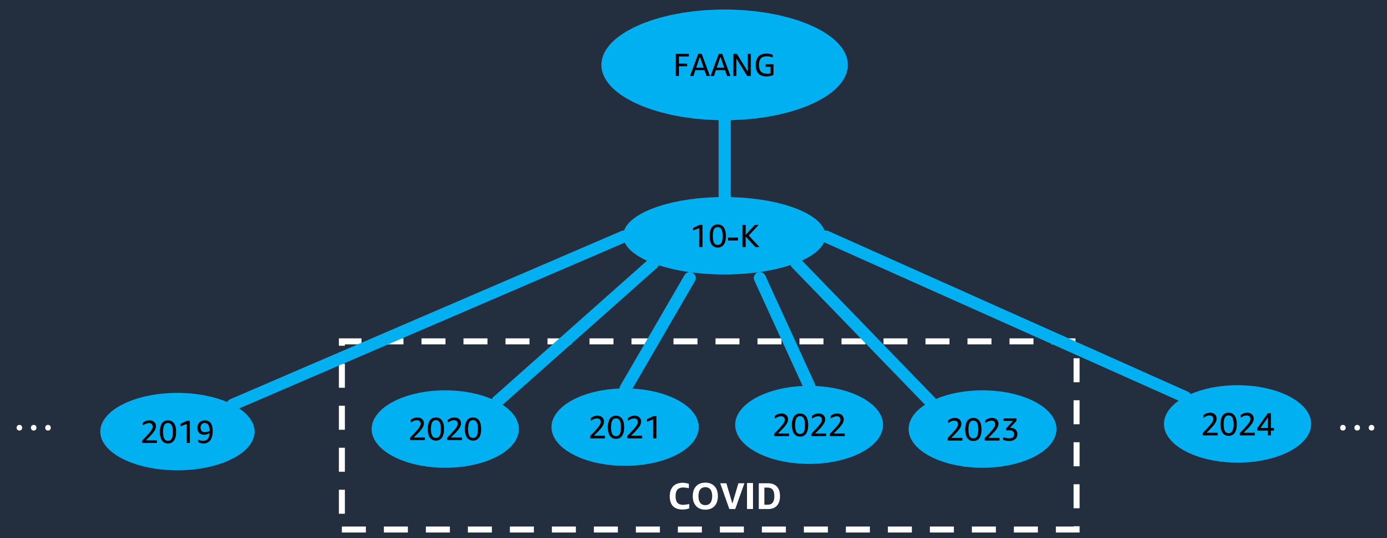
Which suppliers are most dependent on this business?



Comparison query

What types of questions does GraphRAG excel at?

How did the risk factors change among FAANG companies, as reported in their 10-K filings, before, during, and after the COVID-19 pandemic?

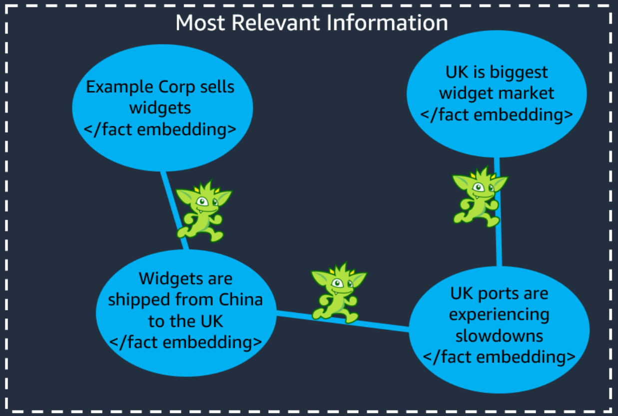


Temporal query

Explainable and Auditable

Dimension	Document	Question
0	-0.0357713	-0.039272
1	-0.000768	0.064294
2	0.054941	0.059986
…		
1023	0.020356	-0.045072

VS



similarity	Dog	Puppy	Cat	Kitten
Dog	1.0			
Puppy	**0.3901**	1.0		
Cat	0.3647	0.1787	1.0	
Kitten	0.2449	0.2151	**0.4386**	1.0

Why is a dog and puppy 0.3901 similar, but a kitten and cat 0.4386?

I'm guessing there is not a single person that can fully explain it.

Demo

Considerations before using GraphRAG

- For many use cases, RAG is “good enough”. **<u>Make sure your requirements require the additional complexity</u>**.

- More computationally expensive
  - In this demo:
    - RAG encodings cost: $0.00001 (4 docs + question)
    - RAG answer cost range: $0.00008 (Llama 3.2 1B) to $0.00060 (Llama 3.2 90B)
    - GraphRAG encodings cost range: $0.0005 (1B) to $0.0039 (90B)
    - GraphRAG answer cost range: $0.0003 (1B) to $0.0020 (90B)

- Graph expertise is not widespread. Prepare for a learning curve.

- Heavy reliance on LLM calls = slower response times

- Consider: Send most calls to RAG and use GraphRAG when you need it.

Key Takeaways

- GraphRAG is quickly becoming recognized as the best methodology for improving the quality and transparency of RAG-style LLM powered solutions.

- If your workload requires traceability, explainability, and/or auditability, GraphRAG provides superior ability to meet those requirements.

- If your workload involves complex questions involving contextual inference, comparing various sources, or comparisons across time, GraphRAG better handles these queries than RAG alone.

- Be aware of the additional costs and complexity involved. Everyone wants the best, but good enough may suffice.

Thank you!

Brian O’Keefe (he/him)

briokeef@amazon.com

brianokeeferochester

Backup slides

Graph models for RAG applications

	Triple/Triple	Keyword	Topic/Lexical	Community Based
Description	Model is based on subject-object-predicate triple extracted from chunks	Model is based on keywords, categories, and labels extracted from chunks	Model is based on chunks	Model Entities and relationships are extracted and hierarchical communities are created
Model Entities	(subject)-[predicate]->(object)	(Chunk)->(Keyword) (Keyword)->(Keyword)	(Source)->(Chunk) (Chunk)->(Topic) (Topic)->(Statement) (Statement)->(Fact)	(Entity)->(Entity) (Entity)->(Cluster) (Cluster)->(Summary)
Best use case	Questions where connectedness in the data is key to relevant data	Questions where expressiveness and metadata are key to relevant data	Questions where relevant data is found by connecting across multiple documents/chunks	Local and Global search questions